Interactive Video Search and Browsing Systems

Marco Bertini, Alberto Del Bimbo, Andrea Ferracani, Daniele Pezzatini
Università degli Studi di Firenze - MICC, Italy
{bertini,delbimbo,ferracani,pezzatini}@dsi.unifi.it

Abstract

In this paper we present two interactive systems for video search and browsing; one is a web application based on the Rich Internet Application paradigm, designed to obtain the levels of responsiveness and interactivity typical of a desktop application, while the other exploits multi-touch devices to implement a multi-user collaborative application. Both systems use the same ontology-based video search engine, that is capable of expanding user queries through ontology reasoning, and let users to search for specific video segments that contain a semantic concept or to browse the content of video collections, when it's too difficult to express a specific query.

1. Introduction

Video search engines are the result of advancements in many different research ares: 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 [12], 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 [10, 11]. These systems are generally desktop applications or have simple web interfaces that show the results of the query as a ranked list of keyframes [7, 12], 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

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 10 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: around 75% of the interviewees considered this "mandatory" and around 20% considered it "desirable". 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 interaction. We argue however, that multi-touch and multi-user interfaces may be effective in certain production environments [8] 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 [13]; in these cases a web based interface that forces people to work separately on different screens may not be the most efficient tool.

Regarding video annotation, some manual annotation tools have become popular also in several mainstream video sharing sites such as YouTube and Viddler, because 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.

[1] http://www.vidivideo.info
[2] http://www.im3i.eu

978-1-61284-433-6/11/$26.00 ©2011 IEEE

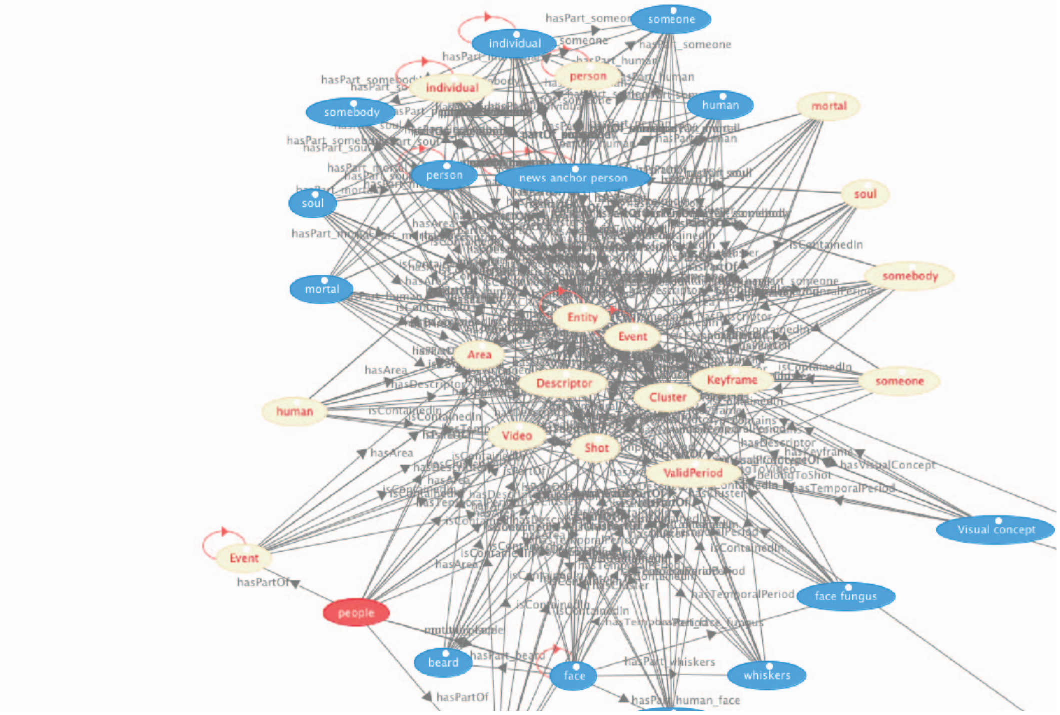

Figure 1. Automatically created ontology structure: view of the relations between some concepts related to "people".

