Use of Content Analysis Tools for Visual Interaction Design

Nazlena Mohamad Ali[1], Hyowon Lee[2] and Alan F. Smeaton[2]

[1]Institute of Visual Informatics, University Kebangsaan Malaysia, Malaysia
[2]CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland
nma@ftsm.ukm.my, {hlee,asmeaton}@computing.dcu.ie

Abstract. Automatic media content analysis in multimedia is a very promising field of research bringing in various possibilities for enhancing visual informatics. By computationally analysing the quantitative data contained in text, audio, image and video media, more semantically meaningful and useful information on the media contents can be derived, extracted and visualised, informing human users those facts and patterns initially hidden in the bit streams of data. Insights into how to transform the emerging technological possibilities from these media analysis tools into usable visual interfaces to help people see visual information in novel ways will be an important contribution to visual informatics. In this paper, we outline some of the more promising content analysis techniques currently being researched in multimedia and computer vision and discuss how these could be used to develop visually-oriented end-user interfaces that support searching, browsing and summarization of the media contents in various usage contexts. We illustrate this with a few example applications that we have developed over the years, all of which designed in such a way as to take advantage of the automatic content analysis and to discover and create novel usage scenarios of consuming visually-oriented media contents.

Keywords: content analysis, visual interaction design, visualization, video browsing

1 Introduction

Automatic media content analysis tools and techniques in the field of multimedia are becoming more and more useful as their accuracy and robustness increase as the result of on-going research effort in this area. An emerging research challenge for interaction and visualization community from this is the question of how to take full advantage of the outcomes of these content analysis methods in designing feasible and usable applications. A myriads of technical possibilities in media content analysis bring benefits to users in helping interpret and navigate those rich media in ways that did not exist before. Some of these experimental applications in this line of work went as far as being evaluated most commonly in the back-end technique or

algorithm level performance or in user evaluation in a controlled lab setting with group of sample users, but increasingly in the form of deployment via the Web or via App store in order to reach wider users and to ascertain more realistic usage feedback over long period of time (for example, SportsAnno [4] and NewsBlaster [10] as web-based campus-wide deployments, iTV system [1] on a cable TV network, and PocketNavigator [16] on Android app market). Taking advantage of these automatic content-based techniques, numerous possible real world application scenarios can be imagined and a large number of novel demonstration systems in varying degrees of completeness have been built with aim to emphasize on user interaction (see [11] for more discussion on this aspect). We expect that this direction of research will bring a new perspective into variety of research agenda in visual informatics that extends interaction and communication based on visual interfaces thus could support user in their understanding, knowledge acquisition and sharing.

In this paper, we take a few promising content analysis tools and techniques in multimedia as examples of soon-to-be some of the core back-end engines of many media applications in the future, and illustrate how we developed visual interfaces that particularly take advantage of these tools to guide and shape the future usage of these applications. A number of novel application examples will be introduced in order to demonstrate the ways in which technical advancements are interpreted and represented as user-centred, visually-oriented end-user features.

The paper is organized as follows. In the next section, a selection of multimedia content analysis techniques is introduced that are continuously researched and developed today. In Section 3, some end-user applications developed within our centre highlighting the use of content analysis in its interaction and visualization design are described, and in particular, we explain in each of these applications how they try to leverage the power of these content analysis in such a ways as to support visually oriented user interaction. We end with a conclusion for summarizing the trend of R&D effort in this area and the future work.

2 Multimedia Content Analysis

In this section we introduce some of the influential and potentially high-impacting content analysis techniques that are currently developing in the multimedia research field, some combinations of which could be used to design visually-oriented end-user applications. Many of these tools are currently being investigated and refined in multimedia and computer vision laboratories, with varying stages of their overall development.

Shot Boundary Detection (SBD) and *Keyframe Extraction* have main goal of segmenting broadcast news into individual camera shots and finding most accurate visual representation of each shot, enabling further structuring of the video content and other operations, thus serving as essential precursors to video indexing and retrieval [5]. A video *shot* refers to the basic unit of retrieval – a short, coherent video sequence that serves as the starting point for the semantic analysis and structuring of content – and relates directly to camera shooting boundaries within the video sequence. Most SBD approaches involve measuring the visual similarity between adjacent or near-adjacent video frames and if these are visually similar, within some

threshold value, then it is likely that they belong to the same shot [19]. Camera and/or object motion will cause adjacent frames to be slightly dissimilar while shot cuts will generally cause a noticeable increase in this dissimilarity. Sudden changes in shots or hard-cuts are relatively easy to identify.

