Multimed Tools Appl
DOI 10.1007/s11042-011-0888-9

Interactive multi-user video retrieval systems

Marco Bertini · Alberto Del Bimbo · Andrea Ferracani · Lea Landucci · Daniele Pezzatini

Abstract In this paper we present two interactive multi-user systems for video search and browsing. The first is composed by web applications which allows multiuser interaction in a distributed environment; such applications are based on the Rich Internet Application paradigm, designed to obtain the levels of responsiveness and interactivity typical of a desktop application. The second system implements a multi-user collaborative application within a single location, exploiting multi-touch devices. Both systems use the same backend, based on a service oriented architecture (SOA) that provides services for automatic and manual annotation, and an ontology-based video search and browsing engine. Ontology-based browsing let users to inspect the content of video collections; user queries are expanded through ontology reasoning. User-centered field trials of the systems, conducted to assess the user experience and satisfaction, have shown that the approach followed to design these interfaces is extremely appreciated by professional archivists and people working on multimedia.

Keywords Video retrieval · Usability · Multimedia ontologies · Natural interaction

M. Bertini (✉) · A. Del Bimbo · A. Ferracani · L. Landucci · D. Pezzatini
Media Integration and Communication Center, University of Florence, Florence, Italy
e-mail: bertini@dsi.unifi.it

A. Del Bimbo
e-mail: delbimbo@dsi.unifi.it

A. Ferracani
e-mail: ferracani@dsi.unifi.it

L. Landucci
e-mail: landucci@dsi.unifi.it

D. Pezzatini
e-mail: pezzatini@dsi.unifi.it

Published online: 15 October 2011

1 Introduction

Video search engines are the result of advancements in many different research areas: audio-visual feature extraction and description, machine learning techniques, as well as visualization, interaction and user interface design. Automatic video annotation systems are based on large sets of concept classifiers [50], typically based on supervised machine learning techniques such as SVMs. The use of these supervised learning approaches requires tools to easily create ground-truth annotations of videos, indicating the objects and events of interest, in order to train appropriate concept classifiers. The current video search engines are based on lexicons of semantic concepts and perform keyword-based queries [47, 49]. These systems are generally desktop applications or have simple web interfaces that show the results of the query as a ranked list of keyframes [38, 50], similarly to the interfaces made common by web search engines. These systems do not let users to perform composite queries that can include temporal relations between concepts and do not allow to look for concepts that are not in the lexicon. In addition, desktop applications require installation on the end-user computer and can not be used in a distributed environment, while the web-based tools allow only limited user interaction. Regarding video annotation, some manual annotation tools have become popular also in several mainstream video sharing sites such as YouTube and Viddler: they extend the popular idea of image tagging, made popular by Flickr and Facebook, applying it to videos. This information is managed in a collaborative web 2.0 style, since annotations can be inserted not only by content owners (video professionals, archivists, etc.), but also by end users, providing an enhanced knowledge representation of multimedia content.

Within the EU VidiVideo[1] and IM3I[2] projects have been conducted two surveys, formulated following the guidelines of the IEEE 830-1984 standard, to gather user requirements for a video search engine. Overall, more than 50 qualified users from ten different countries participated: roughly half of the users were part of the scientific and cultural heritage sector (e.g. academy and heritage audio-visual archives) and half from the broadcasting, entertainment and video archive industry. The survey campaigns have shown a strong request for systems that have a web-based interface: about 75% of the interviewees considered this “mandatory” and about 20% considered it “desirable”, mainly to increase the interoperability of the archives with systems inside and outside each organization. Regarding the backend, the European Broadcasting Union is working on the introduction of service oriented architectures for asset management [25]. Users also requested the possibility to formulate complex composite queries and to expand queries based on ontology reasoning. Since many archives use controlled lexicons and ontologies in their annotation practices, they also requested to have an interface able to show concepts and their relations.

As for multi-touch interfaces, they have been popularized by personal devices, like smartphones and tablet computers, and are mostly designed for single user

[1] http://www.vidivideo.info

[2] http://www.im3i.eu

interaction. We argue however that multi-touch and multi-user interfaces may be effective in certain production environments [42] like TV news, where people with different roles may have to collaborate in order to produce a new video, thanks to the possibility to achieve a common view with mutual monitoring of others' activities and verbal and gestural utterances [55]; in these cases a web-based interface that forces people to work separately on different screens may not be the most efficient tool.

Some interviews to potential end-users have been conducted with the archivists of RAI, the Italian public broadcaster, in order to collect suggestions and feedbacks about their workflow. Nowadays, RAI journalists and archivists can search the corporate digital libraries through a web-based system. This application provides a simple keyword based search on textual descriptions of the archived videos. Sometimes these descriptions are not very detailed or very relevant to the video content, thus making the document difficult to search or even to find. Moreover the cognitive load required for an effective use of the system often makes the journalists delegate their search activities to the archivists that could be not familiar with the specific topic and therefore could hardly choose the right search keyword. A system that allows different professionals to collaborate interactively and at the same time may thus improve the current workflow.

For these reasons we designed a multi-modal system [24] which can provide both remote and co-located interaction in order to fit the needs of multimedia professionals. The remote interaction is performed by the web-based interface while the co-located one is allowed thanks to the multi-touch collaborative interface; both the interactive modalities exploit the same backend system, developed using a Service Oriented Architecture that exposes services to annotate (manually and automatically), search and browse media collections.

In this paper we present an interactive multi-user video retrieval framework composed by: (i) a SOA-based system that provides services for video semantic and syntactic annotation, search and browsing for different domains (possibly modeled with different ontologies) with query expansion and ontology reasoning; (ii) web-based interfaces for interactive query composition, archive browsing, annotation and visualization; and (iii) a multi-touch interface featuring a collaborative natural interaction application. The usability of the system has been assessed in field trials performed by professional video archivists and multimedia professionals, proving that the system is effective, efficient and satisfactory for users.

The paper is organized as follows: Section 2 describes the state-of-the-art related to the main topics of this work, in Section 3 the architecture and the main functions of the framework are discussed, Section 4 presents the architecture of the interfaces; Section 5 reports on the usability assessment techniques used and on the outcomes of the field trials. Finally, conclusions are drawn in Section 6.

2 Related work

2.1 Semantic annotation

A thorough review of concept-based video retrieval has been presented in [48]. The approach currently achieving the best performance in semantic concept annotation

in competitions like TRECVid [47] or PASCAL VOC [26] is based on the bag-of-words (BoW) approach, as demonstrated by the results of [52]. The bag-of-words approach has been initially proposed for text document categorization, but can be applied to visual content analysis [19, 31, 45], treating an image or keyframe as the visual analog of a document that is represented by a bag of quantized descriptors (e.g. SIFT), referred to as visual-words. A comprehensive study on this approach has been presented in [62], considering the classification of object categories; a review of action detection and recognition methods has been recently provided in [4]. Since the BoW approach disregards the spatial layout of the features, some researchers have proposed several methods to incorporate the spatial information to improve classification results [6]: in [35] it has been proposed to perform pyramid matching in the two-dimensional image space, while in [13] has been proposed the construction of vocabularies of spatial relations between features.

Regarding CBIR we refer the reader to the complete review presented in [22].

2.2 The web-based user interfaces

ALIPR—Automatic Linguistic Indexing of Pictures in Real-time—is an automatic image annotation system [36] that has been recently made public, to let people have their pictures annotated using computer generated tags validated by human annotations. PARAgrab is a Web image search system [32] that exploits visual features and textual meta-data for search.

A web based video search system, with video streaming delivery, has been presented in [30], to search videos obtained from PowerPoint presentations, using the associated metadata. A crowd-sourcing system for retrieval of rock'n'roll multimedia data has been presented in [51], which relies on online users to improve, extend, and share, automatically detected results in video fragments. Another approach to web-based collaborative creation of ground truth video data has been presented in [61], extending the successful LabelMe project from images to videos. A mobile video browsing and retrieval application, based on HTML and JavaScript, has been shown in [14]. In [3] an interactive video browsing tool for supporting content management and selection in postproduction has been presented, comparing the usability of a full-featured desktop application with a limited web-based interface of the same system.

2.3 Multi-touch interface for multimedia retrieval

Single display groupware (SDG) is a research field regarding technology solutions that support work among small, co-present groups through one shared display. The display provides group members with a shared context and focus of attention, supporting co-located cooperative work [53]. The main topics in SDG regard solutions for usability issues about digital access information, data visualization and representation [21, 40], and the evaluation methods [2]. Most SDG projects address the creation and manipulation of digital content using tabletops, while a few works have considered the use of interactive tables to support collaborative search of digital content [56]. Collaborative activities are an important characteristic of the information searching process in task-based information retrieval [29]. Some solutions exploit tangible tokens [11, 57]; other projects allow remote instead of co-located collaboration [16]. Morris et al. in [37] describe the TeamSearch project, an application

supporting co-present collaborative search of metadata-tagged digital content on an interactive table. ShowMe [20] is a a system that exploits personalization to support information browsing and discovery in museum contexts. An interesting approach is used in the Fischlar-DT system [46], that exploits keyframes on a tabletop interface as the users' unit of searching and the system's unit of retrieval. Fischlar-DT is a tabletop application running on DiamondTouch table [23]. Fischlar-DT is designed for two users to collaboratively search for video shots from a video archive of about 80 h of broadcast TV news. The video archive is automatically indexed off-line, using a common shot boundary definition and keyframes provided by TRECVid. The paper presented also a systematic user evaluation of a collaborative multimedia search tool, used by a pair of searchers to satisfy some information seeking goal.

The main difference in our work is the natural interaction approach exploited in the system, which aims to make it become a self-explanatory multi-user interactive tabletop. Our visualization technique avoids standard GUIs as windows, digital keyboard and so on, exploiting ontology structures of data which allow users to perform queries just through selection and browsing. The structured disposition of the query results can be also re-arranged in order to refine them to reach the goal more efficiently.