In this paper we present an integrated system composed by *i)* a video search engine that allows semantic retrieval by content for different domains (possibly modelled with different ontologies) with query expansion and ontology reasoning; *ii)* web-based interfaces for interactive query composition, archive browsing, annotation and visualization; *iii)* a multi-touch tangible interface featuring a collaborative natural interaction application.

2. The System

The search engine The Orione search engine 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. 1. The ontology is modelled following the Dynamic Pictorially Enriched Ontology model [3], 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 required 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.

Automatic video annotations are provided by a system based on the bag-of-words approach, that exploits SIFT, SURF and MSER visual features and the Pyramid Match Kernel [5]; these annotations can be complemented by manual annotations added through a web-based interface.

The web-based user interface The web-based Sirio search system[3], based on the Rich Internet Application paradigm (RIA), does not require any software installation,

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

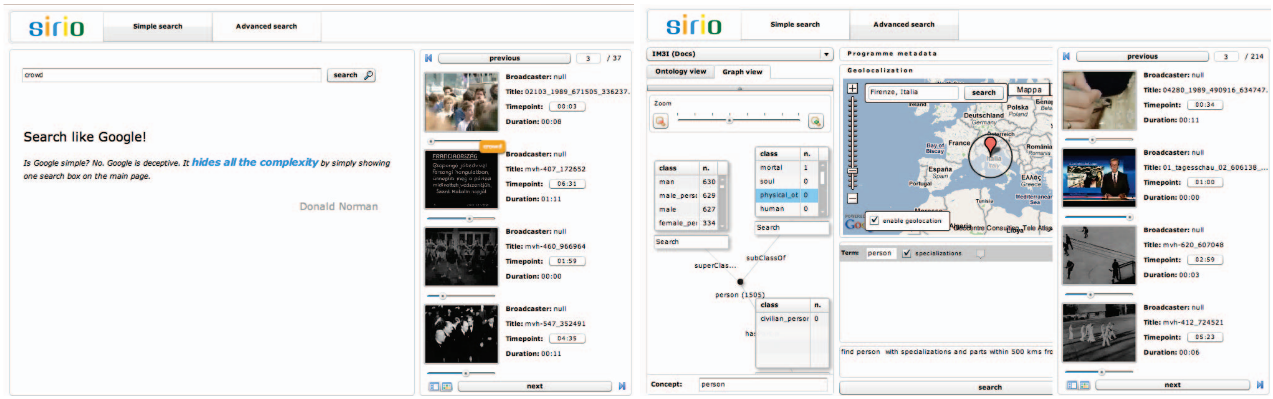

Figure 3. The Sirio web-based user interfaces: *left)* simple user interface; *right)* advanced user interface with ontology graph and geo tagging.

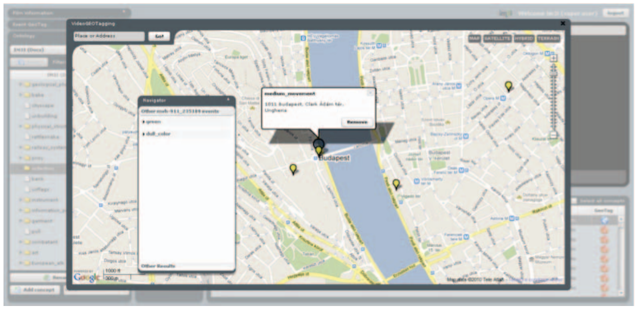

Figure 2. Manual video annotation system: adding geographical annotations to a visual concept

is extremely responsive. It is complemented by a system for manual annotation of videos (Fig. 2), developed with the aim of creating, collaboratively, manual annotations e.g. to provide geographical annotations and ground truth annotations that can be used for correcting and integrating automatic annotations or for training and evaluating automatic video annotation systems. Sirio is composed by two different interfaces shown in Fig. 3: a simple search interface with only a free-text field for Google-like searches, 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 GUI interface allows also to inspect and use a local view of the ontology graph, when building queries, to better understand how one concept is related to the others and thus suggesting to the users possible changes of the composition of the query.