Once a video sequence has been segmented into a sequence of shots, representative keyframes can be extracted for each shot as keyframes serve as a visual shot summary. What constitutes a *good* keyframe and how to recognize a good keyframe remains an open question and current techniques generally rely on a variety of heuristic methods. For example, the simplest approach often involves selecting the first, middle, or last frame in the shot. More sophisticated techniques, however, can take account of in-shot camera movement by selecting the frame where camera movement stops.

Reliable and accurate shot boundary detection and keyframe extraction techniques are considered an "enabling technology" that can be used as the first step for more sophisticated and advanced administration of content analysis tools to media contents, and so far have led to the development of a number of systems that focus on providing users with more effective sequence navigation and shot browsing features.

Another research on content analysis is related to *face detection and recognition*. Because much of our media consumption is based on people directly known to us (e.g. friends and family) or those from media (e.g. celebrity and politicians), the content analysis tools that automatically detect faces and label their names appearing in the media contents have been considered valuable and have a relatively long history of research. Many of currently researched face detection algorithms are based on classifiers where the system is trained using example face data then the regions of image contents are transformed into feature vectors and determined whether the region falls into the 'face exists' category or not. Detected faces are then compared to each other in terms of their feature vector similarity in order to establish/predict whether two faces are of the same person. While working reasonably well for domain-constrained environment (e.g. comparing frontal faces of crime suspects from a convicts database), the overall accuracy and reliability in general context is not high enough to be used in an unsupervised setting (e.g. comparing faces in Flickr database). Additional contextual cues such as 'body patch', location or time could dramatically enhance the reliability of face detection or recognition algorithms [14].

Other work on content analysis is *scene detection and classification.* Depending on the genre of the media content, different techniques optimised for that particular genre can be used. For example, *news story segmentation* is a special case of scene detection applied to broadcast TV news. TV news is typically very well-structured typically starting with a short highlight of the day's news followed by an anchorperson(s) introducing the first news story then a reporter appearing from the scene then returning back to the anchorperson(s) to start the second news story, and so on. Using some of these news broadcast conventions and a combination of other related detection methods such as shot boundary detection and face detection), a TV news programme can be automatically segmented into individual news stories. *Topic detection tracking (TDT)* then tries to identify the topics in the segmented news stories and establish similarity and relatedness amongst the stories.

The scene detection and classification in a generic video content such as movies try to detect *events* based on a number of audiovisual features from movie creation

principles [8,9]. These features were used in the event class detections such as a description of the audio content, where the audio (speech or music) are placed into a specific class; measurement of the amount of camera movement; measurement of the amount of motion in the frame; measurement of the editing pace; and measurement of the amount of shot repetition. For example, work has been done using an approach to detect events in a movie and classified them into three classes based on film grammar as follows:

- *Dialogue*— contains a conversation among characters (one or more people)
- *Exciting* — contains something exciting for the audience (car chase, fighting etc.)
- *Montage* — contains strong musical background as in montage, emotional and musical events

In making a movie, a director follows a certain universal film grammar. For example, he will use a static camera to give the audience a low distraction, relaxed viewing-mode and to give more focus. On the other hand, faster pace editing and high level of camera movement can be used to create an exciting feeling for viewers, give high impact and increased stimulation levels. In addition, background music is used a lot as a medium for creating an emotional response among the viewers. Based on these criteria, a summary of certain measurements are used as the basis in the scene detections.