3 The framework

The interactive multi-user video retrieval systems presented in this paper are part of a complex framework for management, annotation, browsing and retrieval of multimedia data. This framework is based on a service oriented architecture (SOA), that allows different applications and clients to exchange data with each other on a shared and distributed infrastructure, permitting to build new digital library applications on the base of existing services [12]. The framework has three layers (Fig. 1): analysis, SOA architecture and interfaces. The analysis layer provides services for syntactic and semantic annotation of multimedia data, using series of user-definable processing pipelines that can be executed on distributed servers. The SOA Architecture layer routes communications between interfaces and analysis services and provides the main repository functions. Finally, the interface layer provides applications and systems for manual annotation, tagging, browsing and retrieval.

Automatic image and video semantic annotations are provided by a system based on the bag-of-words approach, following the success of this approach for scene and object recognition [27, 45, 60]; the system exploits SIFT, SURF and MSER visual features, that have become the de facto standards because of their good performance and (relatively) low computational cost, and the Pyramid Match Kernel [28], that is robust to clutter and outliers, and is efficient thanks to its linear time complexity in matching. Audio annotation is based on a fusion of timbre features like ZCR, MFCCs, chroma and spectral features and SVM classifiers. These annotations can be complemented and corrected by manual annotations added through a web-based interface (Fig. 3).

CBIR retrieval is performed using MPEG-7 features for visual data and rhythm and pitch features for audio data. In particular have been used Scalable Color, Color Layout and Edge Histogram descriptors to capture different visual aspects with compact and low computational cost descriptors [7].

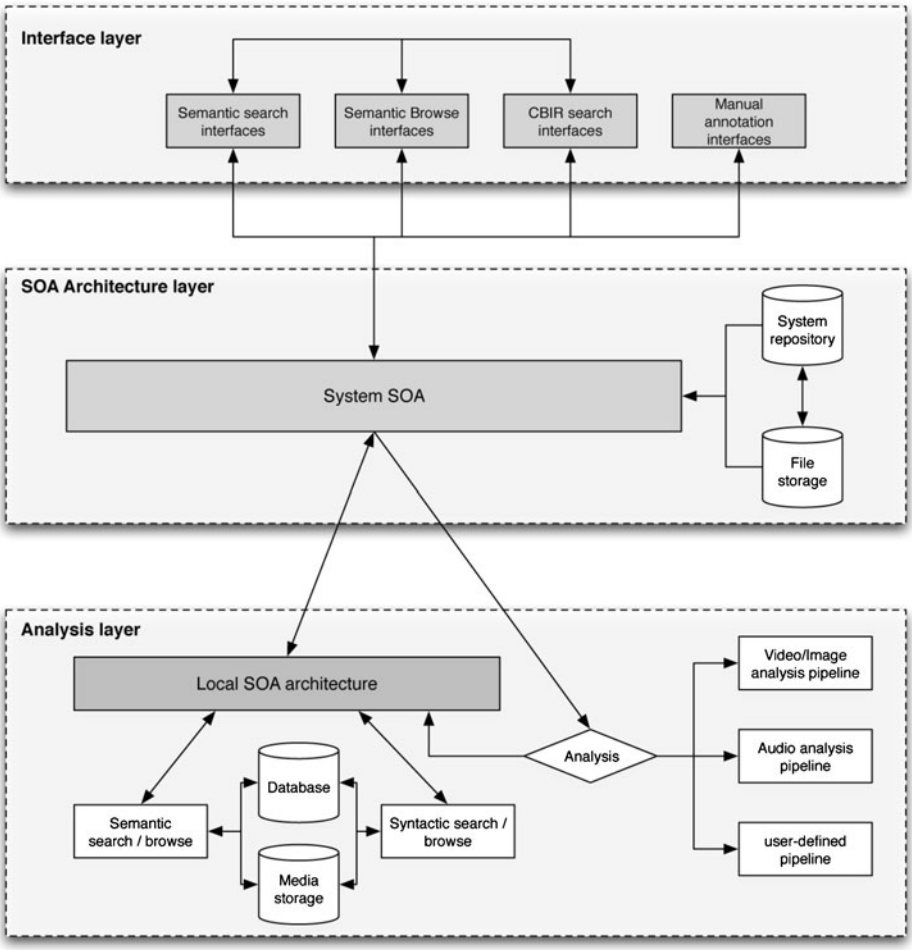


Fig. 1 Overall view of the framework architecture and main layers

3.1 The search engine

The Orione search engine[3] permits different query modalities (free text, natural language, graphical composition of concepts using Boolean and temporal relations and query by visual example) and visualizations, resulting in an advanced tool for retrieval and exploration of video archives for both technical and non-technical users. It uses an ontology that has been created semi-automatically from a flat lexicon and a basic light-weight ontology structure, using WordNet to create concept relations (*is_a*, *is_part_of* and *has_part*) as shown, in a fragment, in Fig. 2. The ontology

[3] A preliminary version of the system was presented in [9].

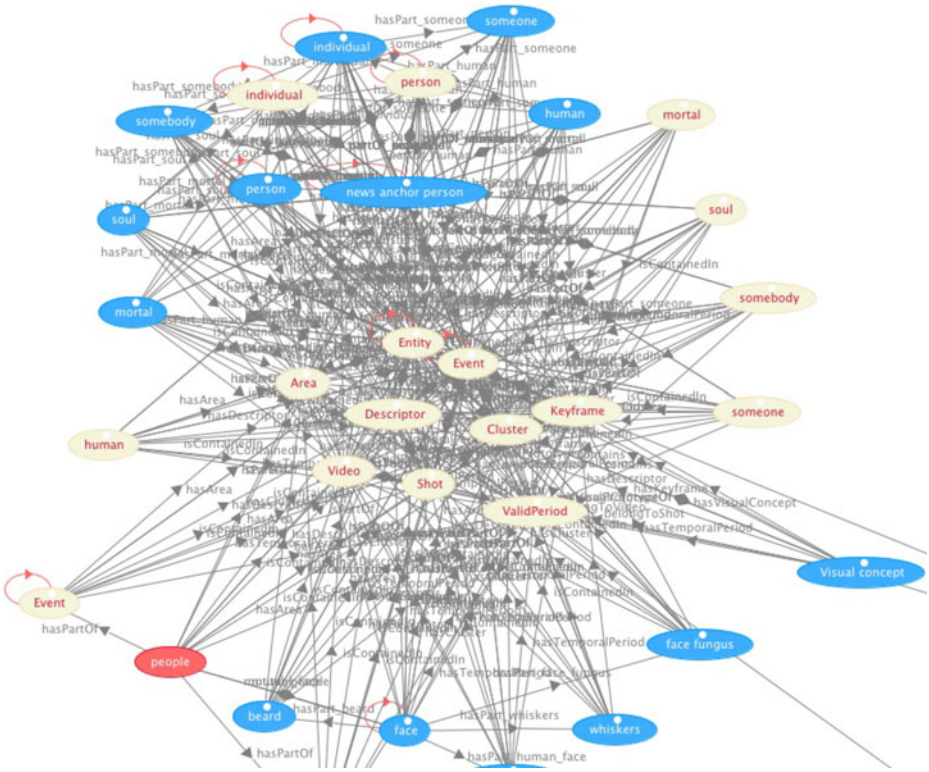


Fig. 2 Automatically created ontology structure used to annotate multimedia content: view of the relations between some concepts related to "people"

is modeled following the Dynamic Pictorially Enriched Ontology model [8], that includes both concepts and visual concept prototypes. These prototypes represent the different visual modalities in which a concept can manifest; they can be selected by the users to perform query by example, using MPEG-7 descriptors (e.g. Color Layout and Edge Histogram) or other domain specific visual descriptors. Concepts, concepts relations, video annotations and visual concept prototypes are defined using the standard Web Ontology Language (OWL) so that the ontology can be easily reused and shared. The queries created in each interface are translated by the search engine into SPARQL, the W3C standard ontology query language. The system extends the queries adding synonyms and concept specializations through ontology reasoning and the use of WordNet. As an example consider the query "*find shots with vehicles*": the concept specializations expansion through inference over the ontology structure permits to retrieve the shots annotated with "vehicle" and also those annotated with the concept's specializations (e.g. "trucks", "cars", etc.). In particular, WordNet query expansion, using synonyms, is enabled when using free-text queries, since it is not desirable to force the user to formulate a query selecting only the terms from a predefined lexicon. As an example consider the case in which a user types the word "automobile": also videos that have been annotated using the word "car" will be returned.