For each result of the query it is shown the first frame of the video clip. These frames are obtained from the video streaming server, and are shown within a small video player. 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. Furthermore another interface (Andromeda), 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.

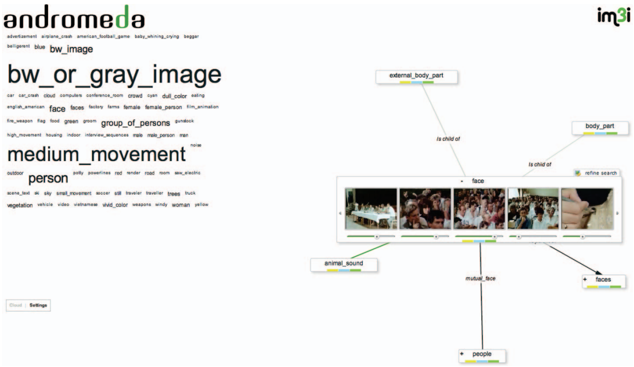

Figure 4. Browsing videos: select concepts in the tag cloud, then browse the video archive or start a search from a concept of the ontology.

This 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 concepts used as annotations (Fig. 4). Users can select a concept from the ontology graph to build a query in the advanced search interface at any moment.

The tangible user interface MediaPick is a system that allows semantic search and organization of multimedia contents via multi-touch interaction. It has an advanced user centered interaction design, developed following specific usability principles for search activities [1], which allows users collaboration on a tabletop [9] about specific topics that can be explored, thanks to the use of ontologies, 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. Some interviews to potential end-users have been conducted with the archivists of RAI, the Italian public broadcaster, in order to study their workflow and collect suggestions and feedbacks; at present RAI journalists and archivists can search the corporate digital libraries through a web-based system. This 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 find. 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. 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.

The user interface adopts some common visualization principles derived from the discipline of Information Visualization [4] and is equipped with a set of interaction functionalities designed to improve the usability of the system for the end-users. The GUI consists of a concepts view (Fig. 5, top), to select one or more keywords from an ontology structure and use them to query the digital library, and a results view (Fig. 5, bottom), which shows the videos returned from the database, so that the user can navigate and organize the extracted contents.

The concepts view consists of two different interactive elements: the ontology graph (Fig. 5, top), to explore the concepts and their relations, and the controller module (Fig. 5, bottom), to save the selected concepts and switch to the results view. 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 present 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 to select the archived concepts and launch the query against the database. The returned data are then displayed within a horizontal list (Fig. 5 bottom), in a new visual component placed at the top of the screen, which contains returned videos and their related concepts. Each video element has three different states: idle, playback and information. In the idle state the video is represented 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, quality, duration, etc.) over the video.

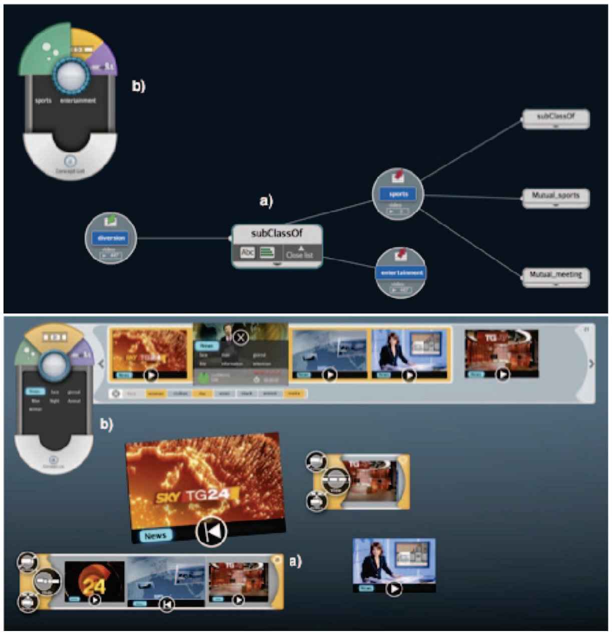

Figure 5. MediaPick system: *top)* **concepts view: a) the ontology graph; b) controller module;** *bottom)***results view: a) video players and user's result picks; b) controller module and results list**

At the bottom of the results 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.