A work on *concept detection* and *activity identification* focusing on identifying what objects, concepts, events or activities are shown or happening in the visual media contents (e.g. explosion, nature, sunset, people eating food, shopping, etc.) is a difficult challenge because what appears in visual media may not even be agreed by two human viewers or indexers with possibly very ambiguous and subjective interpretation at work in the process. However, using machine learning algorithms to train the system with positive examples (i.e. those parts of video that contain a concept X) and negative examples (i.e. those that don't contain it), reasonable accuracy can be achieved depending on the training data size and the nature of the contents. In the annual TRECVid which aims to advance the automatic indexing of digital video contents, detection of high-level concepts such as "people marching on the street" or "airplane taxing in the airport" have been exercised for a number of years, pushing the level of accuracy each year.

Study of Lifelogging also have utilised concept detection within Lifelog data such as photo collection generated by passive capture devices such as SenseCam [6]. The SenseCam is a small wearable personal device which automatically captures up to about 2,500 images per day. This yields a very large personal collection of images, or in a sense a large visual diary of a person's day. Automatic and intelligent techniques are necessary for effective structuring, searching and browsing of such a large image set for locating important or significant events in a person's life [6]. The identification of the activity can be detected by machine learning techniques, we can classify the lifelog photos in terms of what they contain for example eating, car, people, indoor, shopping or driving.

The techniques described in this section are, even though bearing an enormous potential for future exploitation, simply a selection of technical content analysis without direct link to its potential usage scenarios and possible real-world

applications. Because not many of these techniques are currently used in existing applications, how to channel these technical possibilities into feasible and usable end-user visualization tools is a significant question. In the next section, we introduce some of our end-user applications that leverage the power of these content analysis and touch on how the various techniques could be used to visualize the results of these content analysis in user-centred way.

3 Using Content Analysis for Visual Interaction

We believe that in order to develop sound, practical and useful new media applications, the perspective from multimedia technology should not be used alone in terms of its progress and experimentation but be combined and balanced with conventional and established work practices of use and the way human users have been carrying out work tasks. This is because the introduction of a new technology should be used to enhance rather than completely overturn established work methods. The discipline of Human-Computer Interaction and especially a series of techniques, for example in usability engineering [13] has been developed in order to identify existing practices from the end-user point of view and to then guide the development of new (interfaces) technology into established work practice.

We now present a variety of example of visual interfaces that leverage some combinations of the content analysis techniques as introduced in the previous section and explain how those various tools could be used to visualize the results of the content analysis.

3.1 Video Browsing with Fischlar System

The Fıschlar Digital Video Library System was developed to support capture, indexing, browsing, searching and summarising of digital video and has been deployed into four separate video content collections for a variety of users and application scenarios [18]. The four versions of the Fıschlar system include TV programs [7], TV news [17], TRECVid video track participation [2] and nursing educational videos [3], each designed to support different content type. Information provided to users in the system interface are based on finding and selecting a video program either using text or metadata. Supported interface elements included a keyframe slideshow, a hierarchical keyframe browser, and a timeline browser.

Fıschlar-News was one of the collections designed to support an archive to the main evening TV news broadcast. It incorporates a number of multimedia and recommendation techniques and was deployed within a University campus for several years, in which the large scale testing and evaluation (performance and usability) has been carried out [17]. Methods used from video content analysis include shot boundary detection, keyframe extraction, news story segmentation, capture of closed captions, and the system allows for text searching, browsing and playback based on news stories. An example of an interface screen shot is depicted in Fig. 1.

Fig. 1. Fischlar Interface

Using the output of news story segmentation, shot boundary detection and keyframe extraction directly on the user-interface, this application was designed with the expectation that the accuracy and robustness of these content analysis techniques will become high enough in near future so that those detected and segmented units could directly be mapped into visual representation elements on the screen.

3.2 Managing Personal Photo Collection with MediAssist

The MediAssist [15] is a novel application that incorporated a number of content analysis to enable users to efficiently search their personal photo archives. Automatically generated contextual metadata and content-based analysis tools (face and building detection) are used, and semi automatic annotation techniques allow the user to interactively improve the automatically generated annotations. Our retrieval tools allow for complex query formulation for personal digital photo collection management. Fig. 2 shows the interface of the application. The user starts with formulating a search query on the left side of the screen, by selecting location and dragging timeline bar as well as time of the year, number of faces in the photos, weather conditions when the photos were taken, and so on. The result of search is displayed on the right side of the screen as a grid of thumbnail photos, and selecting one of the results then presents an enlarged display of the photo as shown in Fig.2.

