“Specifications of generic social and real world descriptors”

D3.2

‘I-SEARCH_D3.2_WP3_UNIGE-v1.61.doc’

Version: 1.61

Distribution Level: PU

- ***Distribution level***
  PU = Public,
  RE = Restricted to a group of the specified Consortium,
  PP = Restricted to other program participants (including Commission Services),
  CO= Confidential, only for members of the I-SEARCH Consortium (including the Commission Services)

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248296

The I-SEARCH Project Consortium groups the following organizations:

Partner Name	Short name	Country
Centre for Research and Technology Hellas / Informatics and Telematics Institute	CERTH	GR
JCP-Consult	JCP	FR
Institut National de Recherche en Informatique	INRIA	FR
Athens Technology Center	ATC	GR
Engineering Ingegneria Informatica S.p.A.	ENG	IT
Google Ireland Limited	Google	IR
University of Genoa	UNIGE	IT
Exalead	EXL	FR
Erfurt University of Applied Sciences	FHE	DE
Accademia Nazionale di Santa Cecilia	ANSC	IT
EasternGraphics GmbH	EGR	DE

Executive Summary:

The basic scope of this document is to formally specify the generic social, emotional and real world descriptors that will be included in the RUCoD file format. Initially, the current usage of real-world descriptors in existing application is examined and then the current status of social, emotional and real world descriptors is discussed. More specifically, the valence-arousal and the energy-velocity spaces are considered for the emotional descriptors; the environmental, empathy, saliency and leadership descriptors along with the geometric group features are examined as candidate social descriptors. Real-world descriptors are also examined and divided into broad categories, i.e. the location descriptors, the time descriptors, the environmental descriptors and the ID descriptors. Special care has been taken in order to support future extensions of real-world descriptors.

The forecasted interaction of the components that involve social, emotional and real-world descriptors is also presented, based on the components identified during the definition of the I-SEARCH System Architecture (D2.2). Additionally, existing standards related to the selected descriptors are also examined in order to select the appropriate representation to be included in RUCoD. Finally, the descriptors that will be used for the needs of I-SEARCH project are formally specified and appropriately associated with the selected Use cases (D2.1).

Document Identity

Title:	Specifications of generic social and real-world descriptors
Subject:	Social and real-world descriptors
Number:	
File name:	I-SEARCH_D3.2_WP3_UNIGE-v1.61.doc
Registration Date:	18 April, 2010
Last Update:	30 May, 2011

Revision History

No.	Version	Edition	Author(s)	Date
1	0	1	Vasilios Darlagiannis & Sotiris Malassiotis - CERTH	18/4/2010
	Comments:	Initial Release, State of the art		
2	0	5	Thanassis Mademlis – CERTH & Xavier Le Bourdon - JCP	15/3/2011
	Comments:	Added initial version for specification of real-world Descriptors		
3	0	6	Alberto Massari - UNIGE	4/4/2001
	Comments:	Social and emotional descriptors		
4	1	0	Thanassis Mademlis – CERTH	14/4/2011
	Comments:	Integration of contributions		
5	1	1	Xavier Le Bourdon – JCP	22/04/2011
	Comments:	Revision & Corrections		
6	1	2	Alberto Massari – UNIGE	25/04/2011
	Comments:	Revision & Corrections		
7	1	3	Thanassis Mademlis – CERTH	28/04/2011
	Comments:	Pre-final Version		
8	1	4	Thanassis Mademlis & Dimitrios Tzovaras–CERTH & Alberto Massari - UNIGE	27/5/2011
	Comments:	Revisions & corrections		
9	1	5	Thanassis Mademlis & Dimitrios Tzovaras - CERTH	30/5/2011
	Comments:	Final Version		
10				
	Comments:			
11				
	Comments:			
12				
	Comments:			
13				
	Comments:			
14				
	Comments:			
15				
	Comments:			

Table of Contents

List of figures

List of tables

List of abbreviation

2D	Two Dimensional
3D	Three Dimensional
3DLD	3D Location Descriptor
3DLD-B	Box 3D Location Descriptor
3DLD-C	3D Location Descriptor with Circle support area
ARML	Augmented Reality Markup Language
BF	Beaufort
BNWD	Beaufort Number Wind Descriptor
CO	Content Object
CTD	Celsius Temperature Descriptor
EXIF	Exchangeable Image File Format
GML	Geo Markup Language
GPS	Global Positioning System
GSD	Generic Sensor Descriptor
HD	Percentage Humidity Descriptor
IPTC	International Press Telecommunications Council
ISO	International Standards Organisation
KML	Keyhole Markup Language
MPQF	MPEG Query Format
MXF	Material eXchange Format
NE	North-East
OGC	Open Geospatial Consortium
RFID	Radio Frequency IDentifier
RUCoD	Rich Unified Content Descriptor
RW	Real-world
UK	United Kingdom
W3C	World Wide Web Consortium
W3CT	W3C Time Descriptor
XML	eXtensive Markup Language

1. Introduction

The I-SEARCH project aims to develop a novel multimedia search engine that will be able to handle specific multimedia types (such as text, image, video and 3D) among with expressive, emotional social and real-world information. Expressive, emotional social and real world information is expected to play an important role during the retrieval phase because these types of information provide a different, user-oriented, perspective in the retrieval. These types of information are more semantic-oriented, characterise the Content Object and the combined retrieval using all low-level, emotional, social and real world descriptors during retrieval is expected to provide high-quality results. More specifically, the adoption of real-world descriptors enables intuitive search and retrieval operations on appropriate mobile devices, which exploit special equipment (such as GPS) and automatically update the results based on the user location. Expressive, emotional and social descriptors provide novel query capabilities in the end-user interface. Emotional and real-world descriptors are initially involved during the content creation procedure, where the existing Content Objects low-level descriptors, will be enriched using emotional and real-world descriptors; all descriptors will be stored using the RUCoD format.
The user will be capable to query the system by exploiting these features which are expected to alter the user perception in search and retrieval purposes. Furthermore, after the presentation of retrieved results to the user, the user will be capable of refining the results by altering the initial query by providing additional emotional, social and real-world features.

The present document describes the specifications of emotional, social and the real-world descriptors that will be included in the Rich Unified Content Descriptor (RUCoD) for the needs of the I-SEARCH project. This document provides a review of the existing methodologies and standards concerning emotional, social and the real-world description of the multimedia elements and selects the appropriate features for describing the content which are expected to greatly enhance the retrieval accuracy and efficiency of the integrated I-SEARCH platform. The feature selection proposed in this document has been majorly guided on the use case scenarios and their future extensions as presented in D2.1.

The deliverable is organised as follows: Initially, related work concerning the utilisation of real-world descriptors in existing systems are briefly reviewed in Section 2, while the interference of social, emotional and real-world descriptors in the I-SEARCH architecture is described in Section 3. In Sections 4, 5 and 6 the existing status regarding the Emotional, Social and Real-world descriptors are presented, respectively and the descriptors that will be utilized in I-SEARCH platform are specified in Section 7. The metric functions utilised for their comparison is provided in Section 8. .

2. Related Work

In the next paragraphs, the state-of-the-art regarding applications and services that consider real-world descriptors is summarised. The gathered information will be considered in the definition of the appropriate descriptors for RUCoD and will guide the development of the social, emotional and real-world descriptor extractors. This information includes context-related approaches, such as context-extraction and context acquisition methods along with the utilization of these descriptors for search and retrieval purposes.

2.1 CONTEXT-BASED APPROACHES

A survey of context-aware mobile computing projects is collected at [1], where a number of context-aware applications are listed, approaches to sense and model the context are discussed, and supporting infrastructures and security and privacy issues are carefully examined.

A comprehensive context classification is provided in [2], where context can be anything from low-level parameters, such as time and temperature, to highly abstract concepts such as intention and social relationship. Using a human user-centred approach, the context information can be roughly expanded along three dimensions:

- Space/physical: the user's outside environment, the user's own activity in reacting to the environment, and the user's internal physical and mental status;
- Time/history: current time-of-day that is conventionally assumed appropriate for some activities, personal and related group's activity schedule, personal activity history and preferences;
- Spiritual/social relationship: regarding people the user will likely care about.

2.2 LANDMARK-BASED AND GEO-REFERENCED CONTEXT CAPTURING APPROACHES

Pedestrian navigation services [3] enable people to retrieve precise instructions to reach a specific location. The localization accuracy of satellite positioning systems is satisfactory for car navigation, but insufficient for pedestrians. Humans orient themselves and navigate through space with the help of cognitive maps – mental representations of the surroundings, which form a model of the world according to the way it is perceived by a specific individual. Humans need salient objects for orientation and navigation. These objects (so-called landmarks) serve as reference points in the environment, which help to structure space and support navigation. Landmarks are stationary, distinct and salient objects or places, which serve as cues for structuring and building a mental representation of the surrounding area.

A system that provides automatic annotation on collections of geo-referenced photos is described in [4]. As a user uploads a photograph a place of origin is estimated from visual features which the user can refine. Once the correct location is provided, tags are suggested based on geographic and image similarity retrieved from a database of 1.2 million images crawled from Flickr. The system effectively mines geographically relevant terms and ranks potential suggestion terms by their posterior probability given observed visual and geo-coordinate features.

A framework for probabilistically modelling geographical information using a GIS database for event and activity recognition in general-purpose consumer images, such as those obtained from Flickr is described in [5]. The proposed framework discriminatively models the statistical saliency of geo-tags in describing an activity or event. This work leverages the inherent patterns of association between events and their geographical venues, while using descriptions of small local neighbourhoods to form bags of geo-tags as our representation.

A very interesting work is described in [6], where smart-phones are used to monitor air quality from sensors, combined with audio and images. The developed applications provide a mapping between the captured media (audio and images), the sensor data with the relevant GPS and time information. A second application has been also developed to allow users interact with the content via a graphical map. Several implementation details are discussed. The Mobile Landscapes project [7] focuses on using handheld devices for urban analysis. Some of the measured quantities within this project include cell phones activity during the day, as well as the density of the cell phones in the urban area of Milan in Italy. Additionally, the paper discusses privacy issues and suggests a taxonomy of location-based services.

eGS [8] is a system that allows users to explore a city area while collaboratively visualizing a common atmospheric pollutant (carbon monoxide) in real-time. Using GPS and an attached sensor on a tablet-PC, a map shows pollutant values as a colour-coded trail as the user moves around the city. Users may take photographs of pollution-significant situations that are referenced against their current map location. The collaboration between the users is based on a piece of software called George Square [9].

GeoPix [10] is a mobile, and location-based application mobile using Google Maps, where users produce geo-referenced pictures that can be shared and accessed online. On-the-road users can use their camera phone equipped with GPS to capture images in a spot and to immediately share them on the web together with their physical location and timestamp, creating "geographic bookmarks".

Panoramio [11] (acquired by Google) is another very similar Web service using Google Maps and linking photos to their geographical location. Users can browse the Google Maps or search with tags. Confluence [12] is a project that provides pictures (and stories about the visits) for each of the latitude and longitude integer degree intersections in the world. The Web interface allows range queries where the user may provide latitude and longitude values and gets as result all the images and stories that fit to the query.

Placeopedia [13] is a web application based on Google Maps where markers representing Wikipedia articles are placed as overlays on the maps. The geo-referenced resources are articles and not photos and it is not aimed to the mobile user. Plazes [14] is a service that comes close to Twitter with geotagging, points of interest, and social networking features. GeoURL [15] is a location-to-URL reverse directory allowing finding URLs by their proximity to a given location.

2.3 CONTEXT EXTRACTION SYSTEMS

Context Watcher is part of the Mobilife Project [16] that aims to implement a system for automatically recording, storing, and using context information related to user's daily life. Context Watcher is able to collect information such as position, speed, body data, weather forecasts as well as content explicitly inserted by the user and make statistic on that data.

ContextCam [17] is a prototype of a consumer video camera that provides point of capture annotation of time, location, person presence and event information associated to recorded video. Both low- and high-level metadata are discovered via a variety of sensing and active tagging techniques, as well as through the application of machine learning techniques that use past annotations to suggest metadata for the current recordings. More particularly, the ContextCam acquires date and time information from an internal clock as well as from GPS satellites. In addition, ContextCam uses ultrasound to triangulate position of active tags around the camera. The ContextCam also captures who is near or around the camera by simply storing the RF pings.

WillCam [18] is a real-time annotation system where photographs are annotated at the capture time. WillCam enables users to capture various information such as location, temperature, ambient noise, and photographer facial expression, in addition to the photo itself. WillCam also helps users express their interest -what object or information in the picture/scene is most important for them- visually.

WillCam collects contextual information of three types.

- Author/photographer information: "who" takes a photograph. Most users are more interested in photographs taken by their families or friends than photographs by strangers.
- Situations: Information about how and where photographs are taken. Photographs taken at famous places or emblematic situations (e.g. "hot" summer, and "noisy" street) may attract users' attention.
- Targets: "what" object or situation in the scene portrayed in a photo is most important for the photographer.