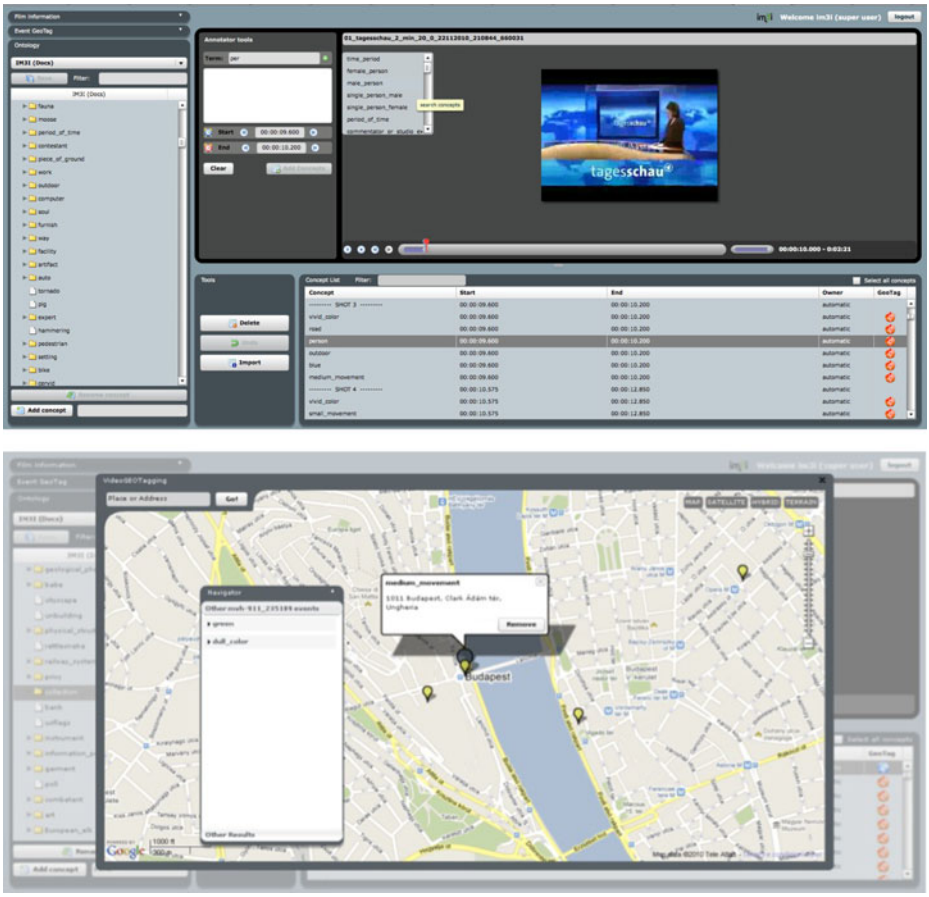


Fig. 3 The Pan web-based manual video annotation system: adding and inspecting annotations (*top*); adding geographical annotations to a visual concept (*bottom*)

The search engine provides also services for browsing the ontology structure and to filter query results using metadata, like programme names or geolocalization information.

3.2 The web-based user interfaces

The web-based search and browse system,[4] based on the Rich Internet Application paradigm (RIA), does not require any software installation and it is composed by several integrated and specialized interfaces. It is complemented by a web-based system for manual annotation of videos (Fig. 3), developed with the aim of creating, collaboratively, manual annotations and metadata, e.g. to provide geolocalization of

[4] http://shrek.micc.unifi.it/im3i/

concepts' annotations. The tool can be used to create ground truth annotations that can be exploited for correcting and integrating automatic annotations or for training and evaluating automatic video annotation systems.

The Sirio semantic search engine is composed by two different interfaces shown in Fig. 4: a simple search interface with only a free-text field for Google-like searches

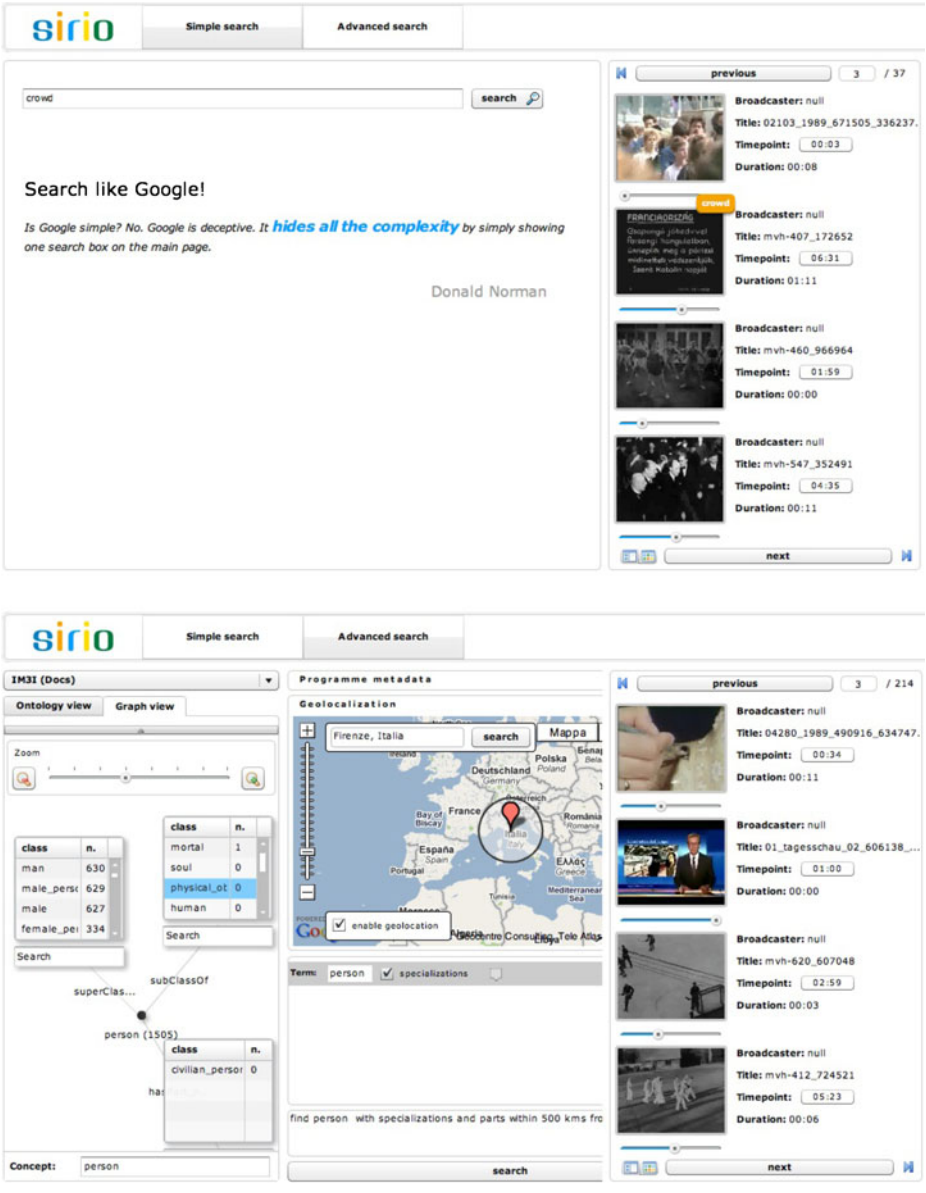


Fig. 4 The Sirio web-based user interfaces: simple user interface (*top*); advanced user interface with ontology graph and geolocalization filtering (*bottom*)

and an advanced search interface with a GUI to build composite queries that may include Boolean and temporal operators, metadata (like programme broadcast information and geo tags) and visual examples. The advanced interface allows also to inspect and use a local view of the ontology graph for building queries. This feature is useful to better understand how one concept is related to the others, thus suggesting possible changes in the composition of the query.

Sirio has two views for showing the list of results: the first presents a list of four videos per page; each video, served by a video streaming server, is shown within a small video player and it is paused in the exact instant of the occurrence of a concept. The other view is an expanded list of results that can show, in a grid, up to thousands of keyframes of the occurrences of the concepts' instances. In both cases users can then play the video sequence and, if interested, zoom in each result displaying it in a larger player that shows more details on the video metadata and allows better video inspection. The extended video player allows also to search for visually similar video shots, using the CBIR search interface Daphnis (Fig. 5).

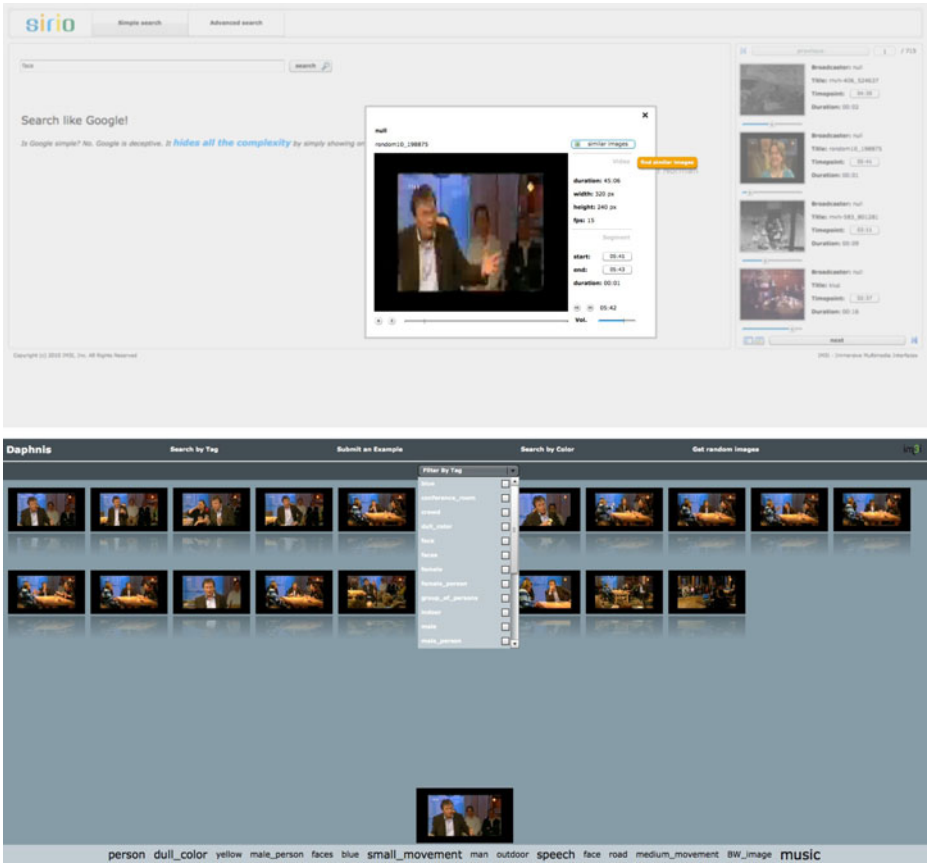


Fig. 5 From semantic search to CBIR search: the Sirio web-based user interfaces: inspection of semantic search results in the Sirio interface (*top*); inspection of CBIR search results in the Daphnis interface (*bottom*)

Furthermore another interface (Andromeda), also integrated with Sirio, allows to browse video archives navigating through the relations between the concepts of the ontology and providing direct access to the instances of these concepts; this functionality is useful when a user does not have a clear idea regarding the query that he wants to make.