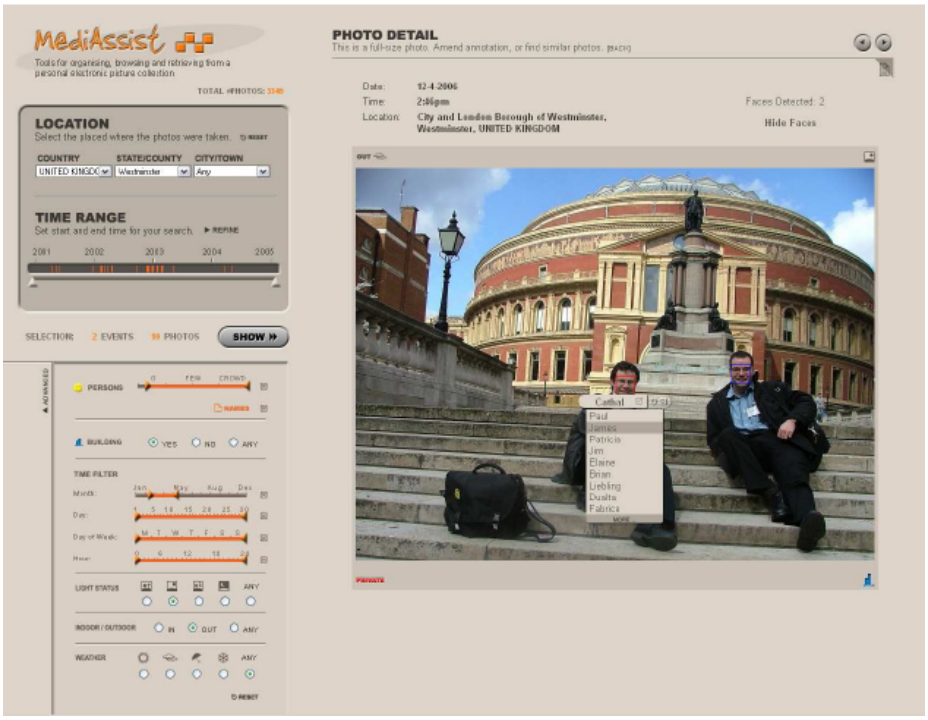

Fig. 2. MediAssist Photo Management System

Recognizing the fact that the detection algorithms incorporated in the system cannot be 100 percent accurate today, a simple user correction mechanism is featured where the user can click on the face of the person and select one of the few predicted names based on the visual similarity of that face within the same event. While content analysis tools are improving steadily, unpredictable photo quality and face orientation, shading, occlusion, and lighting condition inevitably causes incorrect labeling of the faces thus a simple and easy manual correction whenever such an error is noticed by the user becomes an important feature in an application such as this.

3.3 Navigating Scene Types with MOVIEBROWSER2

We developed the MOVIEBROWSER2 [11,12] to incorporate the use of a number of content analysis techniques, particularly those that identify movie scene boundaries and categorise them into exciting, montage and dialogue scenes. MOVIEBROWSER2 uses several recent multimedia technologies to automatically process digital video content but at the same time we used a usability engineering process to relate these techniques to the real tasks of real users in their real environments. The application domain we work in is film studies where students need to study movie contents and analyse movie sequences and was deployed in the university for a duration of a semester. Fig. 3 shows the main interface of MOVIEBROWSER2 application.

When a user selects a movie (i.e. “Shrek” in Fig. 3), the movie’s content is visually presented to the user. In Fig. 3, the whole duration of the movie is represented as 3-band horizontal timeline near the top of the screen, in three different colours, each representing different types of scenes. For example, the green band represents those scenes with *Dialogues*; the pink band represents those scenes with *Montage*; the yellow band represents *Exciting* scenes and it is also noticeable the *Exciting* keyframe scenes appear throughout the movie. Clicking on any of these colour blocks on the band will jump the keyframe list (below the timeline) to that scene, and clicking on the keyframe will start streaming the video from that point onwards.