The continuous recording and storing of the entire audio–visual lifetime of a person together with all digital information that the person perceives or creates is investigated in [19]. In such a scenario, it is challenging to find information because of the large amounts of data. In order to address this issue, three approaches of creating relevant meta-information are investigated, in particular: i) the user’s physical activity, e.g., sitting, standing, walking, or shaking hand, ii) from acceleration sensors, the social environment, e.g., in a restaurant or lecture, from audio, and iii) interruptibility from multiple sensors.

TagTansu [20] is an image capture and annotation system, which can easily capture pictures of clothes and add annotations to support creating a picture database of clothes. TagTansu consists of hook sensors and capture components attached inside the doors of a wardrobe. The hook sensors estimate the type and the weight of cloths. TagTansu generates annotations for the captured cloth images that include this additional information.

Memory Karaoke [21] is a mobile reminiscing tool, which targets the retelling of recent daily experiences by providing contextual cues such as location, time, and photos. Such a system can augment episodic memory exercises such as reminiscing and storytelling.

To compute an estimate of the current location on the phones without the need for extra hardware (such as a GPS unit), the system relies on a software-implemented place detection algorithm based on the PlaceLab initiative. This algorithm relies on recognizing nearby GSM towers as seen from the phone. Such location systems are now integrated in modern mobile systems (iOS, Android, etc)

2.4 SEARCHING BASED ON REAL WORLD DESCRIPTORS

A discussion about the issues involved in integrating Web search with a GIS database is provided in [22]. More specifically, the issues of extraction, knowledge discovery, and presentation are covered. A system has been implemented that obtains characteristic experiences from weblog entries. Such characteristic experiences can be extracted for different locations. For knowledge discovery, the cognitive significance of geographic objects according to content gathered from the Web is measured.

G-Portal [23] is a web portal that aims to provide digital library services over geospatial and georeferenced content found on the World Wide Web. G-Portal adopts a map-based user interface to visualize and manipulate the distributed geospatial and georeferenced content. Annotation capabilities are also supported, allowing users to contribute geospatial and georeferenced objects as well as their associated metadata.

World Explorer [24] is a visualization tool that can help expose the content of images, using a map interface to display derived tags as well as the original photo items. It has been applied on a large set of photos from Flickr. The selected photos are associated with user-defined unstructured text labels (i.e., tags). Tags are analyzed and associated with the geo-referenced Flickr images to generate aggregate knowledge in the form of “representative tags” for arbitrary areas in the world. Primary tags are defined to characterize geographical areas, which are shown with the visual tool over the map.

GATE [25] attempts to give spatial semantics to web pages by assigning them place names. More particularly, GATE is a named entity extraction software, which has been designed to perform place name extraction using a US Census gazetteer, proposing in parallel a rule-based place name disambiguation method and a place name assignment method capable of assigning place names to web page segments. The rule-based disambiguation method uses self-features, near context, perfect match, extraction patterns, and spatial distance to systematically determine a gazetteer place name to each ambiguous place name.

A search engine that automatically derives spatial context from unstructured Web resources and allows for location-based search is presented at [26]. The developed crawler applies heuristics to crawl and analyze Web pages that have a high probability of carrying a spatial relation to a certain region or place; the location extractor identifies the actual location information from the pages; the developed indexer assigns a geo-context to the pages and makes them available for a later spatial Web search.

SPIRIT [27] is an EU Project that has (among others) addressed the following issues:

- a multi-modal user interface providing textual input and interactive map feedback of the context of retrieved documents;

- machine learning techniques for the extraction of geographical context from web documents and for generating metadata providing spatial context
- spatial indices for web collections

The problem of collecting geographically-aware pages using collaborative crawling strategies is studied in [28]. A number of collaborative crawling strategies is discussed by considering features like URL address of page, content of page, or extended anchor text of links. It is concluded that the URL based and the extended anchor text based crawling strategies have the best overall performance.

2.5 DESCRIPTION FORMATS

There is a number of standards for description and exchange of location data on the Web. These range from simple coordinate-oriented ones specifying latitude and longitude such as vCard [29], Microformats or W3C Geo [30] to more powerful formats able to express additional concepts like lines, boxes, polygons etc. such as Dublin Core Metadata [31] the Geography Markup Language (GML) [32] or the KML format [37] of Google Earth. The description of a location is typically accomplished in two ways: specification of a globally unique coordinate tuple (i.e., longitude, latitude and optional height, usually in the WGS84 frame of reference) or a named hierarchical description. GeoWeb [34] is a distributed database of metadata in the form of DNS hierarchy of servers to finding data on the Web.

3. Generic Emotional, Social and Real-World Descriptors in I-SEARCH architecture

In this section, the role of generic emotional, social and real-world descriptors in the I-SEARCH architecture is presented according to the architecture. More specifically, the components that will involve emotional, social and real-world descriptors are identified in the layered architecture of the project and in the sequel their interaction with other components is briefly presented.

3.1 THE COMPONENTS IN THE LAYERED ARCHITECTURE OF I-SEARCH

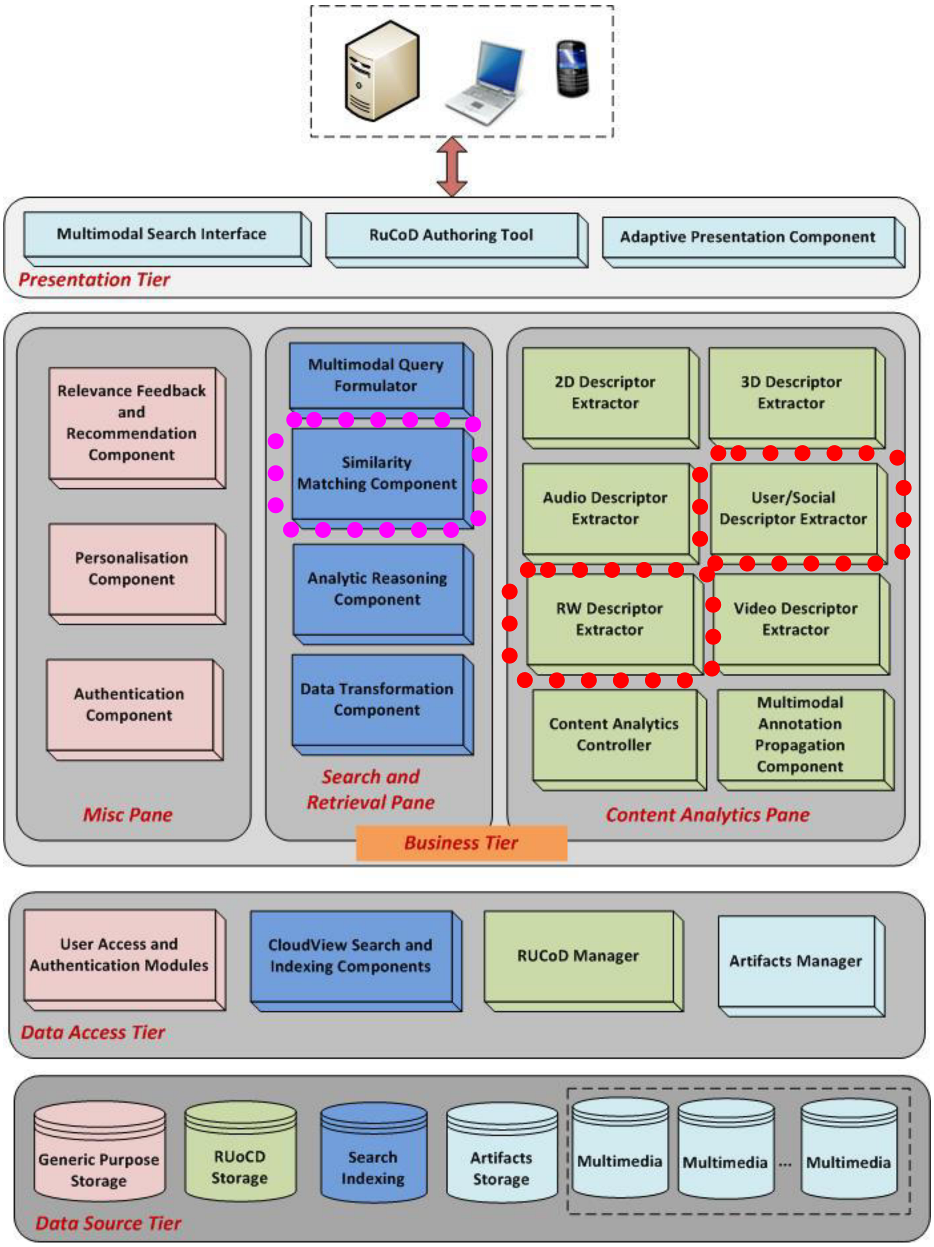

Figure 1: The User, Social and Real – World Features in the Layered Architecture of I-SEARCH

Figure 1 depicts the layered architecture of the I-SEARCH project, where all components are present, according to the I-SEARCH architecture. The emotional, social and real world descriptors are involved in three distinct architecture components, which are marked in red and pink in Figure 1:

- The Real World Descriptor Extraction Component
- The User Social Descriptor Extraction Component
- The Similarity Matching component

More specifically, the Real World Descriptor Extraction Component is responsible for the creation of the real-world part of the RUCoD record. In the same manner, the User Social Descriptor Extraction Component undertakes the extraction of the social and emotional descriptors and returns the social and emotional part of the RUCoD.

The similarity matching component is equipped with all appropriate matching methods that are required for low-level, real-world, emotional and social descriptors. This module is required during the search in order to perform the integrated similarity matching procedure.

3.2 INTERACTIONS WITH OTHER I-SEARCH COMPONENTS

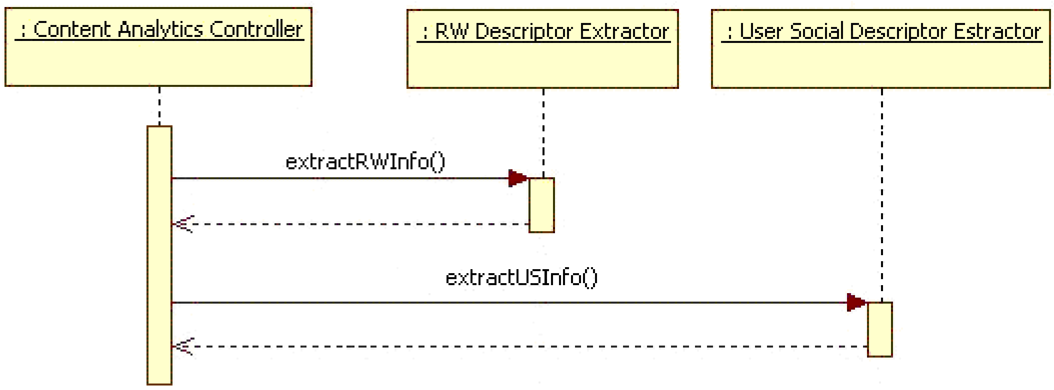

Figure 2: Interaction of Real-World and User Social Descriptor Extractor with the I-SEARCH platform

According to the I-SEARCH architecture, when a new content object arrives at the I-SEARCH platform, the Content analytics controller is triggered in order to produce the appropriate RUCoD record. A content object is composed by a set of multimedia files and associated text and metadata that represent the same object. The Content analytics controller requests from every descriptor extractor the relative RUCoD parts. Both real-world descriptor and user social descriptor extractors are interacting with the Content analytics controller (CAC), which orchestrates the RUCoD formulation, according to Figure 2. Finally, the RUCoD object of the Content object is formulated by CAC by combining all fragments received from the extractors and is forwarded to the RUCoD manager for indexing. Detailed interactions between the User-Social and Real-world Descriptor extractor components will be presented in D4.2; however the functionality of these extractors adopts the same concept. The functions extractRWInfo() and extractUSInfo() will get a Content Object as input and will return to the CAC the RUCoD fragment that includes the relative descriptors of the given Content Object.

4. Generic Emotional Descriptors

In this section, the conceptual model of the emotional descriptors is initially presented and then the existing two well-known emotional spaces (valence-arousal and energy-velocity) are presented.

4.1 Conceptual Model and Core framework

The conceptual model of the emotional descriptors is based on the assumption that the emotions conveyed to the user are built from physical stimuli that can be measured. Multiple modalities (music, video, computer animation) allow the transmission of expressiveness parameters from one domain to another – for example from music to computer animation, or from dance to music. This implies the development of a layered conceptual framework for affect processing that splits up the problem into different sub-problems.