The Andromeda interface is based on some graphical elements typical of web 2.0 interfaces, such as the tag cloud. The user starts selecting concepts from a tag cloud, than navigates the ontology that describes the video domain, shown as a graph with different types of relations, and inspects the video clips that contain the instances of the annotated concepts (Fig. 6). Users can select a concept from the ontology graph to build a query in the advanced search interface at any moment.

3.3 The multi-touch user interface

The MediaPick multi-touch UI is a system that allows semantic search and organization of multimedia contents. It has an advanced user-centered design, developed following specific usability principles for search activities [1]. It allows user collaboration [43] on specific contents organized into ontologies that can be explored from general to specific concepts. Users can browse the ontology structure in order to select concepts and start the video retrieval process. Afterwards they can inspect the results returned by the Orione video search engine and organize them according to their specific purposes.

As context of use we considered a television network in which archivists and editors work together to create news reports. Normally the workflow, in this case, starts from the editor asking the archivist to retrieve some pictures and videos

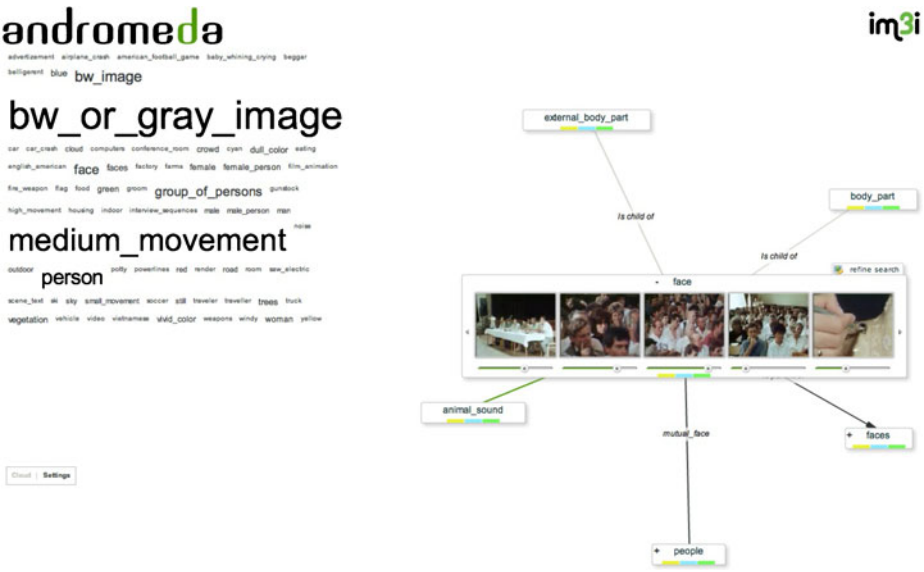


Fig. 6 Browsing videos: select concepts in the tag cloud, then browse the video archive or start a search from a concept of the ontology

regarding a particular topic; the archivist proposes a set of results from which the editor can choose the most suitable for the news report. Our idea is to simplify such workflow in order to make it more efficient and collaborative.

The goal of the MediaPick design is to provide to the broadcast editorial staff an intuitive and collaborative interface to search, visualize and organize video results archived in huge digital libraries with a natural interaction approach. We focused on collaborative applications where the users' goal is to complete a task effectively without degrading the face-to-face experience [41]. Unlike traditional desktop or web applications, multi-user interfaces provide a way for exploiting collaborative dynamics, sharing the same contents on the same workspace, enriching the interaction with the face-to-face communication aspect. Such kind of interaction encourages discussion, ideas sharing and brain storming (Fig. 7).

The user interface adopts some common visualization principles derived from the discipline of Information Visualization [15] and is equipped with a set of interaction functionalities designed to improve the usability of the system. The GUI consists of a concepts view (Fig. 8, top) and a results view (Fig. 8, bottom). The first view allows to select an use one or more keywords from an ontology structure to query the digital library; the second one shows the videos returned from the database that the user can navigate, select, drag and organize on the workspace.

The concepts view consists of two different interactive elements: the ontology graph to explore the concepts and their relations (Fig. 8, top, a), and the controller module to save the selected concepts and switch to the results view (Fig. 8, top, b). The user chooses the concepts as query from the ontology graph. Each node of the graph consists of a concept and a set of relations. The concept can be selected and then saved into the controller, while a relation can be triggered to list the related concepts, which can be expanded and selected again; then the cycle repeats. The related concepts are only shown when a precise relation is triggered in order to minimize the number of visual elements displayed at the same time in the interface. After saving all the desired concepts into the controller, the user is able to change the state of the interface and go to the results view. Also the results view shows the controller to let the user select the archived concepts and launch the query against the database. The returned data are then displayed within a horizontal list in a new visual component placed at the top of the screen which contains returned videos and their related concepts (Fig. 8, bottom, b). Each video element has three different states: idle, playback and information. In the idle state the video is represented

Fig. 7 Examples of collaborative use of MediaPick

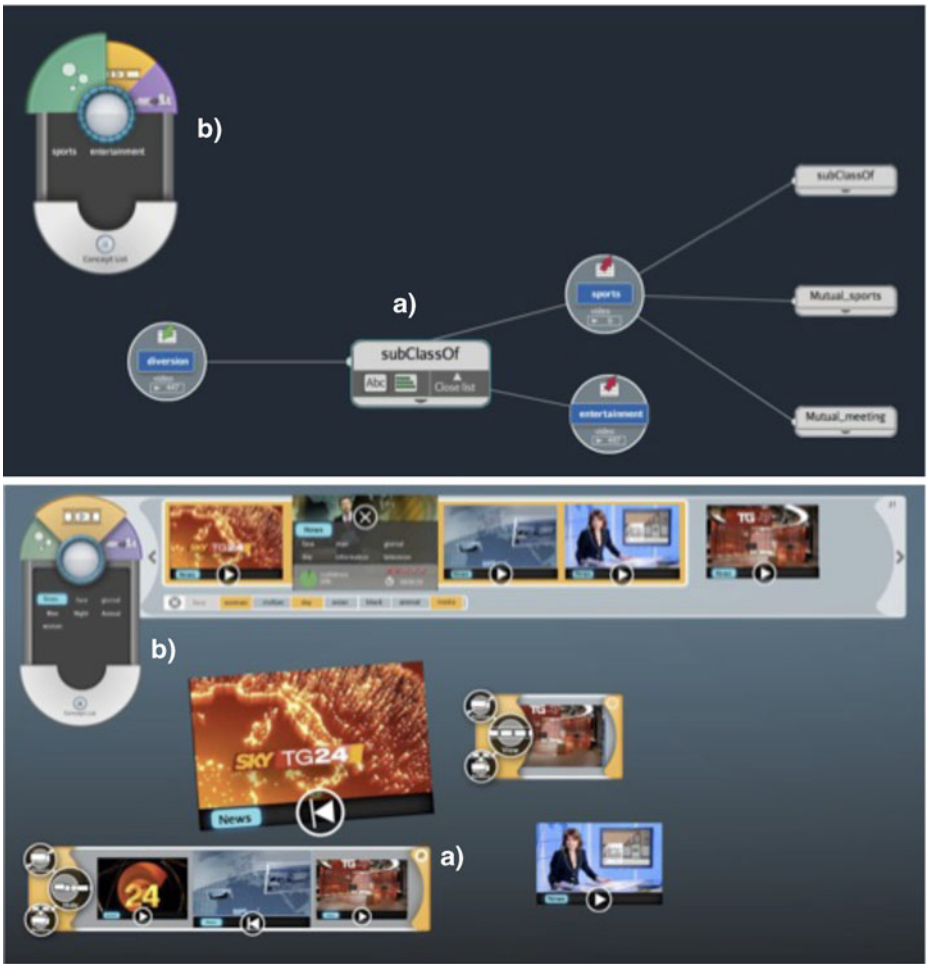


Fig. 8 MediaPick system: concepts view (*top*): **a** the ontology graph; **b** controller module; results view (*bottom*): **a** video players and user's result picks; **b** controller module and results list

Table 1 Gestures and actions association for the multi-touch user interface

Gesture	Actions
Single tap	– Select concept
	– Trigger the controller module switch
	– Select video group contextual menu options
	– Play video element
Long pressure touch	– Trigger video information state
	– Show video group contextual menu (Fig. 9a)
Drag	– Move video element
Two-fingers pinch	– Resize videos (Fig. 9b)
Two-fingers framing	– Group videos (Fig. 9c)

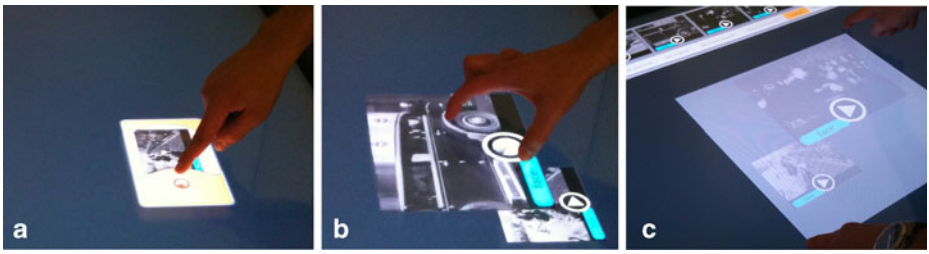


Fig. 9 Gestures and actions: **a** long pressure touch to show video information; **b** two finger pinch to resize a video; **c** two fingers framing to group videos

with a keyframe and a label visualizing the concept used for the query. During the playback state the video starts playing from the frame in which the selected concept was annotated. A longer touch of the video element activates the information state, that shows a panel with some metadata (related concepts, confidence, duration, etc.).