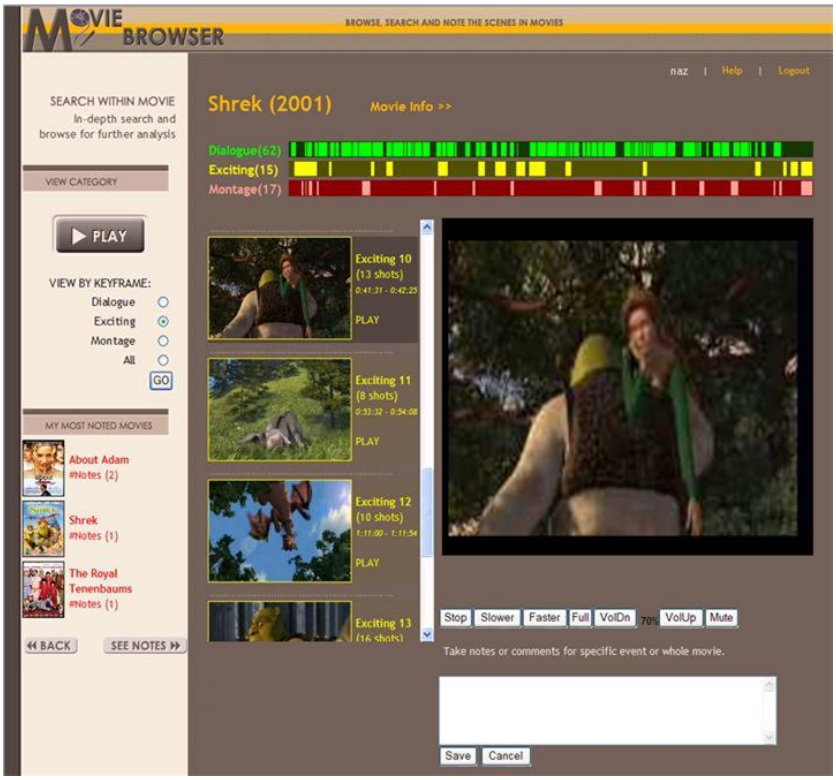

Fig. 3. MOVIEBROWSER2 Main Interface

An experiment has been carried out on how the visual display on automatic content analysis is beneficial to the user as reported in [11]. In this experiment, a film studies students were chosen as a sample users and they were given a task to browse and complete the short essays. The findings from the essay outcome revealed that there are slightly improving or better results which is also supported by the remarks from the module lecturer that shows students have more variability (more opinions, expressions) in their written essay when using MOVIEBROWSER2. Our findings also show that satisfaction levels are higher after using the newly introduced tool with

higher mean scores in all aspects of statements given as compared to when using a conventional standard player.

3.4 Reviewing a Day with My Visual Diary

My Visual Diary, a SenseCam image management system, is an application for SenseCam use that resolves some of the problems of managing the exceedingly large number of SenseCam photos [6]. The system employs a number of content-based image analysis techniques to automatically structure and index the captured photos in such a way that the owner of the photos can easily search and browse the large amount of SenseCam photos through a web-based interface.

Usage scenario involves a user wearing the SenseCam over a long period of time (say a few years), and a few nights a week tries to review what happened during the past week. The main challenge of this kind of Lifelogging scenario is the huge number of photos that have to be managed: typically resulting in 2,000 – 3,000 photos per day, or if worn every day, over 1 million photos per year, going through individual photos and manually tagging or annotating them for future access is out of the question. Using our event detection to group the photos into meaningful events and calculating the visual uniqueness of each event, the application automatically compose an appealing visual montage of a day's happenings. Fig. 4 shows a screen shot where a user selected one particular date from the mini calendar on the top left of the screen, and that day's visual summary is presented on the main part of the screen. Currently 19 most important events of the day are shown, each photo representing a key photo from an event. The size of the photos is in proportion to the uniqueness of the event relative to the rest of the events that happened that day. Thus, for example, the largest photo at the bottom of the screen is the most unique event that happened that day by this user.