The conceptual framework aims to clarify the possible links between physical properties of a particular modality and the affective/emotive/expressive meaning that is typically associated with these properties. At the lower level lies the syntactical layer, that represents the analysis and synthesis of physical properties; it is applied in different modalities (audio, movement and animation) and hence will analyze different low-level descriptors in order to propagate to the next level a semantic information. On top of the syntactical layer lies the semantic layer, containing descriptions of affects, emotions and expressiveness, in a modality-unaware way (i.e. disconnected from the modality used to compute them). At the top level lies the final affective/emotive/expressive meaning, defined as the mapping between the syntactical space and the labels that user assign to emotions (e.g., happy, sad, angry/aggressive, hesitant, meditative).

The current proposal of emotional descriptors at this stage is focusing on the valence-arousal circumplex model [45]. During the project, two directions will be considered for possible extension of the emotional descriptors: (i) the addition of a third dimension, such as Power (e.g. Wundt), as proposed by the literature in psychology; (ii) the inclusion of emerging emotion models based on non-utilitaristic emotions, such as the "aesthetic emotions" proposed by Scherer and Zentner for "music emotions" [61],[62].

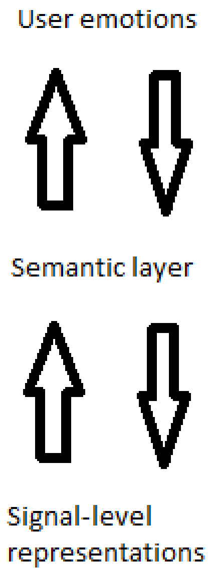

Figure 3: The layers of the conceptual framework of emotions

Transitions between these layers allow for cross-modal communication of emotions: for example, analyzing the movements of a dancer makes the transition from the syntactical layer to a point in the semantic space; but the same point in the semantic space corresponds to a set of syntactical features in a different domain, e.g. music. Thus, it is possible to manipulate in real time a music piece so that it matches the same emotional content of the dancer's movements, or generate an avatar displaying sadness in his posture when a sad music is played.

It is important that the RUCoD format can store descriptors from all three levels; descriptors coming from the physical layer are useful to allow mono-modal search (when both the query and the result come from the same modality), while descriptors from the semantic and affective layer allow higher level queries like "find a music matching the mood of my movements" (semantic layer) or "look for sad songs" (affective layer).

The RUCoD descriptors corresponding to the physical layer have been described in the deliverable for task 3.1, and the algorithms for their computation will be identified in task 4.1. The descriptors for the semantic layer are described in this section of this deliverable, and their algorithms will be identified in task 4.2.
The queries at the level of the affective layer can be made possible by re-using the values of the descriptors at the semantic layer, if the mapping between the two layers has the needed characteristics; in any case RUCoD will allow the user to manually specify the affective attributes using one of the most commonly used vocabularies, presented in this deliverable.

4.1.1 Minimal representations at the syntactic level: the Bounding Triangle

Expressive gesture analysis in I-SEARCH will consider a bounding triangle identified by the position of hands and head. Recent studies showed that this minimal representation of human upper-body movements provide sufficient information to a system to distinguish automatically between meaningful groups of emotions [63] related to the four quadrants of the valence/arousal space. Evaluation of emotion recognition performance based on this minimal representation further assessed that human observers could discriminate between high and low arousal emotions.

4.1.2 Entropy-based measures for a dynamic analysis of expressive behaviour

The analysis of features dynamics related to human behaviour requires specific computational tools: behavioural signals can be analyzed using traditional approaches based on time and frequency domains. However, such measures fail to account for central properties of human movement dynamics: (i) the non- linearity (small perturbations can cause large effects) and (ii) non-stationarity (the statistical properties change with time). Recent entropy-based measures have been developed to address this issue and to extract information from behavioural time series not contained in traditional methods based on mean and variance [60].

One of these measures, Sample Entropy, computes the conditional probability that sequences of behavioural signals similar for m points will remain similar when they are extended with one more measure point (m+1). High values of Sample Entropy indicate that the user introduced a change in his movements.

4.2 VALENCE-AROUSAL

The affective space is derived from the “circumplex model of affect” proposed by the psychologist Russell [45] and later expanded by Thayer [44], who organized the emotions in terms of pleasure-displeasure (valence) and arousal. By using this semantic space, the semantic layer can associate 4 basic emotions (happy/exuberant, angry/anxious, sad/depressed and calm/content) via a simple mapping with each quadrant [44].

The valence domain goes from “displeasure/tense” (left side) to “pleasure/calm” (right side); the arousal domain goes from “resting/tired” (bottom part) to “active/energetic” (top part). A third dimension is often noticed, but its meaning is less clearly specified, although it is often associated with the concept of power/dominance [67], [68].

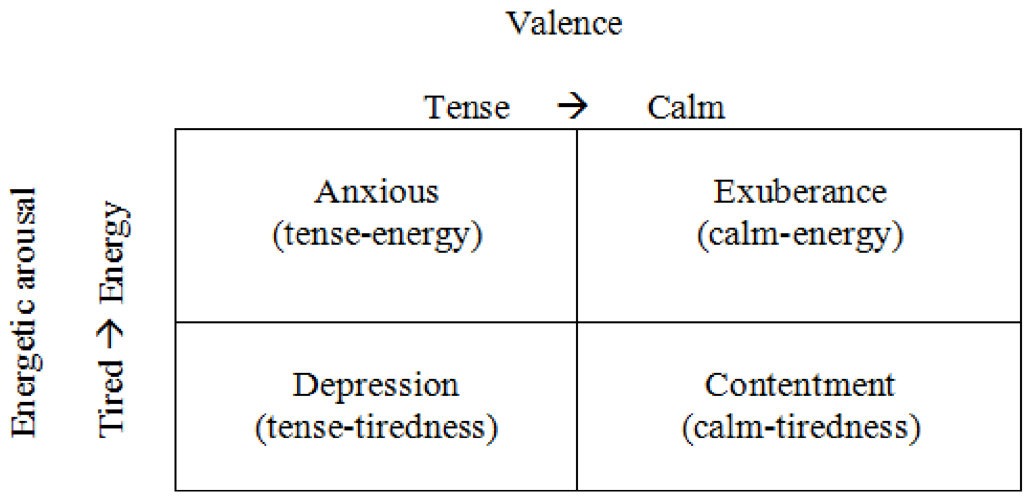

Figure 4: Valence-Arousal plane with the four mood clusters from Thayer's model

4.3 KINAESTHETIC SPACE OR ENERGY-VELOCITY

This space has been successfully used for the analysis and synthesis of musical performance [46] and is derived from factor analysis of perceptual evaluation of different expressive music performances. Listeners tend to use these coordinates as mid-level evaluation criteria.

The most evident correlation of energy-velocity dimensions with syntactical features is legato-staccato versus tempo. The robustness of this space is confirmed in the synthesis of different and varying expressive intentions in a musical performance by using a control based on timing and on the dynamics of the notes. The kinetics (velocity) domain goes from “light” to “heavy”, while the energy domain goes from “hard” to “soft”.

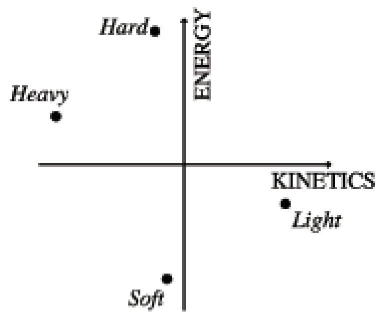

Figure 5: The Kinaesthetic space of Energy-Velocity

4.4 RELATED STANDARDS

In this subsection, standard W3C EmotionML which is related to emotion representation is presented

4.4.1 W3C EmotionML

The W3C has launched in 2002 the Multimodal Interaction Activity group to standardize XML dialects to support new scenarios allowed by the new technologies (e.g. interaction with speech, pen-based input, gestures):

- mobile applications
- wall mounted interactive displays
- home entertainment systems
- automotive telematics

The group is currently working on the final specification of the EmotionML [66] XML dialect, recently released in the Last Call phase of the approval procedure. The dialect will allow the embedding of emotion-related annotations in several other XML languages:

- EMMA (Extensible MultiModal Annotation) e.g. controlling the facial parameters of an avatar to express the requested emotion
- SSML (Speech Synthesis Markup Language) e.g. manipulating the synthetic voice used to read a text in order to express the requested emotion
- SMIL (Synchronized Multimedia Integration Language) e.g. to instruct the rendering agent about the emotion accompanying the media items

A separate effort focuses on the categorization of the human emotions ("Vocabularies for EmotionML").
The currently vocabularies defined in literature are:

1) Ekman's "big six" basic emotions
    a. Anger
    b. Disgust
    c. Fear
    d. Happiness
    e. Sadness
    f. Surprise
2) Cowie's everyday emotions
    a. Affectionate
    b. Afraid
    c. Amused
    d. Angry
    e. Bored
    f. Confident
    g. Content
    h. Disappointed
    i. Excited
    j. Happy
    k. Interested
    l. Loving
    m. Pleased
    n. Relaxed
    o. Sad
    p. Satisfied
    q. Worried
3) OCC (Orthny, Clore, Collins) categories
    a. Admiration
    b. Anger
    c. Disappointment
    d. Distress
    e. Fear
    f. Fears-confirmed
    g. Gloating
    h. Gratification
    i. Gratitude
    j. Happy-for
    k. Hate
    l. Hope
    m. Joy
    n. Love
    o. Pity
    p. Pride
    q. Relief
    r. Remorse
    s. Reproach
    t. Resentment

    u. Satisfaction
    v. Shame
4) FSRE (Fontaine, Scherer, Roesch, Ellsworth) categories
    a. Anger
    b. Anxiety
    c. Being hurt
    d. Compassion
    e. Contempt
    f. Contentment
    g. Despair
    h. Disappointment
    i. Disgust
    j. Fear
    k. Guilt
    l. Happiness
    m. Hate
    n. Interest
    o. Irritation
    p. Jealousy
    q. Joy
    r. Love
    s. Pleasure
    t. Pride
    u. Sadness
    v. Shame
    w. Stress
    x. Surprise
5) Frijda's categories
    a. Anger
    b. Arrogance
    c. Desire
    d. Disgust
    e. Enjoyment
    f. Fear
    g. Humility
    h. Indifference
    i. Interest
    j. Resignation
    k. Shick
    l. Surprise

An alternative approach is using a limited number of dimensions that define a space inside which the single emotions are defined as contiguous areas.

1) Mehrabian's PAD dimensions
    a. Pleasure
    b. Arousal
    c. Dominance
2) FSRE (Fontaine, Scherer, Roesch, Ellsworth) dimensions
    a. Valence
    b. Arousal

   c. Potency
   d. Unpredictability
3) The intensity dimension
   a. Intensity

5. Generic Social Descriptors

The social descriptors are utilized to capture the types of interaction between users; unlike the preceding section, where the task of the descriptor was to capture the emotions that either a media item or the movements of a user can convey to an observer, the social descriptors have the task of representing the dynamics of the whole group of users.
Such a group could be composed by well-synchronized members who show the same affective features described in the previous section, or it could have a minority part that shows different characteristics.
In this section, the conceptual model is presented and then the existing methods for describing social features are analyzed. More specifically, these descriptors are describing the entertainment, the empathy, the saliency, the leadership and the group geometric features.
In the I-SEARCH platform such social descriptors will be used for building the query rather than during the analysis of the Content Objects in the database; even if it is possible to place in the database media items where a group of users is performing (e.g. a group of singers singing *a cappella*, or a group of dancers in a video clip), and as such suitable for the extraction of these descriptor, the type of queries that would be allowed by them would be similar to “find me all video clips where the dancers are not well synchronized”.
We want to focus instead on the dynamic nature of the descriptor when computed on a real, non-professional group of users and use it to drive or tune the extraction of other descriptors that allow more meaningful queries.

5.1 Conceptual Model and Core framework

Social descriptors are used to represent the interaction between multiple users. As such they are focusing more on the users performing the query, rather than on the Content Object represented by the RUCoD instance.
When the social descriptors are computed by examining the users performing a query, they can either be stored as-is in the RUCoD representing the query or used to pre-process other low-level descriptors. For instance, a group of users could be allowed to perform a query only when they reach a minimum level of synchronization, and when this happens only the low-level descriptors extracted from the leader of the group are placed in the query RUCoD.

5.2 Entrainment and Empathy Descriptors

The term entrainment comes from the fields of natural sciences and medicine, and it is related to the co-ordination or synchronization (both intended and unintended) that occurs between human beings. Pikovsky et al. [47] define it as “an adjustment of rhythms of oscillating objects due to their weak interaction”. Entrainment and related phenomena can be studied focusing on different kinds of synchronisation (phase synchronisation, general synchronisation, complete synchronisation), on different experimental conditions (e.g., “passive” or “active” experiments), and on physical observables (e.g. physiological data, motor behaviour data, gesture, audio signals).