At the bottom of the result list there are all the concepts related to the video results. By selecting one or more of these concepts, the video clips returned are filtered in order to improve the information retrieval process. The user can select any video element from the results list and drag it outside. This action can be repeated for other videos returned by the same or other queries. Videos placed out of the list can be moved along the screen, resized or played. A group of videos can be created by collecting two or more video elements in order to define a subset of results. Each group can be manipulated as a single element through a contextual menu: it can be expanded to show a list of its elements or released in order to ungroup the videos. The groups of videos created by users during the search process represents the final result of the working session. As such, it can be stored on the database or shared with other end-users.

All the user interface actions mentioned above are triggered by natural gestures, reported in Table 1. Some gestures and relatives feedbacks are shown in Fig. 9.

Regarding the workflow described above, we tried to exploit multi-touch interaction and ontology-based search to offer to the archivists a tool that could help them in their tasks. In particular, the exploration of the ontology concepts (represented with a graph metaphor) could help archivists to find suitable keywords even if he is not familiar with the topic of the search. Moreover, organizing video results into groups could help the editor when he has to choose contents for the video report.

4 Interfaces architecture

Following [44], in order to design the interfaces we took into account the following requirements for the development of a HCI system.

Time to learn Necessary length of time for learning how to use the interface.

Retention over time How well can a user remember the interface after some time has passed. The closer the syntax of the operations match the user's understanding,

the easier it will be to remember how to use the interface. If the time to learn is fast, then the retention is less important.

Speed of performance It is linked to the speed of the user interface (not the speed of the software). It is the number of characters to type, buttons to press, hands touch or movement to execute to carry out an operation. This characteristic normally conflicts with the time to learn requirement: often faster systems are the harder ones to learn.

Real-time response The time required by the system for feedback purposes; in the case of human-computer interfaces, a system operates in real-time if the user does not perceive a delay between her/his action and the system's response.

Rate of errors by users The rate of errors produced by a user can be due to a bad structure of the user interface. It is affected by factors such as consistency or arrangement of screens in GUIs.

Subjective satisfaction It refers to how comfortable users feel with the software. This criterion is fairly difficult to measure, and it depends on the user's individual preferences and background.

Speed of performance and real-time response requirements are obviously influenced not only by the available computational resources, but also by the backend response speed and by the typology and amount of multimedia contents. Moreover, the real-time response is defined also from a feedback point of view: delays due to server access and query execution are communicated to users through visual feedbacks; in this way the system tells them it has received their requests and users can confidently wait for its response.

The rate of errors and subjective satisfaction have been taken into account in all the software life cycle and the application has been intensively tested during the developing phase and after the first release.

Finally, since the time to learn is the major requirement influencing the interface design in natural interaction systems, the interfaces were developed in order to be self-explanatory, thus letting users learn from their own experimentation and experience.

MediaPick exploits a multi-touch technology chosen among various approaches experimented in our lab since 2004 [5]. Our solution uses an infrared LED array as an overlay built on top of an LCD standard screen (capable of a full-HD resolution). The multi-touch overlay detects fingers and objects on its surface and sends information about touches using the TUIO [33] protocol at a rate of 50 packets per second. MediaPick architecture is composed by an input manager layer that communicates trough the server socket with the gesture framework and the core logic. The latter is responsible of the connection to the web services and media server as well as the rendering of the GUI elements on the screen (Fig. 10).

The input management module is driven by the TUIO dispatcher: this component is in charge of receiving and dispatching the TUIO messages sent by the multi-touch overlay to the gesture framework through the server socket (Fig. 10). This module is able to manage the events sent by the input manager, translating them into commands for the gesture framework and the core logic.

The logic behind the multi-touch interfaces needs a dictionary of gestures which users are allowed to perform. It is possible to see each digital object on the surface

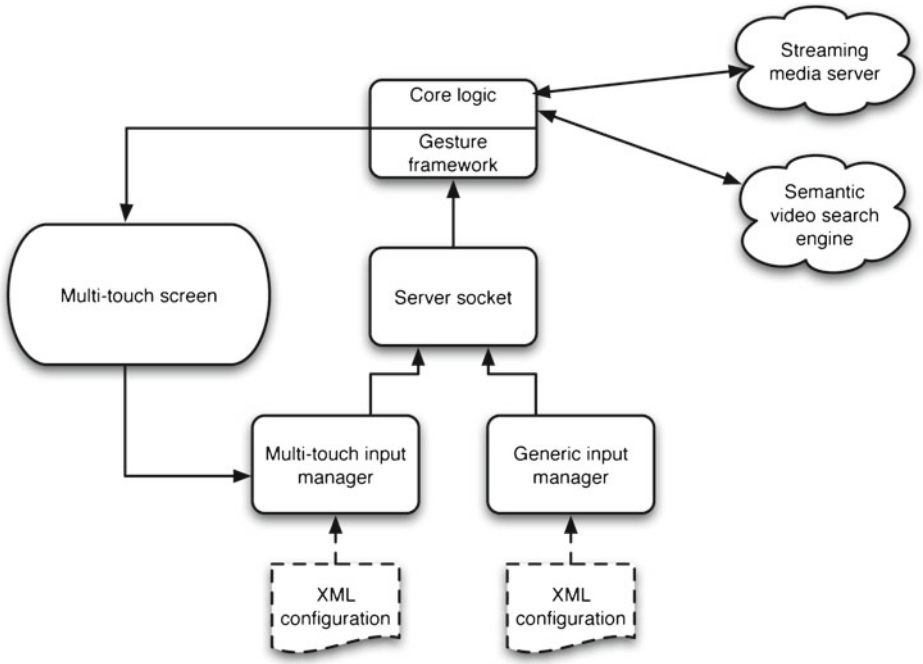


Fig. 10 Multi-touch interface system architecture

like an active touchable area; for each active area a set of gestures is defined for the interaction, and each gesture is linked to the active area in which it is enclosed. For this reason each active area has its own set of touches and allows the gesture recognition through the interpretation of their associated behavior.

The system backend and the search engine are currently based on open source tools (i.e. MySQL database, Apache Tomcat application server and Red 5 video streaming server) or freely available commercial tools (Adobe Media Server has a free developer edition). Videos are streamed using the RTMP video streaming protocol.

The Orione search engine is developed in Java and supports multiple ontologies and ontology reasoning services. Ontology structure and concept instances serialization have been designed so that inference can be execute simultaneously on multiple ontologies, without slowing the retrieval; this design allows to avoid the need of selecting a specific ontology when creating a query with the Google-like interface in Sirio or in MediaPick. The engine has also been designed to fit into a service oriented architecture, so that it can be incorporated into the customizable search systems, other than Sirio and MediaPick, that have been developed within IM3I project. Audio-visual concepts are automatically annotated, using the IM3I automatic annotation engine, or manually annotated using the Pan web application (Fig. 3). The search results are produced in RSS 2.0 XML format, with paging, so that they can be used as feeds by any RSS reader tool and it is possible to subscribe to a specific search. Both the web-based interfaces and the multi-touch interface have been developed in Flex+Flash, according to the Rich Internet Application paradigm.

5 System evaluation

According to ISO, usability is defined as the extent that a user can utilize a product effectively, efficiently and satisfactorily in achieving a specific goal. However, researchers in the field of usability have defined a variety of views on what usability is and how to assess it. Some follow ISO standards on quality models (ISO 9126) as in [18], others follow user-centered design (ISO 9241) or user-centered approaches [59]. Moreover, it has been observed that usability guidelines are often not comprehensive, e.g. large portions of recommendations are specific of each guidelines [10], and thus studies on usability of digital video library systems have used a variety of methods and guidelines [17].

The methodology used in the field trials of the two systems follows the practices defined in the ISO 9241 standard and of the guidelines of the U.S. Department of Health and Human Sciences [58], and gathered: (i) observational notes taken during test sessions by monitors, (ii) verbal feedback noted by test monitors and (iii) a survey completed after the tests by all the users. The usability tests have been designed following the task-oriented and user-centered approaches presented in [54, 58]; in particular the test aimed at evaluating the main characteristics of usability (as defined in ISO 9126 and ISO 9241) that are: (i) understandability, i.e. if the user comprehends how to use the system easily; (ii) learnability, i.e. if the user can easily learn to use the system; (iii) operability, i.e. if the user can use the system without much effort; (iv) attractiveness, i.e. if the interface looks good; (v) effectiveness, i.e. the ability of users to complete tasks using the system; and (vi) satisfaction, i.e. users' subjective reactions to using the system. Before the trials, users received a short interactive multimedia tutorial of the systems, as in [17].

The questionnaires filled by the users contained three types of questions [34]: factual types, e.g. the number of years they have been working in a video archive or the types of tools commonly used for video search; attitude-type, to focus the respondent's attention to inside themselves and his response to the system; option-type, to ask the respondent what they think about features of the systems, e.g. what is the preferred search mode. The answers to the attitude-type questions followed a five point Likert scale questionnaire, addressing frequency and quality attitudes.

5.1 Web usability test

The web-based interfaces have been tested to evaluate the usability of the system in a set of field trials. A group of 18 professionals coming from broadcasting, media industry, cultural heritage institutions and video archives in Italy, The Netherlands, Hungary and Germany have tested the system on-line (running on the MICC servers), performing a set of pre-defined tasks, and interacting with the system. These activities were:

Task 1: Concept search.

In this task users had to perform two search activities using the advanced and simple search interfaces. Each search activity was composed by sub-activities, e.g. using the ontology graph to inspect concepts' relations, filter results using metadata, etc.. Each sub-activity was evaluated but, for the sake of brevity, we report only the overall results.