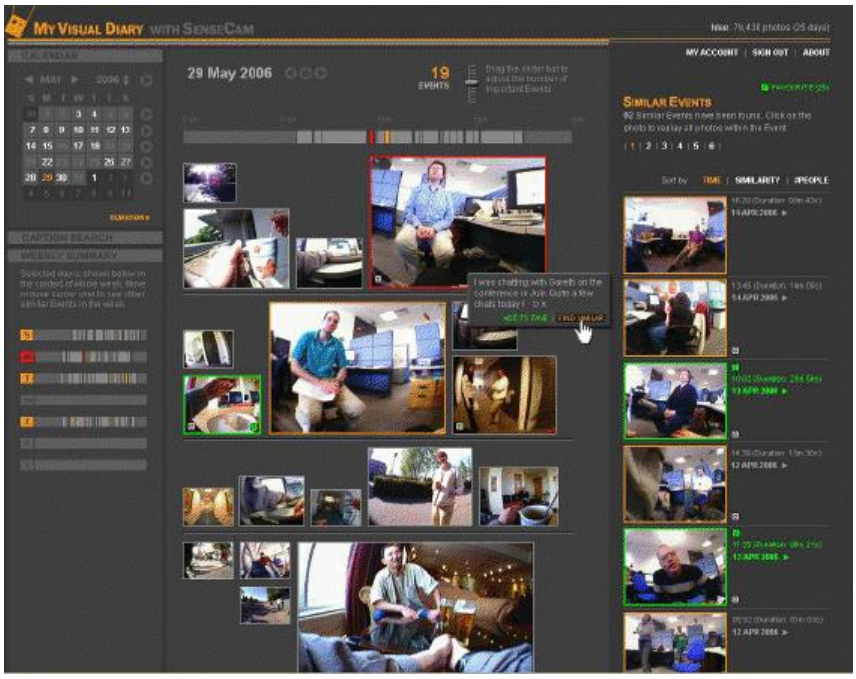

Fig. 4. SenseCam Photo Browser Interface

This application allows a very quick glance of a day, not by intensively flipping through thousands of photos of that day, but by intelligently grouping the photos into events, selecting most representative photo from each event, and identifying more important (or unique) events then presenting them in a static, one page template.

4 Conclusion

Automatic content analysis is one of the most dynamic and technically aggressive research area in multimedia. Most of the techniques being researched in this field have been initially conceived from a computational perspective and technical possibilities rather than from an end-user perspective and their needs thus they generally tend to lack the grounding into the real-world scenarios and situations. However, the potential for exploitation as the understanding of these tools grows and matures is staggering. From the perspective of a visual interaction designer, an automatic detection and content analysis can be a significant advantage to help come up with novel ways of supporting the envisaged end-users by creating visual interfaces that otherwise could not be implemented. While understanding the user perspectives and their requirements is an important element for visual informatics, taking advantage of these emerging content analysis tools and techniques and exploring novel visualization provisions that end-users would normally not ask or demand in their conventional usage context could open up many innovative new ways

of discovering or creating visual interfaces afforded by these emerging tools. Once a novel visualization is designed, then a series of well-established usability engineering methods could be used to see how these new interfaces could be tailored, adapted and customized for the specific wishes and needs of the users.

Our future work includes refining the applications as described in this paper by more rigorously employing the usability engineering methods, especially longitudinal study methods such as diary method and ethnographic studies in order to understand people's adoption and appropriation of the novel applications in their lives. Such a study will reveal many new challenges back to the content analysis streams of research and guide their future directions and agenda, then the technical tools that come out under those agenda will more likely be valuable and useful for incorporating into subsequent new batch of visual applications.

Acknowledgement We would like to thank those involved directly or indirectly on the projects that have been carried out in CLARITY: Centre for Sensor Web Technologies, Dublin City University.