For instance, at a music concert where a string quartet is playing, a highly trained ensemble can reach high levels of gesture co-ordination and as such is able to achieve empathic behaviour with its audience, conveying them the desired emotion.

5.2.1 Phase Synchronization

The descriptor is computed starting from the measures of the behaviour of each user (either a raw stream of accelerometer data, or the computation of a higher level index such as Motion Index, Contraction Index, etc.) during a window of a few seconds. Starting from this vector, a Recurrence index is computed, showing how many times the state vector returns close to a previous value. After normalizing the index on a range [0,1], the autocorrelation between two vectors belonging to different users gives the probability that the state of the system recurs at a given time τ is provided by:

$$\hat{p}_x(\varepsilon,\tau) = RR_\tau(\varepsilon) = \frac{1}{N-\tau}\sum_{i=1}^{N-\tau}\Theta(\varepsilon - \|\vec{x}_i - \vec{x}_{i+\tau}\|)$$

Finally, the phase synchronization index is obtained as the correlation coefficient computed around the normalized generalized autocorrelation functions computed for the users.

5.3 Saliency Descriptors

Behaviour saliency can be modelled starting from recent results obtained in the field of *computational attention* [56][57]. In this framework, salient behaviour is understood as a behaviour capturing the attention of the observer. The saliency index can be computed on any low -level (e.g., speed) and mid-level (e.g., Motion Index) features and indifferently on *n*-number of participants. Motion features can be compared in the spatial context of the current video frame (e.g., one participant with respect to the others in a group) and analyzed on a varying sliding time window (e.g., participant's movement over time). The Saliency index is based on the so-called self-information. Let us note m_i a message containing an amount of information (e.g., a certain value of a motion index). This message is part of a message set M (e.g., the Motion index values over time). The saliency of m_i is defined by :

$$S(m_i) = -\log(p(m_i))$$

where $p(m_i)$ is the occurrence likelihood of the message m_i within the message set M.

5.4 Geometric group features

Geometric forms can be created to approximate group configurations. In some cases, it can consist in applying existing individual spatial cues (e.g., bounding rectangle or triangle) to a collective situation where the entire group is analyzed instead of a single individual. Usual measure can be performed (e.g., degree of contraction/expansion) to gain insight on the behaviour of the group behaviour, considered as a single organism. Generic geometric form can be created to adapt to more complex group situation (e.g., polygon relating heads of a string quartet musicians). Additional features can then be devised to adapt to the specificity of group studies. For example the stability of the Centre-of-Gravity or polygon area may inform on the group cohesion (e.g., participants are coordinating/compensating movement of the other participants in order to keep the polygon area most stable). Rigidity of formation is also a candidate feature to distinguish between form configuration (see work by Khan et al [69]).
These descriptors can be very significant in the analysis of the behaviour of a group, e.g. a group of dancers in the I-SEARCH use case.

(a) (b)

Figure 6: Examples of polygons connecting members of a group

5.5 LEADERSHIP DESCRIPTORS

Dominance is "the set of expressive, relationally based communicative acts by which power is exerted and influence achieved", by means of forcefulness, monopolizing or involvement.

Two strategies can be devised to measure dominance:

(i) by looking at how, starting from a state where all the members in a group are synchronized, a divergence occurs and then disappears; dominance is seen here as a feature related to the direction of synchronization

(ii) by analyzing how some member in the group succeed in controlling and regulating their own behaviour and the group activity; dominance is seen here as a feature related to behaviour complexity

5.5.1 Dominance as a feature related to the direction of synchronization

The direction of the synchronization reveals if the non-synchronized member of the group was a "master" that forced the other members to follow him in the new pattern, or a "slave" that wasn't fully following the leader. In the former case, a user acts as an external force driving the other/s to reach synchronization; in the latter case all the users combine to bring about synchronization and no driver can be detected.

To compute it starting from two vectors containing the measure of an affective descriptor, the number of recurrence occurring in a short window of time are counted

$$c^{\tau}(y|x) = \sum_{j=1}^{m_y} \sum_{i=0}^{m_x} J_{ji}^{\tau} \qquad c^{\tau}(x|y) = \sum_{i=1}^{m_x} \sum_{j=0}^{m_y} J_{ij}^{\tau}$$

where:

$$J_{ij}^{\tau} = \begin{cases} 1 & if\ 0 < t_i^x - t_j^y < \tau \\ \frac{1}{2} & if\ t_i^x = t_j^y \\ 0 & otherwise \end{cases} \qquad J_{ji}^{\tau} = \begin{cases} 1 & if\ 0 < t_j^y - t_i^x < \tau \\ \frac{1}{2} & if\ t_i^x = t_j^y \\ 0 & otherwise \end{cases}$$

The relative delay between the two series is then given by

$$q_\tau = \frac{c^\tau(y|x) - c^\tau(x|y)}{\sqrt{m_x m_y}}$$

where q_τ is between -1 and 1; the absolute value indicates how much leadership is strong, whereas the sign indicates who is the leader.

5.5.2 Dominance as a feature related to behaviour complexity

Analysis of dominance is here based on the theoretical framework of multi-scale entropy (MSE), a non-linear technique to quantify the behaviour *complexity*, i.e., the information expressed by the body movement dynamics over multiple time scales. Recent studies suggest that the dominant person is the one which behaviour complexity is:

(i) relatively low with respect to the complexity of others in the group
(ii) highly correlated with the group activity

The dominant person (*leader*) appears as the one able to "integrate" others' activity and to decrease the total entropy of the group.

Considering a time series, the computation of the *Complexity Index* (CI) comprises three distinct processes deriving from the Multi-Scale Entropy method:

(i) a coarse-graining procedure to represent the system's dynamics at different time scales
(ii) the quantification of the degree of irregularity of each coarse-grained time series through the application of Sample Entropy (SampEn)
(iii) A complexity index (CI) of the time series is calculated by integrating the SampEn values obtained for the different time scales.

Considering a group of 3 users, computation for the dominant person identification starts from the three vectors containing the kinematics measurement of the users (e.g., position) plus one vector containing the kinematics measures relative to the polygon approximating the group (e.g., perimeter, centre of gravity of the polygon).

The identification of the dominant user (*leader*) is given by:

(i) compute the Complexity Index for each vector
(ii) select the users displaying the lowest Complexity Index (*potential leader*)
(iii) compute the *Pearson's product moment correlation* of the Complexity Index of each potential leader with the Complexity Index of the polygon's centre of gravity
(iv) select the user displaying the highest correlation (*leader*)

6. Real-world Descriptors

Real-world descriptors are mainly utilized to describe information that is gathered by contextual factors during acquisition, such as location, date, time and environmental conditions. In this section, the real-world descriptors are appropriately divided in order to capture different features and the existing approaches for representing this information are discussed.

6.1 Motivation for using real-world data for retrieval

Real-world information refers to annotations of the multimedia using information that is gathered by contextual factors during acquisition. These data are location (accurate position or area), time and environmental descriptors (temperature, lighting conditions, etc). The use of advanced set of accompanying features allow for a more intelligent and user-centric approach towards data search, retrieval and presentation. The utilization of the real-world descriptors is expected to enrich and purify the retrieval results in a way that is closer to the user preferences (such as location, intentions, etc).

A variety of real-world descriptors can be used in order to assist the search and retrieval procedure. These descriptors can be classified in the following categories:

1. Location-related descriptors.
2. Time-related descriptors.
3. Environmental descriptors.
4. ID descriptors.

In the rest of this chapter, the various forms of the aforementioned descriptors are presented. Then, the descriptors that will be adopted by the I-SEARCH project are specified.

6.2 Location Descriptors

Location descriptors are well known and some forms are utilised in everyday life. The most common forms of the Location descriptors are:

- Absolute position using earth coordinates, i.e. in the form of longitude, latitude and altitude, or equivalently in the form of eastings, northings and heightings.
- Absolute civil addressing schema, used in everyday life for finding a location (e.g.: 22, Broadway Grove, York, UK).
- Geographic tags that represent single points, or large areas (such as “the Sagrada familia church” or “the city of London”).
- Relative position from a landmark (e.g. 35ft NE from Eiffel Tower).

Although all of the above location descriptors (except for GPS) are very common, comprehensive and easy to remember even by non-expert users, they can be highly inaccurate or ambiguous. Concerning the absolute civil addressing schema for example, a lot of places exist (especially uninhabited areas and small villages), where no addressing schema is officially defined; thus, it is impractical to use the absolute civil addressing in a retrieval project where uninhabited areas should be covered. Regarding the relative position schemes, their adoption encapsulates the utilisation of another absolute scheme to assist them. Additionally, the relevant position can potentially be highly inaccurate.

In contrast, the utilization of an absolute position using earth coordinates is user-unfriendly and very accurate. This oxymoron can be overcome by utilizing earth coordinates in the retrieval process in order to achieve high accuracy, while the user will be capable to use either the absolute position schema or a civil addressing schema. An automatic mechanism will be used (such as OpenStreet Maps API or Google Geolocation API) to translate a civil address in the form of earth coordinates.

Classical representations consider the multimedia objects as localized with a single point. However, this schema does not allow representing a factor we consider as important: coverage. For example, a documentary on London should not be represented as a single point at the city-hall of London but as a shape covering the city of London. Apart from the single point representation, the location descriptor that will be utilized in the I-SEARCH project should be capable to represent areas. Thus, the single point representation is extended in the 3D location Descriptor with circular support area, defined the central point and the radius of the area. Moreover, the Box 3D Location Descriptor is also defined in order to support rectangular areas.

The direction of the camera is also a useful piece of information. For example, a user at the same place may point his camera at the White House or at the Washington National Monument. Without furtherer complex image treatment, the location and the direction allow the search algorithm to infer the subject of the picture. The direction is composed of three values: heading (or yaw), tilt (or pitch) and roll. These values are obtained via embedded electronic compass and accelerometers.

Finally, the location of the user alone is not always relevant for searching though multimedia objects, but is easy to obtain via sensors. The location of the subject is more relevant but is not always available, and may be too subjective. Thus, we allow the location in the descriptor to be the location of the camera and/or the location of the main subject.

A crucial part of the system is the distance between two earth points. The distance that will be adopted utilizes only the longitude and the latitude information of two spherical points and provides the distance between the points ignoring the anaglyph. The distance is based on the “haversine formula”, which is an equation giving great-circle distances between two points on a sphere from their longitudes and latitudes. It is a special case of a more general formula in spherical trigonometry, the law of haversines, relating the sides and angles of spherical "triangles". The distance between two points $P_1(lon_1, lat_1)$ and $P_2(lon_2, lat_2)$ is computed as follows:

$$D = 2R\arcsin\left(\sqrt{h}\right)$$
$$h = haver\sin(lat_2 - lat_1) + \cos(lat_2)\cos(lat_1)haver\sin(lon_2 - lon_1)$$

where R is the radius of the sphere and

$$haver\sin(\theta) = \sin^2\frac{\theta}{2}$$

The earth will be approximated by a sphere of radius R = 6371 km.

6.3 TIME DESCRIPTORS

The term “Time Descriptor” can be probably distinguished from the “Date Descriptor”; however, both descriptors represent the same space; in fact time is subdivision of Date. Furthermore, the existing standards do not distinguish the date and the time and represent both with a single description. Thus, for the needs of the I-SEARCH project, this notation will be followed and both time and date will be represented in a unified descriptor. In the rest of this document, the term “Time Descriptor” refers to both time and date.

Time descriptor, in contrast to Location, is much stricter and does not easily allow ambiguities; however, from another point of view, a time descriptor allows the expression of a time-range with a single descriptor, e.g. the statement Thursday, 3rd March, 1983 describes a time-range of 24 hours, or the statement around 2 o’clock tomorrow describes an undefined period that includes 14:00 of Sunday 21st March 2011. Another issue is related to the different time-zones, daylight savings and the different calendars (such as Julian and Gregorian calendar) utilized by some countries. Nowadays, the vast majority of the countries follow the Gregorian calendar and thus, the issue concerning calendars is highly associated with historical facts, rather than current and future events (for example, the Russian Revolution has started in 24 Oct 1917 according to Julian calendar, which is 7 Nov 1917 according to Gregorian calendar). In order to provide a unified solution for the aforementioned problems, the selected time descriptor will follow the Gregorian calendar and the description will include information about the time-zone. Additionally, in cases of an expression of a time-range in the form of a single descriptor, the median of the range will be used.

Another significant feature of the time descriptor is that allows easy extraction of correlations between different multimedia, such as periodic phenomena and more intuitive high-level information, using appropriate methods over the time descriptors.