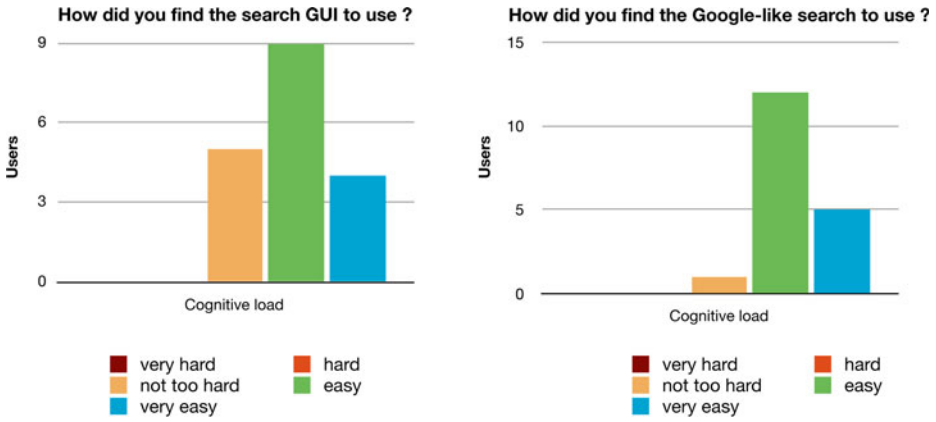


Fig. 11 Web-based interfaces usability tests: usability of the advanced search and simple search interfaces in task 1

Task 2: Concept browsing.

In this task users had to browse the ontology, searching for concepts and concepts' instances (i.e. video clips) related to "person".

Task 3: CBIR search.

In this task users had to perform a semantic-based search for "face" or "person" using the advanced search interface and then search for shots containing an "anchorman" or "Angela Merkel" based on visual similarity.

Test monitors have recorded observational notes and verbal feedbacks of the users; these notes have been analyzed to understand the more critical parts of the system and, during the development of the IM3I project, they have been used to redesign the functionalities of the system. In particular, a search interface based on natural language queries has been dropped from the system because users found it too difficult to be used and not providing enough additional functionalities w.r.t. the advanced and simple search interfaces.

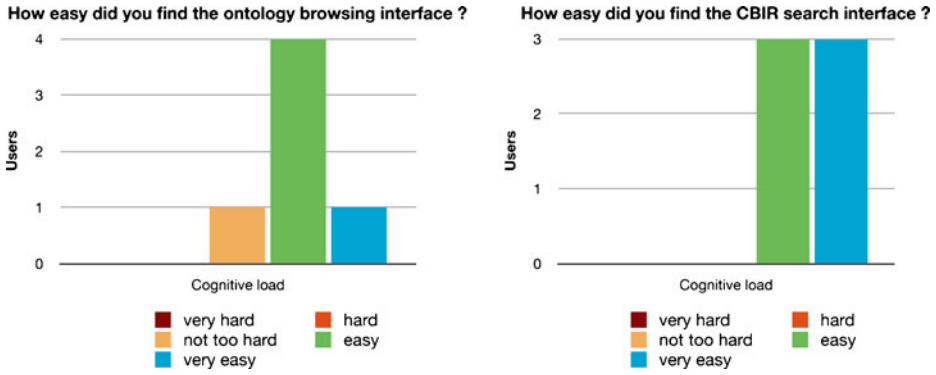


Fig. 12 Web-based interfaces usability tests: usability of the Andromeda and Daphnis interfaces in tasks 2 and 3

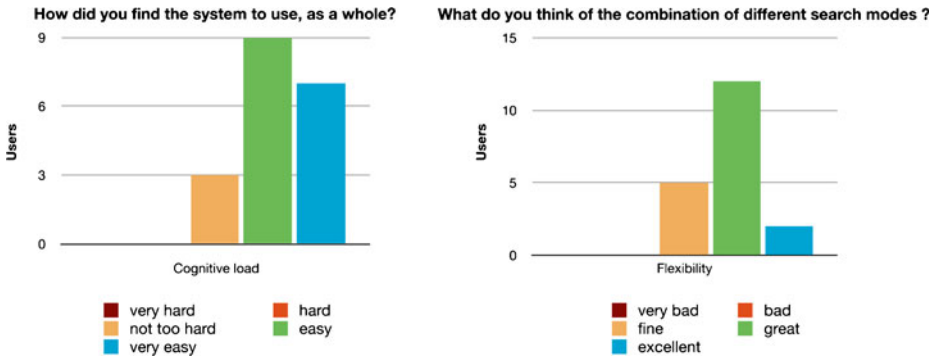


Fig. 13 Overview of web-based interfaces usability tests: overall usability of the system, usability of the combination of search modalities

Figure 11 reports evaluations for the advanced and simple search modalities used in task 1. Figure 12 reports results for tasks 2 and 3. These tests were carried on only by six participants, that work as archivists in a national video archive. However, as noted in [39, 58], this number of participants is enough for such usability tests.

Figure 13 summarizes two results of the tests. The overall experience is very positive and the system proved to be easy to use, despite the objective difficulty of interacting with a complex system for which the testers received only a very limited training. Users appreciated the combination of different interfaces. The type of search interface that proved to be more suitable for the majority of the users is the Sirio advanced interface because of its many functionalities that are suitable for professional video archivists.

5.2 MediaPick usability test

In order to evaluate the *time to learn* [44] of MediaPick, an important feature for a system whose type of interaction is still quite uncommon, we tested both trained and untrained users: in the first case the test was conducted after a brief introduction and explanation of the interaction language (about two minutes), in the second one without any introduction. During the introduction the main gestures of the

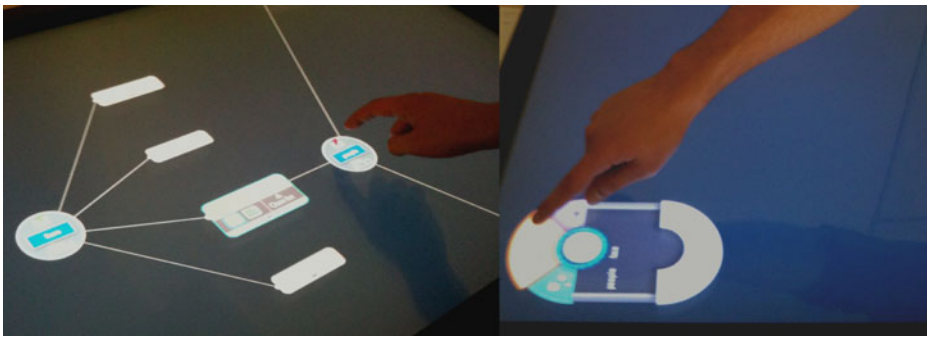


Fig. 14 MediaPick evaluation test—task 1: select a concept (*left*) and perform the query (*right*)

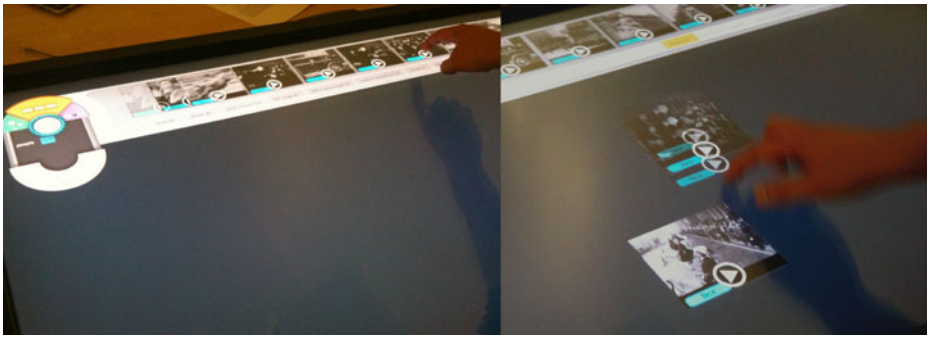


Fig. 15 MediaPick evaluation test—task 2: filter results (*left*) and drag the result in the workspace (*right*)

interaction language have been explained to users: in particular how to select, drag, zoom, collect, activate and filter results. With the knowledge of such gestures users should be able to perform all the requested tasks. After the first system release we tested it with twenty users (thirteen trained and seven untrained users) with good multimedia skills, both professionals and graduated students (attending the Master in Multimedia Content Design, University of Florence). We provided them with a task-oriented questionnaire in order to evaluate the system performance, according to the criteria used for the trials of the web-based interfaces and to those presented in [44]. Users had to perform three tasks, that were:

Task 1: concept selection.

In this task users have to search end select a particular concept in the graph, in order to perform the query (Fig. 14).

They are requested to: select the concept "face", open the relation "mutual", order the result list alphabetically and finally select the concept "people" and perform the query.

Average performing time 28″ (trained), 1′ and 20″ (untrained). **Not performed**: none.

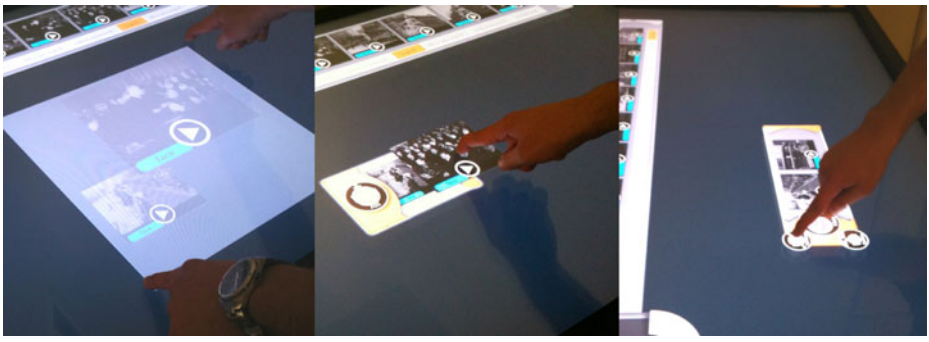


Fig. 16 MediaPick evaluation test—task 3: group video items (*left*), add new item into the group (*center*) and ungroup items (*right*)

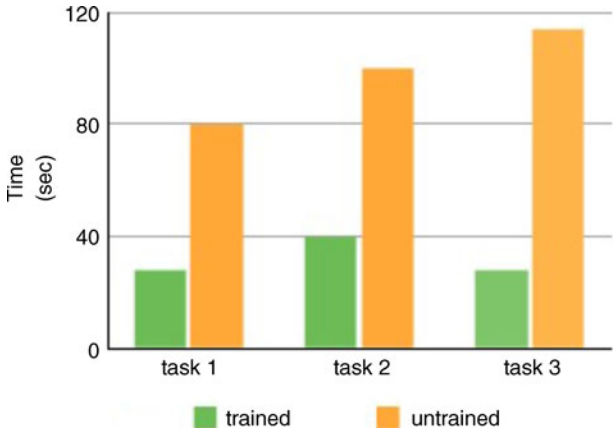


Fig. 17 MediaPick evaluation time results for the three tasks

Task 2: browse filtered results.