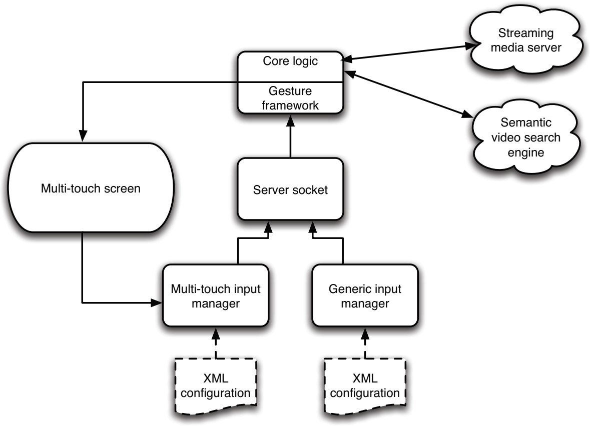

Figure 6. Multitouch interface system architecture

3. Architecture

The system backend and the search engine are currently based on open source tools (i.e. Apache Tomcat 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 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. 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 are developed within IM3I and euTV projects. Audio-visual concepts are automatically annotated using either IM3I and euTV automatic annotation engines. 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 interface and the multi-touch interface have been developed in Flex+Flash, according to the Rich Internet Application paradigm.

Multitouch MediaPick exploits a multi-touch technology chosen among various approaches experimented in our lab since 2004 [2]. 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 can detect fingers and objects on its surface and sends information about touches using the TUIO [6] 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 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. 6). 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. 6). This module is able to manage the events sent by the input manager, translating them into commands for the gesture framework and 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 like an active touchable area; for each active area a set of gestures is defined for the interaction, thus it is useful to link each touch with 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. All the user interface actions mentioned above are triggered by natural gestures shown in the Tab. 1.

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
Drag	- Move video element
Two-fingers drag	- Resize video - Group videos

Table 1. Gesture/Actions association for the multi-touch user interface

4. Usability test

The video search engine (Sirio+Orione) has been tested to evaluate the usability of the system, in a set of field trials. A group of 19 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 queries and activities, and interacting with the system. The methodology used follows the practices defined in the ISO 9241 standard, and gathered: *i)* observational notes taken during test sessions by monitors, *ii)* verbal feedback noted by test monitor and *iii)* an online survey completed after the tests by all the users. The observational notes and verbal feedbacks of the users have been analyzed to understand the more critical parts of the system and, during the development of the IM3I project, have been used to redesign the functionalities of the

978-1-61284-433-6/11/$26.00 ©2011 IEEE

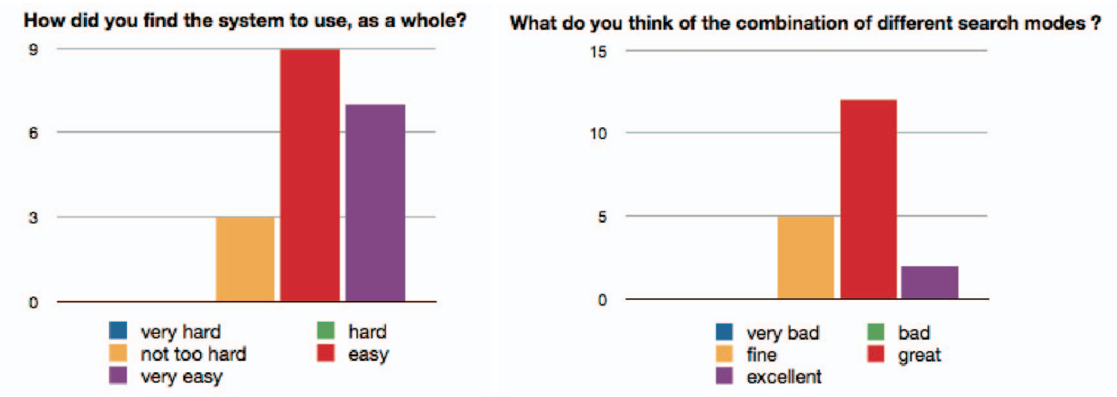

Figure 7. Overview of usability tests: overall usability of the system, usability of the combination of search modalities.

system. Fig. 7 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 interaction that proved to be more suitable for the majority of the users is the advanced interface, because of its many functionalities.