6.4 ENVIRONMENTAL DESCRIPTORS

Concerning the environmental descriptors, there are multiple independent factors that can be described. Most of them are highly related to the weather. These are:

- Temperature: The temperature can be described using one of the two most common different standards that are utilised in Physics, i.e. Celsius or Fahrenheit. Absolute relationships exist that transform any measurement from any format to any format. There are also other temperature scales, such as Kelvin and Rankine scales, which are used for scientific purposes, and Delisle, Newton, Réaumur and Rømer scales, which are historical temperature scales. None of them has been widely adopted by the end-users. In the I-SEARCH project, the Celsius Temperature descriptor will be used, due to its great acceptance; however, in order to be compatible to the users that are familiar with Fahrenheit temperature scale (i.e. USA), the GUI should provide the user with the option to query the system with Fahrenheit values, which will be automatically converted to Celsius.

- Wind: The wind can be described using different metrics, such as Beaufort or wind speed (in meters per second). Unfortunately Beaufort scale is empirical and, thus, every value in Beaufort scale is assigned to a range of wind speeds, e.g. 3 BF means wind with speed in range 3.4–5.4 m/s. However, the Beaufort number well known, very familiar to the users and can be acquired from weather forecast and historical resources; thus, it is selected as the real-world descriptor in the I-SEARCH project.

- Humidity: There are three different metrics for Humidity: the "*absolute humidity*" which represents the total mass of the water vapor in the air per cubic meter, the "*relative humidity"* which expressed in % representing the amount of water vapor in the air at a specific temperature compared to the maximum water vapor that the air is able to hold in that temperature and the *"specific humidity"* which represents the ratio of water vapor to air (including water vapor and dry air) in a particular mass. For the needs of our user scenarios, the *relative humidity* will be adopted mainly due to its acceptance and the easy of use by the users.

- Conditions: This represent the general conditions, like "Cloudy", "Clear" or "Raining". For international and automatic support, a subset of the METAR standard will be used. The METAR standard is used in aviation and used keys such as "RA" for Rain, "NSC" for no cloud. The subset that we will use is described in the next section.

- Lighting: This descriptor represents the lighting conditions, i.e. how light is the scene. This information can be described by various ways, such us textual descriptions that express the lighting conditions or a single real number in a predefined range that represents all the lighting range (from absolute dark to absolute light). Although lighting conditions depend on weather conditions (cloudy, clear etc), they are also affected by many other environmental parameters such as the time of the day, the area and temporarily events. Thus, lighting conditions will be included as a separate descriptor in the I-SEARCH architecture, following the numerical format in the range [0,1], where 0 represents the absence of any light.

- Environment type: This descriptor represents the type of the environment. This can be simplistic textual descriptions such as "Indoor environment" and "outdoor environment", or more complicated textual descriptors such as "kitchen", "sea", "basketball field" and "garden". This descriptor is highly user-oriented and, thus, contains highly personal subjectivity on the descriptions.

6.5 ID DESCRIPTORS

ID Descriptors are widely utilized in order to automatically identify the marked product, based on a unique ID that is also exists in a supporting database; the label, which is automatically captured, is the 'ID' of the product in the database.

The most common identification descriptor is the well-known barcode. The barcode utilizes the Universal Product Code (UPC) scheme which encodes 12 bits in an optical machine-readable representation. The barcode system is nowadays extended in the Radio-Frequency

IDentifier (RFID). The basic aim of the RFID is to replace the barcode in the future. The major difference between the RFID and barcode lies in the way that the "ID" is stored in the product: while the barcode is printed on the product, the RFID is stored in an electronic device and is automatically read by the "scanner" using electromagnetic waves in the RF. The data inside the RFID follow a format that is incompatible with UPC.

A major issue of identification descriptors lies in the fact that access to the backend database is required, in order to gather potentially useful information regarding the product. Nevertheless, the RFID by itself can provide generic information, such as the origin country of the product and (probably) its type.

6.6 Related standards (Spatio-temporal)

In this subsection, the existing standards which are related to real-world descriptors are presented. The examination of the related existing standards will provide to the project a valuable knowledge that will be exploited during the RUCoD format specification. The majority of the standards presented in this subsection are related to location and time descriptors

6.6.1 Dublin Core

The Dublin Core [1] Standard utilizes textual description for every field using a standardised vocabulary of fifteen properties. Additionally, some self-documenting data strings are used whenever possible. Some indicative real-world descriptors in the Dublin Core format are:

- 49d4.5N 123d6.45W 45m
- 49d44'32"N 123d06'21W 120ft
- 49 44 32 N 123 06 21 W 45 m

A very useful element of the Dublin Core Standard concerning the location related and date-related real-world information is the Coverage element, which describes the spatial and temporal characteristics of the object or resource; Coverage is the key element for supporting spatial or temporal range searching on document-like objects that are spatially referenced or time-referenced. Spatial information may be given in numeric form (e.g. degrees) or in text (e.g. "Paris, France"). Temporal information may also be given in numeric form or in text.
The Coverage element can be defined initially as:

- coverage.periodName
- coverage.placeName
- coverage.x
- coverage.y
- coverage.z
- coverage.t
- coverage.polygon
- coverage.line

Inheritance may be extended such that the x,y,z,t qualifiers reference beginning and ending points in space and time to "bound" a coverage:

- coverage.x.min
- coverage.x.max

A numeric grouping subelement can be added at the end to keep the correct x with the correct y:

- coverage.x.min.1
- coverage.x.max.1
- coverage.y.min.1
- coverage.y.max.1
- coverage.t.min.1
- coverage.t.max.1
- coverage.x.min.2
- coverage.x.max.2
- coverage.y.min.2
- coverage.y.max.2
- coverage.t.min.2
- coverage.t.max.2

where groupings 1 and 2 describe two distinct bounding rectangles of coverage that may be discontinuous.

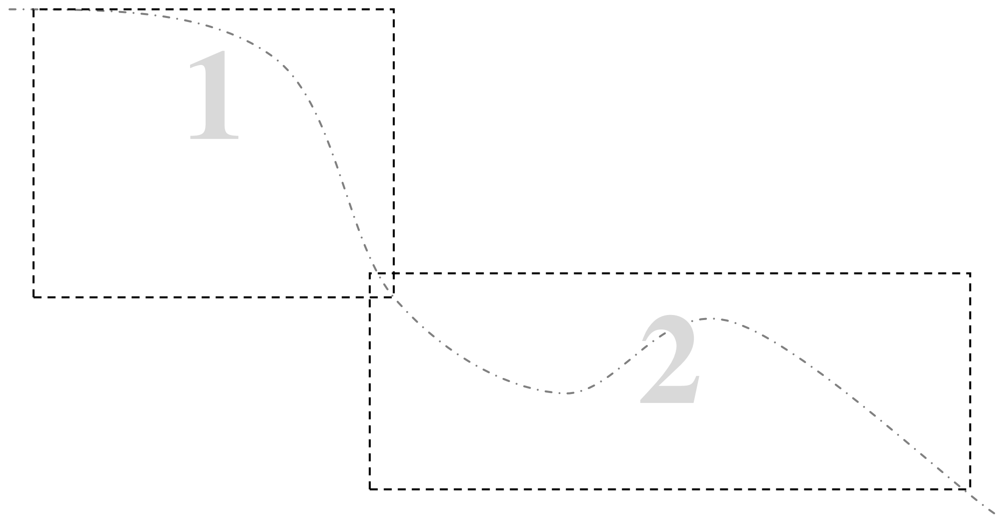

Figure 7 Schema to illustrate spatio-temporal represenation of Dublin-Core

6.6.2 GML

The Geography Markup Language (GML) [36] is a standard to express geographical features. It is an XML grammar defined by the Open Geospatial Consortium (OGC). This standard is well known and widely used.

The main use of the GML standard is to define shapes (e.g. points, lines, circles, polygons). These shapes can then be used to describe the location of things. For example, points can define the location of photos:

```
<PhotoCollection xmlns="http://www.myphotos.org" xmlns:gml="http://www.opengis.net/gml"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.myphotos.org
      MyGoodPhotos.xsd">
     <items>
         <Item>
```

```
            <name>Lynn Valley</name>
            <description>A shot of the falls from the suspension bridge</description>
            <where>North Vancouver</where>
            <position>
                <gml:Point srsDimension="2" srsName="urn:ogc:def:crs:EPSG:6.6:4326">
                       <gml:pos>49.40 -123.26</gml:pos>
                 </gml:Point>
             </position>
        </Item>
    </items>
</PhotoCollection>
```

More complex shapes can be defined to represent the coverage or the precision of the location. For example, the GEOPRIV [35] standard use circles or sphere to define the location obtained via a GPS, and wedges for the location of a user obtained via a directional antenna.

6.6.3 GMLJP2

The primary role of GMLJP2 [33] is to transform a JPEG 2000-image [39] into a usable geographic image. The latter is accomplished by providing the description of the image geometry and radiometry in the GMLJP2. Additionally, the GMLJP2 data can be used for the encoding of coverage metadata, geographic features, annotations, annotation and feature styling, and supporting components for coordinate reference systems and unit of measure definitions. This format is dedicated to aerial imagery.

6.6.4 KML

KML [37] is an XML format used to describe geographic properties. KML is an open standard officially named the KML Encoding Standard (OGC KML) and is maintained by the Open Geospatial Consortium, Inc. (OGC).

KML format complements the GML standard and follows almost the same grammar with it. KML has been adopted and made popular by Google. In a KML file, the real-world information is described in the <Camera> element, where location information is stored. Table 1 presents detailed information about the child-elements of the <Camera> element.

Attribute	Description
<longitude>	Longitude of the virtual camera (eye point). Angular distance in degrees, relative to the Prime Meridian. Values west of the Meridian range from −180 to 0 degrees. Values east of the Meridian range from 0 to 180 degrees.
<latitude>	Latitude of the virtual camera. Degrees north or south of the Equator (0 degrees). Values range from −90 degrees to 90 degrees.
<altitude>	Distance of the camera from the earth's surface, in meters. Interpreted according to the Camera's <altitudeMode> or <gx:altitudeMode>.
<heading>	Direction (azimuth) of the camera, in degrees. Default=0 (true North). (See diagram.) Values range from 0 to 360 degrees.
<tilt>	Rotation, in degrees, of the camera around the X axis. A value of 0 indicates that the view is aimed straight down toward the earth (the most common case). A value for 90 for <tilt> indicates that the view is aimed toward the horizon. Values greater than 90 indicate that the view is pointed up into the sky. Values for <tilt> are clamped at +180 degrees.
<roll>	Rotation, in degrees, of the camera around the Z axis. Values range from −180 to +180 degrees.

Table 1: <Camera> child-elements in KML standard [37]

The <tilt> and <heading> attributes can be relevant for picture retrieval. The <roll> attribute does not seem to be relevant.

Another, possibly useful KML element is the <PhotoOverlay> item, which is utilised to describe the position of views as big 2D screens (plan, cylinder or sphere) in the real-world. Table 2 presents detailed information about the child-elements of the <PhotoOverlay> element

Attribute	Description
<rotation>	Adjusts how the photo is placed inside the field of view. This element is useful if your photo has been rotated and deviates slightly from a desired horizontal view.
<ViewVolume>	Defines how much of the current scene is visible. Specifying the field of view is analogous to specifying the lens opening in a physical camera. A small field of view, like a telephoto lens, focuses on a small part of the scene. A large field of view, like a wide-angle lens, focuses on a large part of the scene.
<leftFov>	Angle, in degrees, between the camera's viewing direction and the left side of the view volume.
<rightFov>	Angle, in degrees, between the camera's viewing direction and the right side of the view volume.
<bottomFov>	Angle, in degrees, between the camera's viewing direction and the bottom side of the view volume.
<topFov>	Angle, in degrees, between the camera's viewing direction and the top side of the view volume.
<near>	Measurement in meters along the viewing direction from the camera viewpoint to the PhotoOverlay shape.
<shape>	The PhotoOverlay is projected onto the <shape>. The <shape> can be one of the following: • rectangle (default), for an ordinary photo • cylinder, for panoramas, which can be either partial or full cylinders • sphere, for spherical panoramas

Table 2: < PhotoOverlay > child-elements in KML standard

The distance (<near>) between the camera and the photo is an artificial attribute, considering the view as a 2D screen. The Field-of-Vision (<xxxxFov>) is an interesting piece of information; it can be used to determine what may be seen by the camera.

6.6.5 Exif

The JPEG format can embed metadata under EXIF standard. Most of the publicly available cameras add some of the EXIF metadata. Table 3 describes all the metadata that are supported by the EXIF standard.