In this task users refine the search results filtering them by a particular concept and then drag the resulting item in the workspace in order to watch it (Fig. 15).

They are requested to: filter the search results by the concept "horse", drag the result item in the workspace, play the video and finally stop it.

Average performing time 40″ (trained), 1′ and 40″ (untrained). **Not performed**: none.

Task 3: group and ungroup results.

In this task users group relevant video items and then ungroup them (Fig. 16).

They are requested to: drag two relevant videos in the workspace, group them into a folder, select another video and add it to the folder created and finally ungroup the videos.

Average performing time 28″ (trained), 1′ and 54″ (untrained). **Not performed**: none of trained users and three out of seven untrained users (42%).

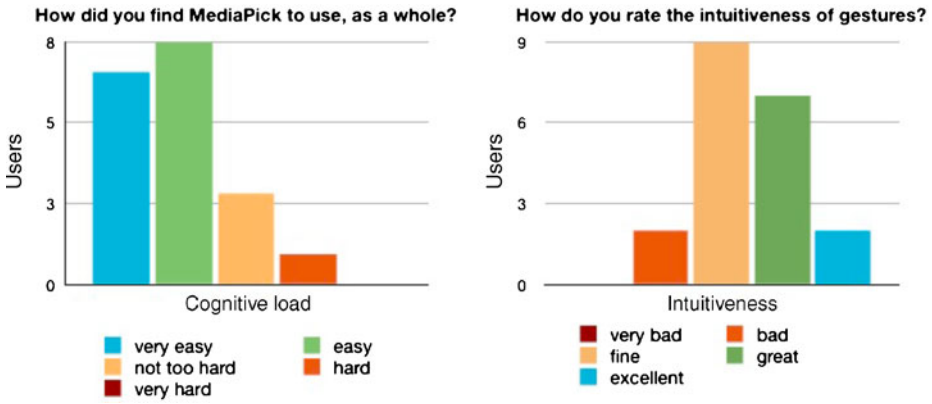


Fig. 18 Overall results of MediaPick usability and experience evaluation

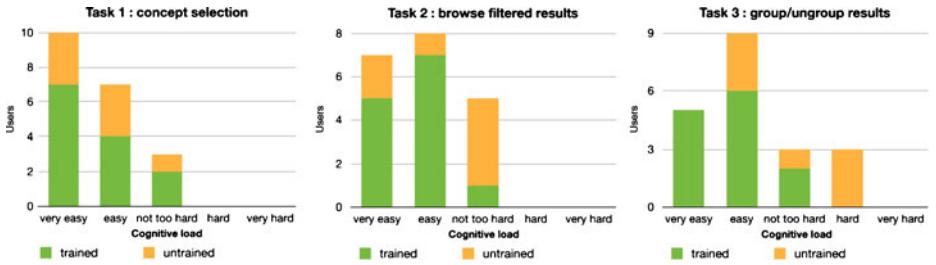


Fig. 19 MediaPick evaluation test results

Analyzing the evaluation results we can argue that, with a brief training before the interaction, users can exploit MediaPick successfully (Figs. 17 and 18). In fact, the results are very encouraging considering that the system is designed to be used by advanced users. Only task 3, that required some familiarity with interactive surfaces, resulted too complex for some of the untrained users (Fig. 19). Most users reported in the section for suggestions and improvements of the questionnaire, that the gesture for the "group" function was not sufficiently intuitive. Almost all the untrained users in the testing phase tried to overlap the videos instead of framing them in order to perform the "group" function. This kind of "overlap" gesture could be useful only when user want to group just two videos: when several videos have to be grouped, such gesture would rise the *speed of performance*. Maybe the solution would be to allow both gestures in order to perform grouping function. Future work for improving MediaPick will include the revision of such gesture in order to make it easier and more intuitive.

6 Conclusions

In this paper we presented two semantic video search systems based on web and multi-touch multi-user interfaces that share a common framework based SOA architecture, and tested by a large number of professionals in user-centered field trials. Usability evaluation has shown that the overall experience and usability of the systems is very good.

Our future work will deal with further development of the interfaces and of the automatic annotation systems of the backend. For the web-based system we will especially consider the new HTML5 technologies. Regarding the multi-touch system, we will improve the interface by modifying those functions considered less intuitive (e.g. group/ungroup function) and adding annotation and browsing functionalities. Finally, we will address scalability in the automatic annotation services.

Acknowledgements The authors thank Giuseppe Becchi for his work on software development. This work was partially supported by the EU IST IM3I project (http://www.im3i.eu/—contract FP7-222267).

References

1. Amant RS, Healey CG (2001) Usability guidelines for interactive search in direct manipulation systems. In: Proc. of international joint conference on artificial intelligence, vol 2, pp 1179–1184

2. Apted T, Collins A, Kay J (2009) Heuristics to support design of new software for interaction at tabletops. In: Proc. of CHI workshop on multitouch and surface computing, pp 1–4
3. Bailer W, Weiss W, Kienast G, Thallinger G, Haas W (2010) A video browsing tool for content management in postproduction. International Journal of Digital Multimedia Broadcasting, Hindawi Publishing Corporation, vol 2010, Article ID 856761, 17 pp. doi:10.1155/2010/856761
4. Ballan L, Bertini M, Del Bimbo A, Seidenari L, Serra G (2011) Event detection and recognition for semantic annotation of video. Multimed Tools Appl 51(1):279–302
5. Baraldi S, Bimbo A, Landucci L (2008) Natural interaction on tabletops. Multimed Tools Appl (MTAP) 38:385–405
6. Behmo R, Paragios N, Prinet V (2008) Graph commute times for image representation. In: Proc. of IEEE conference on computer vision and pattern recognition (CVPR), pp 1–8
7. Bertini M, Del Bimbo A, Nunziati W (2006) Video clip matching using MPEG-7 descriptors and edit distance. In: Proc. of international conference on image and video retrieval (CIVR), pp 133–142. Tempe, AZ, USA
8. Bertini M, Del Bimbo A, Serra G, Torniai C, Cucchiara R, Grana C, Vezzani R (2009) Dynamic pictorially enriched ontologies for digital video libraries. IEEE MultiMedia 16(2): 42–51
9. Bertini M, D'Amico G, Ferracani A, Meoni M, Serra G (2010) Sirio, Orione and Pan: an integrated web system for ontology-based video search and annotation. In: Proc. of ACM international conference on multimedia (ACM MM), pp 1625–1628
10. Bevan N, Spinhof L (2007) Are guidelines and standards for web usability comprehensive? In: Jacko J (ed) Human–computer interaction. Interaction design and usability. Lecture notes in computer science, vol 4550. Springer, Berlin, pp 407–419
11. Blackwell AF, Stringer M, Toye EF, Rode JA (2004) Tangible interface for collaborative information retrieval. In: Proc. of CHI conference on human factors in computing systems (CHI), pp 1473–1476
12. Brettlecker G, Milano D, Ranaldi P, Schek HJ, Schuldt H, Springmann M (2007) ISIS and OSIRIS: a process-based digital library application on top of a distributed process support middleware. In: Proc. of 1st international conference on digital libraries: research and development (DELOS), pp 46–55
13. Bronstein AM, Bronstein MM (2010) Spatially-sensitive affine-invariant image descriptors. In: Proc. of European conference on computer vision (ECCV), pp 197–208
14. Bursuc A, Zaharia T, Prêteux F (2010) Mobile video browsing and retrieval with the ovidius platform. In: Proc. of ACM international conference on multimedia (ACM MM), pp 1659–1662
15. Card SK, Mackinlay JD, Shneiderman B (eds) (1999) Readings in information visualization: using vision to think. Morgan Kaufmann, San Mateo
16. Chen F, Eades P, Epps J, Lichman S, Close B, Hutterer P, Takatsuka M, Thomas B, Wu M (2006) Vicat: visualisation and interaction on a collaborative access table. In: Proc. of the first IEEE international workshop on horizontal interactive human-computer systems. IEEE Computer Society, Washington, DC, pp 59–62
17. Christel M, Moraveji N (2004) Finding the right shots: assessing usability and performance of a digital video library interface. In: Proc. of ACM international conference on multimedia (ACM MM), pp 732–739
18. Chua B, Dyson L (2004) Applying the ISO 9126 model to the evaluation of an e-learning system. In: Proc. of the 21st ASCILITE conference, pp 184–190
19. Chum O, Philbin J, Sivic J, Isard M, Zisserman A (2007) Total recall: automatic query expansion with a generative feature model for object retrieval. In: Proc. of international conference on computer vision (ICCV), pp 1–8
20. Collins A of Sydney U (2010) Making the tabletop personal: employing user models to aid information retrieval/Anthony Collins ... [et al]. University of Sydney, School of Information Technologies, Sydney, NSW. http://sydney.edu.au/engineering/it/~tr/
21. Czerwinski M, Robertson G, Meyers B, Smith G, Robbins D, Tan D (2006) Large display research overview. In: Proc. of CHI extended abstracts on human factors in computing systems, pp 69–74. Association for Computing Machinery (ACM) Press, Montreal, Canada. doi:10.1145/1125451.1125471
22. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40:5:1–5:60
23. Dietz P, Leigh D (2001) Diamondtouch: a multi-user touch technology. In: Proc. of ACM symposium on user interface software and technology (UIST), pp 219–226