References

1. Bernhaupt, R., Obrist, M., Tscheligi, M.: Usability and usage of iTV services: lessons learned in an Austrian field trial. ACM Computers in Entertainment. Vol 5, no 2 (2007)
2. Browne P., Czirjek C., Gaughan G., Gurrin C., Jones G., Lee H., Marlow S., Mc-Donald K., Murphy N., O'Connor N.E., O'Hare N., Smeaton A.F., Ye J.. Dublin City University Video Track Experiments for TREC 2003. In Proceedings of the TRECVid Workshop, Gaithersburg (2003)
3. Gurrin C., MacNeela P., Smeaton A.F., Lee H., Browne P., McDonald K.. Fıschlar-Nursing, Using Digital Video Libraries to Teach Nursing Students. In WBE 2004 - IASTED International Conference on Web-Based Education, pp 111–116 (2004)
4. Lanagan, J., Smeaton, A.F.: SportsAnno: What Do You Think? In: Proc. of Large-Scale Semantic Access to Content (Text, Image, Video and Sound) RIAO (2007)
5. Lee H., Smeaton A.F., O'Connor N., Smyth B.: User Evaluation of Físchlár-News: An Automatic Broadcast News Delivery System. TOIS - ACM Transactions on Information Systems, Vol. 24, No. 2, pp. 145--189(2006)
6. Lee H., Smeaton A.F., O'Connor N., Jones G., Blighe M., Byrne D., Doherty A., Gurrin C.: Constructing a SenseCam Visual Diary as a Media Process. Multimedia Systems Journal, Special Issue on Canonical Processes of Media Production, Vol. 14, No. 6, pp. 341--349 (2008)
7. Lee H., Smeaton A.F.: Designing the User Interface for the Fıschlar Digital Video Library. Journal of Digital Information, 2(4) (2002)
8. Lehane B., O'Connor N. E., Lee H., Smeaton A.F.: Indexing of Fictional Video Content for Event Detection and Summarisation. EURASIP Journal on Image and Video Processing, 2007:1–15(2007)
9. Lehane B., O'Connor N.E., Smeaton A.F., Lee H.: A System For Event-Based Film Browsing, volume 4326/2006, pp. 334--345 (2006)
10. McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkovo, A., Sable, C., Schiffman, B. Sigelman, S.: Tracking and summarazing news on a daily basis

with Columbia's Newsblaster. In: Proc. of the Human Language Technology Conference, (2002)
11. Mohamad Ali, N., Smeaton A.F.: Are Visual Informatics Actually Useful in Practice: A Study in a Film Studies Context. Lecture Notes in Computer Science, In: Visual Informatics: Bridging Research and Practice, Volume 5857/2009, pp. 811--821 (2009)
12. Mohamad Ali, N., Smeaton A.F., Lee H., Brereton P.: Developing, Deploying and Assessing the Usage of a Movie Archive System. Lecture Notes in Computer Science, In: Human-Computer Interaction. Interacting in Various Application Domains, Volume 5613/2009, pp. 567--576 (2009)
13. Nielsen J.. Usability Engineering. Academic Press, Inc., (1993)
14. O'Hare N., Smeaton A.F.: Context-Aware Person Identification in Personal Photo Collections. IEEE Transactions on Multimedia, Special Issue on Integration of Context and Content for Multimedia Management, Vol 11: Issue 2, pp. 220--228 (2009)
15. O'Hare N., Lee H., Cooray S., Gurrin C., Jones G., Malobabic J., O'Connor N., Smeaton A.F., Uscilowski B.: MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections. CIVR2006 - 5th International Conference on Image and Video Retrieval. Springer Lecture Notes in Computer Science, Vol. 4071 / 2006, pp. 529--532 (2006)
16. Pielot, M., Boll, S.: Tactile Wayfinder: Comparison of Tactile Waypoint Navigation with Commercial Pedestrian Navigation Systems. The 8th International Conference on Pervasive Computing (2010)
17. Smeaton A.F., Gurrin C., Lee H., McDonald K., Murphy N., O'Connor N.E., O'Sullivan D., Smyth B., Wilson D.: The Fıschlar-News-Stories System: Personalised Access to an Archive of TV News. In RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, pp. 3--17 (2004)
18. Smeaton A.F., Lee H., McDonald K.: Experiences of Creating Four Video Library Collections with the Fıschlar System. International Journal on Digital Libraries, 4(1):42–44 (2004)
19. Smeaton A.F., Over P., Doherty A.: Video Shot Boundary Detection: Seven Years of TRECVid Activity. Computer Vision and Image Understanding, Vol. 114, No. 4, pp. 411--418 (2010)