Description	Ref.	Type	Size
GPS tag version	GPSVersionID	BYTE	4
North or South Latitude	GPSLatitudeRef	ASCII	2
Latitude	GPSLatitude	RATIONAL	3
East or West Longitude	GPSLongitudeRef	ASCII	2
Longitude	GPSLongitude	RATIONAL	3
Altitude reference	GPSAltitudeRef	BYTE	1
Altitude	GPSAltitude	RATIONAL	1
GPS time (atomic clock)	GPSTimeStamp	RATIONAL	3
GPS satellites used for measurement	GPSSatellites	ASCII	Any
GPS receiver status	GPSStatus	ASCII	2
GPS measurement mode	GPSMeasureMode	ASCII	2
Measurement precision	GPSDOP	RATIONAL	1
Speed unit	GPSSpeedRef	ASCII	2
Speed of GPS receiver	GPSSpeed	RATIONAL	1
Reference for direction of movement	GPSTrackRef	ASCII	2
Direction of movement	GPSTrack	RATIONAL	1
Reference for direction of image	GPSImgDirectionRef	ASCII	2
Direction of image	GPSImgDirection	RATIONAL	1
Geodetic survey data used	GPSMapDatum	ASCII	Any
Reference for latitude of destination	GPSDestLatitudeRef	ASCII	2
Latitude of destination	GPSDestLatitude	RATIONAL	3
Reference for longitude of destination	GPSDestLongitudeRef	ASCII	2
Longitude of destination	GPSDestLongitude	RATIONAL	3
Reference for bearing of destination	GPSDestBearingRef	ASCII	2
Bearing of destination	GPSDestBearing	RATIONAL	1
Reference for distance to destination	GPSDestDistanceRef	ASCII	2
Distance to destination	GPSDestDistance	RATIONAL	1
Name of GPS processing method	GPSProcessingMethod	UNDEFINED	Any
Name of GPS area	GPSAreaInformation	UNDEFINED	Any
GPS date	GPSDateStamp	ASCII	11
GPS differential correction	GPSDifferential	SHORT	1

Table 3: Supported Metadata by the EXIF Standard

However, having only the current position and the destination may lead to inaccurate data retrieval. These metadata may also be useless to describe a stored media.

6.6.6 geoURI

The recent RFC 5870 *A Uniform Resource Identifier for Geographic Locations ('geo' URI)* is a standard to define a link to a georeferenced point. Example: “geo:22.300;-118.44”. The metadata defined through this standard are: latitude, longitude, altitude, uncertainty, and coordinate reference system. This format is still young and poor compared to other standards.

6.6.7 ARML

Most Augmented Reality browsers display points of interest (POIs) around your current location. Each point is characterized by several attributes like geo-coordinates (longitude, latitude, altitude), name, description, images, address and other information. ARML is an initiative to standardize the way points of interest are described.[41]
The ARML specification is based on KML and extends it with AR specific structure.

6.6.8 MPQF

MPQF: “Provides a standardized interface for multimedia content retrieval systems (e.g.: MPEG-7 databases)”. It still belongs to ISO/IEC 15938 for historical reasons but is totally independent from MPEG-7. Concerning the real-world metadata, the next sample illustrates the usage of a SpatialQuery which allows a certain spatial relation to be used as a condition. Let us assume that a multimedia service provides a large set of images showing different animals. A user might be interested in finding all images that show a cat and a dog and in which the cat is on the dog’s left side. This example assumes that the multimedia service is capable of extracting the necessary information (object recognition, low- and/or high level features, etc.) in order to detect the animals.

Code 6 Sample Spatial Query request

```
<MpegQuery>
 <Query>
  <Input>
   <QFDeclaration>
    <Resource xsi:type="MediaResourceType"
            resourceID="animalA">
     <MediaResource>
      <MediaUri>http://db.mpqf.mpeg/cat.jpg</MediaUri>
     </MediaResource>
    </Resource>
    <Resource xsi:type="MediaResourceType"
            resouceID="animalB">
     <MediaResource>
      <MediaUri>http://db.mpqf.mpeg/dog.jpg</MediaUri>
     </MediaResource>
    </Resource>
   </QFDeclaration>
   <OutputDescription mediaResourceUse="true"/>
    <TargetMediaType>image/jpeg</TargetMediaType>
    <Condition xsi:type="SpatialQuery">
     <SpatialRelation
        sourceResource="animalA"
        targetResource="animalB" relationType=
        "urn:mpeg:mpqf:cs:SpatialRelationCS:2008:left"/>
    </Condition>
   </QueryCondition>
  </Input>
 </Query>
</MpegQuery>
```

Figure 8: A query formulated under MPQF standard

6.6.9 MXF

Material eXchange Format (MXF) is a container format for professional digital video and audio media defined by a set of SMPTE standards. This format is exhaustive and includes a plethora of real-word location-related descriptions. Additionally, it has not fully implemented yet. The detailed description of the standard is out of the scope of this document. The curious reader is referred to the MXF documentation [43].

6.6.10 IPTC

IPTC introduced a standard to annotate pictures. The main difference between IPTC and EXIF is that the latter is designed to contain automatic metadata while the former is designed to contain manual metadata.

Attribute	Length (characters)
City	32
Province State	32
Country	64

Table 4: IPTC metadata related to physical location

These locations are free text. This may lead to issues such as misspelling. Also, these attributes are language-dependent (eg. London/Londres). This information may be retrieved from the GPS position and a reverse geocoding database. Table 4 presents the metadata included in the IPTC that are related to physical location.

6.6.11 GeoWeb

GeoWeb is a distributed database of location metadata in the form of DNS hierarchy of servers to finding data on the Web.

6.7 RELATED STANDARDS (WEATHER)

Very few global initiatives exist to define a standard to represent weather. WheatherML is a standardization initiative that started in 2000 but that has been aborted. Nowadays, each weather service uses its own format, without any global standard.
The only global standard that can be easily parsed is the METAR format, used for the aviation.

6.7.1 METAR

METAR is a weather-related standard used in aviation [70]. It describes the hourly surface weather observations using the METAR international standard code format. The standard utilized a complex text-based dictionary, most of which is rather very discriminative considering the needs of the I-SEARCH project. An indicative subset of METAR standard that could be potentially exploitable by the project is presented in Table 5.

Key	Description
RA	Rain
SN	Snow
GR	Grêle
DZ	Drizzle
PL	Ice pellets
GS	Grésil
SG	Snow grains
IC	Ice crystals
UP	Unknown

Table 5: Indicative subset of METAR standard

7. Specification of Social, Emotional and Real World Descriptors

In this section, the social, emotional and real world descriptors that will be used in the I-SEARCH platform are formally specified based on the discussion of the previous chapter.

7.1 Generic Emotional Descriptors Specification

The emotional descriptors that will be utilised in the I-SEARCH platform are the valence and the arousal descriptors, specified in Table 6 and Table 7 respectively, along with the average valence-arousal Descriptor defined in Table 8.

7.1.1 Valence-Arousal Descriptor specification

Layer-1 Type	U-Descriptor	
Layer-2 Type	Emotion	
Title	Valence	
Acronym/Short Name	VAL	
Overview	An indication of whether the media item could be labelled as "positive" (high values) or "negative" (low values)	
Details	**Number of Frames**	N
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	N * 4	
Matching method	No matching expected; the sequence of values can be used for future extensions	

Table 6: Valence Descriptor Specification

Layer-1 Type	U-Descriptor	
Layer-2 Type	Emotion	
Title	Arousal	
Acronym/Short Name	ARO	
Overview	An indication of whether the media item could be labeled as "energetic" (high values) or "calm" (low values)	
Details	**Number of Frames**	N
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	N * 4	
Matching method	No matching expected; the sequence of values can be used for future extensions	

Table 7: Arousal Descriptor Specification

Layer-1 Type	U-Descriptor
Layer-2 Type	Emotion
Title	AvgValenceArousal
Acronym/Short Name	AVA

Overview	The position in the Valence-Arousal plane: as seen in section 4.2, the 2-D coordinate can also be used to express the indication of whether the entire media item could be labeled as happy, sad, calm or angry	
Details	**Number of Frames**	1
	Descriptor Dimension	2
Type of descriptor	Real Number (4 bytes).	
Total bytes required	8	
Matching method	Euclidean distance	

Table 8: Average Valence-Arousal Descriptor Specification

7.1.1.1 Valence-Arousal Descriptors in Use Cases

Table 9 provides the association of the selected descriptors for valence and arousal in the use-cases:

Use Case	Type of user search
UC1: Music retrieval	Emotional features of the gestures of the user are matched against the emotional features of the audio files in the I-SEARCH database
UC2: Social retrieval of music	Emotional features of the gestures of the leader of the dancers are matched against the emotional features of the audio files in the I-SEARCH database
UC3: Search for Furniture	N/A
UC4: Search for Multimedia objects	N/A
UC5: Search for Product Model	N/A
UC6: Search for 3D model	N/A
UC7: Search for Avatar	N/A

Table 9: Relevance of Valence-Arousal descriptor to I-SEARCH use cases

By adopting the valence-arousal descriptor, the I-SEARCH platform can compare the emotional features of music pieces with the emotional features of the gestures of the users.
It would also be possible to include in the search other modalities (e.g. video) by defining how low-level descriptors of those modalities are to be mapped to the valence-arousal perceptual dimensions.

7.1.2 Kinaesthetic Descriptor specification

The kinaesthetic descriptors that will be utilised in the I-SEARCH platform are the kinetics and the energy descriptors, specified in Table 10 and Table 11 respectively, along with the average kinetics - energy descriptor defined in Table 12.

Layer-1 Type	U-Descriptor	
Layer-2 Type	Emotion	
Title	Kinetics	
Acronym/Short Name	KIN	
Overview	An indication of whether the media item could be labeled as "light" (high values) or "heavy" (low values)	
Details	**Number of Frames**	N
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	N * 4	
Matching method	No matching expected; the sequence of values can be used for future extensions	

Table 10: Kinetics Descriptor Specification

Layer-1 Type	U-Descriptor	
Layer-2 Type	Emotion	
Title	Energy	
Acronym/Short Name	ENE	
Overview	An indication of whether the media item could be labeled as "hard" (high values) or "soft" (low values)	
Details	**Number of Frames**	N
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	N * 4	
Matching method	No matching expected; the sequence of values can be used for future extensions	

Table 11: Energy Descriptor Specification

Layer-1 Type	U-Descriptor	
Layer-2 Type	Emotion	
Title	AvgKineticVelocity	
Acronym/Short Name	AKV	
Overview	The average values of the Kinetics/Energy values	
Details	**Number of Frames**	1
	Descriptor Dimension	2
Type of descriptor	Real Number (4 bytes).	
Total bytes required	8	
Matching method	Euclidean distance	

Table 12: Average Kinetics-Energy Descriptor Specification

Currently it is not forecasted to use the aforementioned Kinaesthetic descriptors in the use cases of the I-SEARCH platform; however, in case where the valence-arousal space is not performing well during the system evaluation, the Kinaesthetic descriptors will be examined as alternative.

7.2 GENERIC SOCIAL DESCRIPTORS SPECIFICATION

In this section the social descriptors that will be adopted in the I-SEARCH platform are formally defined. These descriptors are related to the synchronization, the saliency, the geometric group features and the leadership. It should be noted that the social descriptors are utilized only during the querying phase on the client side of the platform, in order to capture accurate queries.

7.2.1 Entrainment and Empathy Descriptors Specification

Table 13 provides the specification of the synchronization descriptor.

Layer-1 Type	U-Descriptor	
Layer-2 Type	User	
Title	Synchronization	
Acronym/Short Name	SYNC	
Overview	An indication of whether the user are synchronized or not	
Details	**Number of Frames**	1
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	4	
Matching method	No matching expected; the value will be used on the client side of the platform	

Table 13: Synchronization Descriptor Specification

7.2.2 Saliency Descriptors Specification

The Saliency Descriptor is formally specified in Table 14 .

Layer-1 Type	U-Descriptor	
Layer-2 Type	User	
Title	Saliency	
Acronym/Short Name	SAL	
Overview	An indication of whether the user movements are capable of capturing the attention of an observer	
Details	**Number of Frames**	1
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	4	
Matching method	No matching expected; the value will be used on the client side of the platform	

Table 14: Saliency Descriptor Specification

7.2.3 Geometric group features Specification

In Table 15 and Table 16 the specification of Group Stability and Group Rigidity Descriptors are provided respectively.

Layer-1 Type	U-Descriptor	
Layer-2 Type	User	
Title	GroupStability	
Acronym/Short Name	GST	
Overview	An indication of whether the movements of the users are coordinated so that the group shape remains stable	
Details	**Number of Frames**	1
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	4	
Matching method	No matching expected; the value will be used on the client side of the platform	

Table 15: Group Stability Descriptor Specification

Layer-1 Type	U-Descriptor	
Layer-2 Type	User	
Title	GroupRigidity	
Acronym/Short Name	GRG	
Overview	An indication of whether the group is organized with rigid segments	
Details	**Number of Frames**	1
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	4	
Matching method	No matching expected; the value will be used on the client side of the platform	

Table 16: Group Rigidity Descriptor Specification

7.2.4 Leadership Descriptors Specification

The Leadership Descriptor is formally specified in Table 17.