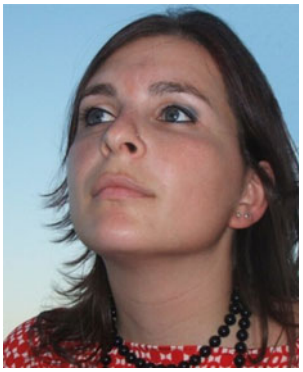
24. Dumas B, Lalanne D, Oviatt S (2009) Multimodal interfaces: a survey of principles, models and frameworks. In: Lalanne D, Kohlas J (eds) Human machine interaction. Lecture notes in computer science, vol 5440. Springer, Berlin, pp 3–26
25. European Broadcasting Union—Common Processes group: EBU Technical—Services Oriented Archictectures. http://tech.ebu.ch/soa. Accessed August 2011
26. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The Pascal Visual Object Classes (VOC) challenge. Int J Comput Vis 88(2):303–338
27. Fergus R, Perona P, Zisserman A (2003) Object class recognition by unsupervised scale-invariant learning. In: Proc. of IEEE conference on computer vision and pattern recognition (CVPR)
28. Grauman K, Darrell T (2007) The pyramid match kernel: efficient learning with sets of features. J Mach Learn Res (JMLR) 8:725–760
29. Hansen P, Järvelin K (2005) Collaborative information retrieval in an information-intensive domain. Inf Process Manag 41:1101–1119
30. Halvorsen P, Johansen D, Olstad B, Kupka T, Tennøe S (2010) vesp: enriching enterprise document search results with aligned video summarization. In: Proc. of ACM international conference on multimedia (ACM MM), pp 1603–1604
31. Jiang YG, Ngo CW, Yang J (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proc. of ACM international conference on image and video retrieval (CIVR), pp 494–501
32. Joshi D, Datta R, Zhuang Z, Weiss WP, Friedenberg M, Li J, Wang JZ (2006) Paragrab: a comprehensive architecture for web image management and multimodal querying. In: Proc. of the international conference on very large data bases (VLDB), pp 1163–1166
33. Kaltenbrunner M, Bovermann T, Bencina R, Costanza E (2005) TUIO a protocol for table-top tangible user interfaces. In: Proc. of international workshop on gesture in human-computer interaction and simulation
34. Kirakowski J (2000) Questionnaires in usability engineering: a list of frequently asked questions, 3rd edn. Available at: http://www.ucc.ie/hfrg/resources/qfaq1.html
35. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proc. of IEEE conference on computer vision and pattern recognition (CVPR), vol 2, pp 2169–2178
36. Li J, Wang JZ (2006) Real-time computerized annotation of pictures. In: Proc. of ACM international conference on multimedia (ACM MM), pp 911–920
37. Morris M, Paepcke A, Winograd T (2006) Teamsearch: comparing techniques for co-present collaborative search of digital media. In: First IEEE international workshop on horizontal interactive human-computer systems, 2006. TableTop 2006, p 8
38. Natsev A, Smith J, Tešić J, Xie L, Yan R, Jiang W, Merler M (2008) IBM Research TRECVID-2008 video retrieval system. In: Proc. of TRECVID workshop
39. Nielsen J (2000) Why you only need to test with 5 users. Available at: http://www.useit.com/alertbox/20000319.html
40. Ryall K, Morris MR, Everitt K, Forlines C, Shen C (2006) Experiences with and observations of directtouch tabletops. In: Proc. of IEEE tabletop the international workshop on horizontal interactive human computer systems, pp 89–96
41. Scott SD, Grant KD, Mandryk RL (2003) System guidelines for co-located, collaborative work on a tabletop display. In: Proc. of European conference computer-supported cooperative work (ECSCW). Helsinki, Finland
42. Shen C (2006) Multi-user interface and interactions on direct-touch horizontal surfaces: collaborative tabletop research at merl. In: Proc. of IEEE international workshop on horizontal interactive human-computer systems
43. Shen C (2007) From clicks to touches: enabling face-to-face shared social interface on multi-touch tabletops. In: Schuler D (ed) Online communities and social computing. Lecture notes in computer science, vol 4564. Springer, Berlin, pp 169–175
44. Shneiderman B (1998) Designing the user interface: strategies for effective human-computer interaction. Addison-Wesley, Reading
45. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: Proc. of international conference on computer vision (ICCV)
46. Smeaton A, Lee H, Foley C, McGivney S (2007) Collaborative video searching on a tabletop. Multimedia Syst 12:375–391
47. Smeaton A, Over P, Kraaij W (2009) High-level feature detection from video in TRECVid: a 5-year retrospective of achievements. In: Multimedia content analysis, theory and applications, pp 151–174

48. Snoek CGM, Worring M (2009) Concept-based video retrieval. FTINR 2(4):215–322
49. Snoek C, van de Sande K, de Rooij O, Huurnink B, van Gemert J, Uijlings J, He J, Li X, Everts I, Nedović V, van Liempt M, van Balen R, Yan F, Tahir M, Mikolajczyk K, Kittler J, de Rijke M, Geusebroek J, Gevers T, Worring M, Smeulders A, Koelma D (2008) The MediaMill TRECVID 2008 semantic video search engine. In: Proc. of TRECVID workshop
50. Snoek CGM, van de Sande KEA, de Rooij O, Huurnink B, Uijlings JRR, van Liempt M, Bugalho M, Trancoso I, Yan F, Tahir MA, Mikolajczyk K, Kittler J, de Rijke M, Geusebroek JM, Gevers T, Worring M, Koelma DC, Smeulders AWM (2009) The MediaMill TRECVID 2009 semantic video search engine. In: Proc. of TRECVID workshop
51. Snoek CG, Freiburg B, Oomen J, Ordelman R (2010) Crowdsourcing rock n' roll multimedia retrieval. In: Proc. of ACM international conference on multimedia (ACM MM), pp 1535–1538
52. Snoek CGM, van de Sande KEA, de Rooij O, Huurnink B, Gavves E, Odijk D, de Rijke M, Gevers T, Worring M, Koelma DC, Smeulders AWM (2010) The MediaMill TRECVID 2010 semantic video search engine. In: Proc. of TRECVID workshop. Gaithersburg, USA
53. Stewart J, Bederson BB, Druin A (1999) Single display groupware: a model for co-present collaboration. In: Proc. of CHI conference on human factors in computing systems (CHI), pp 286–293
54. Taksa I, Spink A, Goldberg R (2008) A task-oriented approach to search engine usability studies. J Softw (JSW) 3(1):63–73
55. Tse E, Greenberg S, Shen C, Forlines C (2007) Multimodal multiplayer tabletop gaming. Computers in Entertainment (CIE) - Interactive TV, Association for Computing Machinery ACM Press 5(2). doi:10.1145/1279540.1279552
56. Twidale MB, Nichols DM, Twidale MB, Nichols DM (1998) Designing interfaces to support collaboration in information retrieval. Computers 10:177–193
57. Ullmer B, Ishii H, Jacob RJK (2003) Tangible query interfaces: physically constrained tokens for manipulating database queries. In: Proc. of INTERACT'03, pp 279–286
58. U.S. Department of Health and Human Sciences (2006) Research-based web design & usability guidelines. Available at: www.usability.gov/guidelines/
59. van Velsen L, König F, Paramythis A (2009) Assessing the effectiveness and usability of personalized internet search through a longitudinal evaluation. In: Proc. of 6th workshop on user-centred design and evaluation of adaptive systems (UCDEAS)
60. Yang J, Jiang YG, Hauptmann AG, Ngo CW (2007) Evaluating bag-of-visual-words representations in scene classification. In: Proc. of int'l workshop on multimedia information retrieval (MIR)
61. Yuen J, Russell B, Liu C, Torralba A (2009) Labelme video: building a video database with human annotations. In: Proc. of int'l conference on computer vision (ICCV), pp 1451–1458
62. Zhang J, Marszałek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vis (IJCV) 73:213–238

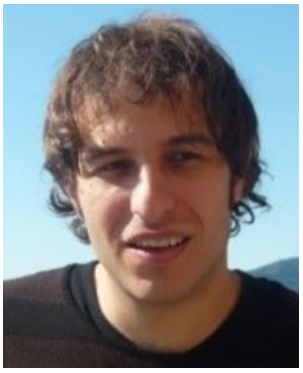
Marco Bertini is an assistant professor in the Department of Systems and Informatics at the University of Florence, Italy. His research interests include content-based indexing and retrieval of videos and Semantic Web technologies. Bertini has a PhD in electronic engineering from the University of Florence. Contact him at bertini@dsi.unifi.it.

Alberto Del Bimbo is a full professor of computer engineering at the University of Florence, Italy, where he is also the director of the Master in Multimedia Content Design. His research interests include pattern recognition, multimedia databases, and human–computer interaction. He has a laurea degree in electronic engineering from the University of Florence. Contact him at delbimbo@dsi.unifi.it.

Andrea Ferracani graduated in Literature from the University of Florence in 2001. Postgraduated from the Master in Multimedia Content Design of the University of Florence in 2004, he's currently working as researcher in the Visual Information and Media Lab at the Media Integration and Communication Center, University of Florence. His research interests focus on web applications.



Lea Landucci received her master laurea in computer engineering at University of Florence in July 2004. Her thesis was on "Multiuser natural interaction environments". Now she's a postdoc working at Media Integration and Communication Centre. Her main interests are on computer vision, interaction design and cognitive psychology and their influence on natural HCI systems.



Daniele Pezzatini graduated in computer engineering at University of Florence in April 2010. His thesis was on "Multi-touch environment for semantic search of multimedia contents". Currently he's working at the Visual Information and Media Lab at Media Integration and Communication Center. His main research interests focus on natural interfaces and environments, web languages and usability.