5. Conclusions

In this paper we have presented two semantic video search systems based on web and multi-touch multi-user interfaces, developed within three EU funded research projects and tested by a large number of professionals in real-world field trials. Our future work will deal with further development of the interfaces, especially considering the new HTML5 technologies, extensive testing of the tangible user interface and thorough comparison of the two systems.

Acknowldgements The authors thank Giuseppe Becchi for his work on software development. This work was partially supported by the projects: EU IST VidiVideo (www.vidivideo.info - contract FP6-045547), EU IST IM3I (http://www.im3i.eu/ - contract FP7-222267) and EU IST euTV (http://www.eutvweb.eu/ - contract FP7-226248).

References

[1] R. S. Amant and C. G. Healey. Usability guidelines for interactive search in direct manipulation systems. In *Proc. of the International Joint Conference on Artificial Intelligence*, volume 2, pages 1179–1184, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[2] S. Baraldi, A. Bimbo, and L. Landucci. Natural interaction on tabletops. *Multimedia Tools and Applications (MTAP)*, 38:385–405, July 2008.

[3] M. Bertini, A. Del Bimbo, G. Serra, C. Torniai, R. Cucchiara, C. Grana, and R. Vezzani. Dynamic pictorially enriched ontologies for digital video libraries. *IEEE MultiMedia*, 16(2):42–51, Apr/Jun 2009.

[4] S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. *Readings in information visualization: using vision to think.* Morgan Kaufmann Publishers Inc., 1999.

[5] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research (JMLR)*, 8:725–760, 2007.

[6] M. Kaltenbrunner, T. Bovermann, R. Bencina, and E. Costanza. Tuio a protocol for table-top tangible user interfaces. In *Proc. of International Workshop on Gesture in Human-Computer Interaction and Simulation*, 2005.

[7] A. Natsev, J. Smith, J. Tešić, L. Xie, R. Yan, W. Jiang, and M. Merler. IBM Research TRECVID-2008 video retrieval system. In *Proc. of TRECVID Workshop*, 2008.

[8] C. Shen. 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*, 2006.

[9] C. Shen. From clicks to touches: Enabling face-to-face shared social interface on multi-touch tabletops. In D. Schuler, editor, *Online Communities and Social Computing*, volume 4564 of *Lecture Notes in Computer Science*, pages 169–175. Springer Berlin / Heidelberg, 2007.

[10] A. Smeaton, P. Over, and W. Kraaij. High-level feature detection from video in TRECVid: a 5-year retrospective of achievements. *Multimedia Content Analysis, Theory and Applications*, pages 151–174, 2009.

[11] C. Snoek, K. van de Sande, O. de Rooij, B. Huurnink, J. van Gemert, J. Uijlings, J. He, X. Li, I. Everts, V. Nedović, M. van Liempt, R. van Balen, F. Yan, M. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J. Geusebroek, T. Gevers, M. Worring, A. Smeulders, and D. Koelma. The MediaMill TRECVID 2008 semantic video search engine. In *Proc. of TRECVID Workshop*, 2008.

[12] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. R. R. Uijlings, M. van Liempt, M. Bugalho, I. Trancoso, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.-M. Geusebroek, T. Gevers, M. Worring, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2009 semantic video search engine. In *Proc. of TRECVID Workshop*, November 2009.

[13] E. Tse, S. Greenberg, C. Shen, and C. Forlines. Multimodal multiplayer tabletop gaming. *Computers in Entertainment (CIE)*, 5(2), April 2007.