Layer-1 Type	U-Descriptor	
Layer-2 Type	User	
Title	Leadership	
Acronym/Short Name	LEAD	
Overview	An indication of whether the group is following a leader	
Details	**Number of Frames**	1
	Descriptor Dimension	1
Type of descriptor	Real Number (4 bytes).	
Total bytes required	4	
Matching method	No matching expected; the value will be used on the client side of the platform	

Table 17: Leadership Descriptor Specification

7.2.5 Application of social descriptors to use cases

Table 18 provides the association of the social descriptors in the use-cases

Use Case	Type of user search
UC1: Music retrieval	N/A
UC2: Social retrieval of music	A group of persons collaborate in constructing the query to identify a music piece that matches the preferences of the majority of the group components
UC3: Search for Furniture	N/A
UC4: Search for Multimedia objects	N/A
UC5: Search for Product Model	N/A
UC6: Search for 3D model	N/A
UC7: Search for Avatar	N/A

Table 18: Relevance of social descriptors to I-SEARCH use cases

In UC2, the following descriptors are useful when the I-SEARCH platform is observing how each user is interacting with the rest of the group:
- Phase Synchronization to assign a dancer to a coherent group of dancers
- Dominance to identify the leader of the coherent group of dancers

Once they have been computed on the client side, the emotional descriptors of the leader are sent to the I-SEARCH database as the actual query.

7.3 REAL WORLD DESCRIPTORS SPECIFICATION

7.3.1 Location Descriptor specification

The real-world Descriptor concerning Location will be expressed in longitude, latitude and altitude in the RUCoD file. Table 19 describes the specification of the Location Descriptor.

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Location Descriptor	
Title	3D Location Descriptor	
Acronym/Short Name	3DLD	
Overview	Consists of the altitude, longitude and latitude.	
Details	**Descriptor Dimension**	3
	Accuracy	10-100 meters
Type of descriptor	Three real numbers	
Total bytes required	24 bytes	
Matching method	Calculate the space difference in the three dimensions.	

Table 19: Location Description Specification

The above Location Descriptor describes a single position; however, in some cases, the user does not remember the specific location and defines an area as the location. Also, the sensors are not perfect and due to their limitation, the accuracy may be reduced. Finally, the media object can have a coverage (ex. City of London). In order to cover these cases, the Location Descriptor is extended in order to support areas instead of single point. Two alternative solutions will be used:

- The Box 3D Location Descriptor which consists of two sets of 3DLD defining the two corner points of the box (Table 20)
- The 3D Location Descriptor with Circle support area which consists of a 3DLD enhanced the radius of the supporting circle (Table 21)

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Location Descriptor	
Title	Box 3D Location Descriptor	
Acronym/Short Name	3DLD-B	
Overview	Consists of two sets (altitude, longitude and latitude)	
Details	**Descriptor Dimension**	6
	Accuracy	Same as 3DLD
Type of descriptor	Six real numbers	
Total bytes required	48 bytes	
Matching method	Calculate the space difference in the three dimensions.	

Table 20: Location (Area) Description Specification

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Location Descriptor	
Title	3D Location Descriptor with Circle support area	
Acronym/Short Name	3DLD-C	
Overview	Consists of the altitude, longitude and latitude and the radius.	
Details	**Descriptor Dimension**	4
	Accuracy	Same as 3DLD
Type of descriptor	Real numbers	
Total bytes required	4 * 8 bytes	
Matching method	Calculate the overlap of the circles.	

Table 21: Location (Area) Description Specification

Finally, Table 22 describes the specification of the Direction Descriptor, which provides the viewpoint of the camera.

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Location Descriptor	
Title	Direction	
Acronym/Short Name	LDD	
Overview	Consists of the heading (0° to +360°), the tilt (-90 to +90°) and roll (-180 to +180°)	
Details	**Descriptor Dimension**	3
	Accuracy	5°
Type of descriptor	Integers	
Total bytes required	3 * 8 bytes	
Matching method	Compare the vectors	

Table 22: Direction Description Specification

An alternative descriptor that will be used in I-SEARCH platform is the Route Location Descriptor. This descriptor represents a complete route, instead of a single point or location, and can be captured automatically by appropriate devices.

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Route Location Descriptor	
Title	Direction	
Acronym/Short Name	RLD	
Overview	Consists of N ordered points of (altitude, longitude and latitude)	
Details	**Descriptor Dimension**	3*N
	Accuracy	Same as 3DLD
Type of descriptor	Real Numbers	
Total bytes required	N*3 * 8 bytes	
Matching method	Find the closest points between two routes.	

7.3.2 Location Descriptors in Use Cases

Table 23 provides the relevance of the selected location descriptors to the use-cases.

Use Case	Type of user search
UC1: Music retrieval	The user requires a display, clustering the results by geographical proximity
UC2: Social retrieval of music	N/A
UC3: Search for Furniture	The end-user requests the dealers that are closer to a specific location. The dealer searches for manufacturers closer to her location
UC4: Search for Multimedia objects	The user looks for an article available within a radius defined by the user.
UC5: Search for Product Model	The end-user requests the dealers that are closer to a specific location.
UC6: Search for 3D model	In the extension of the use case, the designer can filter the results according to her current position
UC7: Search for Avatar	N/A

Table 23: Relevance of location descriptors to the use-cases.

The 3D location descriptor and its extensions are perfectly fit in the use cases that require a location real-world descriptor. More specifically, 3D Location Descriptor will be utilized to represent the position of the user (UC3, UC4, UC5), while the 3D Location Descriptor with Circle support area will be used to represent the dealers and their supporting region. All Location Descriptor will be exploited during clustering of songs by geographical proximity.

7.3.3 Time Descriptors specification

Concerning the description of the time, a variety of methods can be found across different countries (e.g. DD-MM-YYYY HH:MM:SS, or MM-DD-YY HH:MM:SS). The main objective is the adoption of a representation that allows easy comparison using full or partial information, such as only years, months, or specific time of day. Thus, the format of W3C (YYYY-MM-DD hh:mm:ss.ss TZ) has been selected for storing the real-world Time Descriptor, where YYYY stands for the year, MM for the month, DD for the day, hh for the hour (in 24 hour format), mm for the minute and ss.ss for the seconds in decimal form. TZ represents the time-zone. Table 24 describes the specification of the real-world Time Descriptor according to the W3C standard [42].

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Time Descriptor	
Title	W3C Time Descriptor	
Acronym/Short Name	W3CT	
Overview	Consists of a string that follows the W3C format for dates and time.	
Details	**Descriptor Dimension**	1
	Accuracy	10 ms
Type of descriptor	String	
Total bytes required	~30 bytes	
Matching method	Calculate the time difference between two time instances.	

Table 24: Specification of Time Descriptor according to W3C

7.3.4 Time Descriptors in Use Cases

Table 25 provides the relevance of the time descriptors to the I-SEARCH use-cases.

Use Case	Type of user search
UC1: Music retrieval	N/A
UC2: Social retrieval of music	N/A
UC3: Search for Furniture	The end-user requests products manufactured after a specific date.
UC4: Search for Multimedia objects	The user search for a content object that is not older than 10 years.
UC5: Search for Product Model	N/A
UC6: Search for 3D model	N/A
UC7: Search for Avatar	N/A

Table 25: Time descriptor in use cases

The adoption of the W3C format allows comparison and supports fully the time-enhanced queries defined in UC3 and UC4.

7.3.5 Environmental Descriptors specification

Table 26 summarizes the specification of the Celsius Temperature Descriptor, Table 27 depicts the specification of Wind Descriptor and Table 28 provides the specification of Humidity descriptor. In the sequel, in Table 29 the specification of Conditions descriptor is provided and in Table 30 the adopted subset of METAR standard. Finally, Table 31 provides the specification of Lighting Descriptor.

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Temperature	
Title	Celsius Temperature Descriptor	
Acronym/Short Name	CTD	
Overview	Consists of a real number.	
Details	**Descriptor Dimension**	1
	Accuracy	degrees
Type of descriptor	Real number greater than -273	
Total bytes required	4 bytes	
Matching method	Calculate the difference between two temperatures.	

Table 26: Specification of Temperature Descriptor using Celsius

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Wind	
Title	Beaufort Number Wind Descriptor	
Acronym/Short Name	BNWD	
Overview	Consists of an integer showing the Beaufort number of the wind.	
Details	**Descriptor Dimension**	1
	Accuracy	1 BN
Type of descriptor	Integer between 0 and 12	
Total bytes required	1 byte	
Matching method	Calculate the difference between two Beaufort Numbers.	

Table 27: Specification of Wind Descriptor using Beaufort

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Humidity	
Title	Percentage Humidity Descriptor	
Acronym/Short Name	HD	
Overview	Consists of an integer showing the percentage of the humidity.	
Details	**Descriptor Dimension**	1
	Accuracy	1%
Type of descriptor	Integer between 0 and 100	
Total bytes required	1 byte	
Matching method	Calculate the difference between two humidity instances	

Table 28: Specification of Humidity Descriptor using percentage

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Conditions	
Title	Weather Conditions Descriptor	
Acronym/Short Name	WCD	
Overview	Consists of a string with keys from METAR, separated with spaces.	
Details	**Descriptor Dimension**	1
	Accuracy	NA
Type of descriptor	String	
Total bytes required	~10 bytes	
Matching method	Look for the presence of the keys	

Table 29: Specification of Conditions Descriptor

Key	**Description**
RA	Rain
SN	Snow
GR	Grêle
FG	Fog
NSC	No cloud
FEW	1/8 to 2/8 of cloud coverage
SCT	3/8 to 4/8 of cloud coverage
BKN	5/8 to 7/8 of cloud coverage
OVC	8/8 of cloud coverage

Table 30: Keys for conditions according to METAR

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Lighting	
Title	Lighting Conditions Descriptor	
Acronym/Short Name	LCD	
Overview	Consists of a real number between 0 and 1.	
Details	**Descriptor Dimension**	1
	Accuracy	NA
Type of descriptor	Real	
Total bytes required	8 bytes	
Matching method	Calculate the difference between two lighting descriptors	

Table 31: Specification of Lighting Descriptor

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Environmental	
Title	General Environmental Descriptor	
Acronym/Short Name	GED	
Overview	Consists of a string value.	
Details	**Descriptor Dimension**	1
	Accuracy	NA
Type of descriptor	String	
Total bytes required	~15 bytes	
Matching method	Calculate the difference between two lighting descriptors	

Table 32: General Environmental Descriptor

7.3.6 Environmental Descriptors in Use Cases

Table 33 provides the relevance of the weather descriptors to the I-SEARCH use-cases.

Use Case	Type of user search
UC1: Music retrieval	N/A
UC2: Social retrieval of music	N/A
UC3: Search for Furniture	N/A
UC4: Search for Multimedia objects	N/A
UC5: Search for Product Model	N/A
UC6: Search for 3D model	In the extension of use case, the designer can filter the results according to environmental conditions
UC7: Search for Avatar	N/A

Table 33: Environmental descriptors in use cases

The selected Conditions Descriptor and Lighting is perfectly fits the expected user query in the extension of UC6. However, the user can query the system with any supported environmental descriptor is more familiar (Conditions, Lighting, Humidity, Temperature, Wind).

7.3.7 Identification Descriptors Specification

Table 34 summarizes the specification of the Radio Frequency Identifier descriptor. The descriptor consists of at least 96bits that identify the tagged object. This tag can be meaningful, if access to appropriate databases is provided.

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	Identification Descriptor	
Title	Radio Frequency Identifier Descriptor	
Acronym/Short Name	RFID	
Overview	Consists of the RFID tag	
Details	**Descriptor Dimension**	96 bits or more
	Accuracy	N/A
Type of descriptor	Binary	
Total bytes required	12 or more	
Matching method	Unknown	

Table 34: Radio Frequency Identifier Descriptor

The RFID descriptor can be potentially utilized in the future extensions of UC6 and UC3 I-SEARCH use cases, where the professional user captures the RFID tag of the desired real object using an appropriate device.

7.3.8 Future extensions of real-world descriptors

It is probable that the real-world information which is available at the moment will be extended in the future in order to support different or application-specific information, which is not documented in the current RUCoD format. In order to support future extensions and application-specific real-world descriptors, the Generic Sensor Descriptor (GSD) is introduced and specified in Table 35. The GSD provides a generic methodology to include any unspecified real-world related descriptor in the RUCoD.

Layer-1 Type	Real-World Descriptor	
Layer-2 Type	General Sensor	
Title	General Sensor Descriptor	
Acronym/Short Name	GSD	
Overview	A general abstract class allowing extending the RW specifications.	
Details	**Descriptor Dimension**	Variable
	Accuracy	Unknown
Type of descriptor	String	
Total bytes required	Variable	
Matching method	Unknown	

Table 35: Generic Sensor Descriptor

8. Conclusions

After examining the use-case scenarios and the system requirements, the appropriate social, emotional and real-world descriptors have been appropriately defined and mapped to the use-cases.

This deliverable presented the definition of social, emotional and real-world descriptors that will be used in I-SEARCH. These descriptors are based on existing state-of-the art features and cover a wide-range of applications. The largest part of the deliverable is focused on describing and evaluating the existing social, emotional and real-world descriptors in order to select the most appropriate in order to be included in the I-SEARCH platform. These descriptors are then formally specified and their matching methods are presented. The selection of the descriptors and their format has been appropriately chosen in order to fulfil the current user requirements according to the extracted use cases. More specifically, the needs for intuitive, highly user-centric and efficient search and retrieval haves been taken into account in order to introduce the appropriate descriptors that provide the desired capabilities. The format of the selected descriptors has been appropriately chosen in order to support the current standards in the field.

This deliverable introduces new challenges to be addressed. The specified descriptors are respectable in their field; however, they should be combined to each other in an early-fusion or post-fusion approach for achieving enhanced multimodal retrieval that integrates both low-level and high-level features. This challenge will be addressed in Deliverable D4.2, where all descriptors will be available.

9. Bibliography

[1] Guanling Chen and David Kotz. A Survey of Context-Aware Mobile Computing Research. Technical Report TR2000-381, Dartmouth College, Computer Science, Hanover, NH, November 2000.

[2] Huadong Wu, Mel Siegel, and Sevim Ablay. Sensor Fusion for Context Understanding. In Proceedings of IEEE Instrumentation and Measurement Technology Conference, May 2002.

[3] A. Millonig and K. Schechtner. Developing Landmark-based Pedestrian Navigation Systems. In Proceedings of 8th International IEEE Conference on Intelligent Transportation Systems, pages 197-202, September 2005.

[4] J. Kleban, E. Moxley, J. Xu, and B.S. Manjunath. Global Annotation on Georeferenced Photographs. In Proceedings of the ACM CIVR'09, July 2009.

[5] D. Joshi and J. Luo. Inferring Generic Activities and Events from Image Content and Bags of Geo-tags. In Proceedings of the ACM CIVR'08,July 2008.

[6] P. Bartie and S. Kingham. Media Mapping: Using Georeferenced Images and Audio to provide supporting information for the Analysis of Environmental Sensor Datasets. In Proceedings of International Conference for Free and Open Source Software for Geospatial (FOSS4G), October 2009.

[7] C. Ratti, R.M. Pulselli, S. Williams, and D. Frenchman. Mobile Landscapes: using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design, 33:727-748, 2006.

[8] P. Rudman, S. North, and M. Chalmers. Mobile Pollution Mapping in the City. In Proceedings of UK-UbiNet workshop on eScience and ubicomp, May 2005.

[9] Barry Brown, Matthew Chalmers, Marek Bell, Malcolm Hall, Ian MacColl, and Paul Rudman. Sharing the square: Collaborative Leisure in the City Streets. In Proceedings of the Ninth European Conference on Computer-Supported Cooperative Work, pages 427-447, September 2005.

[10] D. Carboni, S. Sanna, and P. Zanarini. GeoPix: Image Retrieval on the Geo Web, from Camera Click to Mouse Click. In Proceedings of MobileHCI 2006, pages 169-172, September 2007.

[11] The panoramio project. http://www.panoramio.com/.

[12] The degree confluence project. http://www. confluence.org/.

[13] The placeopedia project. http://www.placeopedia.com/.

[14] The plazes service. http://plazes.com/.

[15] The geourl icbm address server. http://geourl.org/.

[16] A. Aftelak, A. Haeyrynen, M. Klemettinen, and S. Steglich. MobiLife: Applications and Services for the User-Centric Wireless World. In Proceedings of IST Mobile and Communications Summit, 2004.

[17] Shwetak N. Patel and Gregory D. Abowd. The ContextCam: Automated Point of Capture Video Annotation. In Proceedings of UbiComp 2004: Ubiquitous Computing, pages 301-318, October 2004.

[18] Keita Watanabe, Koji Tsukada, and Michiaki Yasumura. WillCam: a digital camera visualizing users' interest. In Proceedings of ACM CHI, pages 2747-2752, April 2007.

[19] Nicky Kern, Bernt Schiele, and Albrecht Schmidt. Recognizing context for annotating a live life recording. Personal and Ubiquitous Computing, 11(4):251-263, April 2007.

[20] K. Tsukada, H. Tsujita, and I. Siio. TagTansu:A Wardrobe to Support Creating a Picture Database of Clothes. In Proceedings of Sixth International Conference on Pervasive Computing, pages 49-52, May 2008.

[21] K.P. Tang, I. Smith, A. Ha, L. Satpathy, and J.I. Hong. Memory Karaoke: Aiding Episodic Memory with Location-Enhanced Mobile Phones. In Proceedings of MobileHCI'07, 2007.

[22] T. Tezuka, T. Kurashima, and K. Tanaka. Toward Tighter Integration of Web Search with a Geographic Information System. In Proceedings of the World Wide Web Conference, May 2006.

[23] Ee-Peng Lim, Dion Hoe-Lian Goh, Zehua Liu, Wee-Keong Ng, Christopher Soo-Guan Khoo, and Susan Ellen Higgins. G-Portal: A Map-based Digital Library for Distributed Geospatial and Georeferenced Resources. In Proceedings of ACM JCDL 02, pages 351-358, July 2002.

[24] Shane Ahern, Mor Naaman, Rahul Nair, and Jeannie Yang. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. In Proceedings of ACM JCDL'07, June 2007.

[25] Wenbo Zong, Dan Wu, Aixin Sun, EePeng Lim, and Dion HoeLian Goh. On Assigning Place Names to Geography Related Web Pages. In Proceedings of ACM JCDL'05, June 2005.

[26] Dirk Ahlers and Susanne Boll. Location-based Web Search, chapter 6, pages 55-66. Springer London, 2007.

[27] C.B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT Project. In Proceedings of SIGIR'02, pages 387-388, August 2002.

[28] W. Gao, H.C. Lee, and Y. Miao. Geographically Focused Collaborative Crawling. In Proceedings of the 15th International Conference on World Wide Web, pages 287-296, May 2006.

[29] F. Dawson and T. Howes. vCard MIME Directory Profile. RFC 2426 (Proposed Standard), September 1998.

[30] W3C Semantic Web Interest Group. Basic Geo (WGS84 lat/long) Vocabulary.

[31] The Dublin Core metadata initiative. http://dublincore.org/.

[32] Open Geospatial Consortium. OpenGIS Geography Markup Language (GML) Encoding Standard, 2007.

[33] GML in JPEG 2000 for Geographic Imagery (GMLJP2) (05-047r2)

[34] Yvan G. Leclerc, Martin Reddy, Lee Iverson, and Michael Eriksen. The GeoWeb A New Paradigm for Finding Data on the Web. In Proceedings of the 20th International Cartographic Conference, August 2001.

[35] RFC 3693 Geopriv Requirements. February 2004

[36] “Geographic information – Geography Markup Language (GML) ”, ISO 19136:2007

[37] OGC KML 2.2 Standard, http://www.opengeospatial.org/standards/kml/

[38] The JPEG Standard, (JPEG ISO/IEC 10918-1 ITU-T Recommendation T.81)

[39] The JPEG 2000 image coding system, ISO/IEC 15444:2005

[40] A Uniform Resource Identifier for Geographic Locations ('geo' URI), RFC 5870, June 2010

[41] Markus Tripp, Martin Lechner, “ARML Specification”, http://www.openarml.org

[42] "Data elements and interchange formats - Information interchange - Representation of dates and times", ISO 8601 : 1988 (E).

[43] Bruce Devlin, Jim Wilkinson, Matt Beard, Phil Tudor, Nick Wells, “The MXF Book: An Introduction to the Material eXchange Format”, Elsevier, 2006

[44] R. E. Thayer, *The Biopsychology of Mood and Arousal*, Oxford University Press, Boston, Mass, USA, 1990.

[45] . J. A. Russell, "A circumplex model of affect," J. Personality Social *Psychol.*, vol. 39, pp. 1161–1178, 1980.

[46] S. Canazza, G. De Poli, A. Rodá, and A. Vidolin, "An abstract control space for communication of sensory expressive intentions in music performance," J. New Music Res., vol. 32, no. 3, pp. 281–294, 2003

[47] A.Pikovsky, M.G. Rosemblum, and J. Kurths. Syncronization: a Universal Concept in Nonlinear Science. Cambridge University Press, Cambridge, 2001.

[48] A.Camurri, B. Mazzarino, G. Volpe (2004) Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library, in A. Camurri, G. Volpe (Eds.), Gesture-based Communication in Human-Computer Interaction, LNAI 2915, pp.460-467, Springer Verlag.

[49] A. Camurri, G. De Poli, M. Leman, G. Volpe (2005), "Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications", IEEE MULTIMEDIA, Vol.12, No.1, pp.43-53, IEEE Computer Society Press

[50] A.Camurri, A.Frisoli (Guest Editors) (2006) Special Issue of Virtual Reality Journal on Multisensory Interaction in Virtual Environments, Vol.10, No.1, Springer.

[51] G.Varni, A.Camurri, P.Coletta, G.Volpe (2008) "Emotional Entrainment in Music Performance", Proc. 8th IEEE Intl Conf on Automatic Face and Gesture Recognition, Sept. 17-19, Amsterdam.

[52] G.Castellano, A.Camurri, M.Mortillaro, K.Scherer, G.Volpe (2008) Espressive Gesture and Music: Analysis of Emotional Behaviour in Music Performance, Music Perception, Vol.25, No.6, pp.103-119, University of California Press.

[53] A.Camurri, C.Canepa, P.Coletta, B.Mazzarino, G.Volpe (2008) "Mappe per Affetti Erranti: a Multimodal System for Social Active Listening and Expressive Performance". Proc. Intl Conf NIME 2008 – New Interfaces for Musical Expression, Casa Paganini, University of Genoa (www.nime.org).

[54] G. Varni, A. Camurri, P. Coletta, G. Volpe, "Toward a Real-time Automated Measure of Empathy and Dominance", Proc. Of International Conference on Computational Science and Engineering, Vancouver, Canada, 2009

[55] D. Glowinski, F. Bracco, C. Chiorri, A. Atkinson, P. Coletta, and A. Camurri. An investigation of the minimal visual cues required to recognize emotions from human upper-body movements. In *Proceedings of ACM International Conference on Multimodal Interfaces (ICMI), Workshop on Affective Interaction in Natural Environments (AFFINE).* ACM, 2008.

[56] M. Mancas. Computational attention: Towards attentive computers. Similar edition, 2007. CIACO University Distributors.

[57] M. Mancas, B. Gosselin, and B. Macq. A three-level computational attention model. In *Proc. of ICVS Workshop on Computational Attention & Applications*, Germany, 2007.

[58] M. Mancas, D. Glowinski, G. Volpe, A. Camurri, P. Bretéché, J. Demeyer,T. Ravet, P. Coletta, Real-time motion attention and expressive gesture interfaces, Journal on Multimodal User Interfaces, 2008

[59] D. Glowinski, P. Coletta, G. Volpe, A. Camurri, C. Chiorri, A. Schenone. Multi-scale entropy analysis of dominance in social creative activities. In *Proceedings of the international conference on Multimedia* (MM '10). ACM, New York, NY, USA, 2010

[60] S. Ramdani, B. Seigle, J. Lagarde, F. Bouchara, P.L. Bernard. On the use of sample entropy to analyze human postural sway data. Medical engineering & physics, 2009.

[61] Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Differentiation, classification, and measurement. *Emotion, 8*(4), 494-521

[62] Juslin, P. N., & Sloboda, J. A. (2010). The past, present, and future of music and emotion research. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 933-955). New York: Oxford University Press.

[63] D.Glowinski, N.Dael, A.Camurri, G.Volpe, M.Mortillaro, K.Scherer (2011) Towards a Minimal Representation of Affective Gestures. IEEE Transactions on Affective Computing (In Press).

[64] C. Strapparava and A. Valitutti: “WordNet-Affect: an Affective Extension of WordNet,” *Proceedings of the International Conference on Language Resources and Evaluation,* pp. 1083-1086, 2004

[65] X. Hu, J. S. Downie, A.Ehmann. Lyric Text Mining in Music Mood Classification, In the 10th International Symposium on Music Information Retrieval (ISMIR 2009), Oct. 2009, Kobe, Japan

[66] EmotionML, http://www.w3.org/TR/emotionml/

[67] Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C.(2007). The World of Emotions Is Not Two-Dimensional. Psychological Science, 18(12), 1050-1057.

[68] Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. Current Psychology, 14(4), 261-292.

[69] S.M. Khan and M. Shah. Detecting group activities using rigidity of formation. Proceedings of the 13th annual ACM international conference on Multimedia, pages 403–406, 2005.

[70] International Civil Aviation Organization, METAR standard, www.icao.int

---End of Document ---