A Novel System for Interactive Live TV

Stefan M. Grünvogel[1], Richard Wages[1], Tobias Bürger[2], and Janez Zaletelj[3]

[1] Cologne University of Applied Sciences, IMP, Betzdorferstr. 2,
50679 Cologne, Germany
{richard.wages,stefan.gruenvogel}@fh-koeln.de
[2] Salzburg Research Forschungsgesellschaft m.b.H., Jakob-Haringer-Strasse 5/III
5020 Salzburg, Austria
tobias.buerger@salzburgresearch.at
[3] University of Ljubljana, Trzaska 25,
1000 Ljubljana, Slovenia
janez.zaletelj@ldos.fe.uni-lj.si

Abstract. For the production of interactive live TV formats, new content and new productions workflows are necessary. We explain how future content of a parallel multi-stream production of live events may be created from a design and a technical perspective. It is argued, that the content should be arranged by dramaturgical principles taking into account the meaning of the base material. To support the production process a new approach for content recommendation is described which uses semantic annotation from audio-visual material and user feedback.

Keywords: interactive live television, recommender system, extraction of meaning, dramaturgical principle, content, multi-stream.

1 Introduction and Set-Up

Interactive digital television demands new kinds of content and new kinds of production workflows. TV-on-demand service providers (like "Zattoo" or "Joost") make it possible to watch any movie, series or news broadcast at any time, independent of the original broadcasting. But there are events (like e.g. elections or football matches), where the fascination of watching is based in the "live" character of the broadcast – anything can happen in the moment of the actual broadcast. Within the project "LIVE" [1] our aim is to enable the production of interactive live TV events.

There are already several approaches to create interactive live TV formats (e.g. BBC's interactive broadcasting of the 2004 Olympics). A successful example is multi-stream sports where the consumers at home are able to choose from different video channels. The streams could come from several cameras, filming the event from different angles (e.g. at a football stadium). But also the case where an event consists of several sub-events and each sub-event is filmed by its own cameras is possible.

In the LIVE project [1] it is the aim to support the parallel production of several multi-channels. The creation of the multi-channels is done by the "video conductor" (VC), who may be a single human or a whole team at the production site.

L. Ma, R. Nakatsu, and M. Rauterberg (Eds.): ICEC 2007, LNCS 4740, pp. 193–204, 2007.

The content on which the work of the VC is based can come from live AV-streams or from the broadcasters archive. The VC has to combine and stitch the streams together to produce several AV-streams in real-time (cf. Figure 1) which can be received by the consumers. We term the viewers at home who interact with the iTV program as "consumers" and the people working at the production site as "professional users" or simply "users."

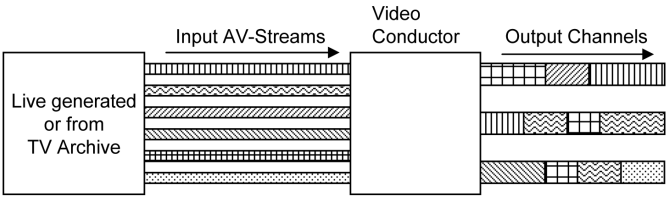

Fig. 1. Creation of multi-channels

Now the interaction issues come into this system in several ways, which distinguishes it from other systems.

The VC does not only produce the live AV-streams, but in addition he will also create transitions points, where the consumer is invited to switch to another channel (cf. Figure 2).

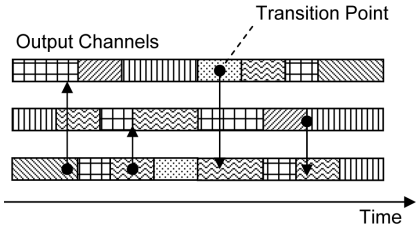

Fig. 2. Multi-channels with transition points

The aim is to prevent a mere "zapping" between channels and instead help the consumer to steer through the content such that an overall dramatic arc persists. How this can be achieved practically is explained in Section 2.

The second type of interaction works by involving both the consumer and the professional users. By using a backchannel from the consumers to the broadcaster, the VC will be able to adapt his style of content stitching, depending on the behaviour of the audience. The technical details of how this works is explained in Section 3.

2 How Will Future Interactive Content Look Like?

2.1 The Need for Dramaturgical Design Principles in Interactive Live TV

When thinking about future interactive content for iTV one has to consider current state-of-the-art approaches to this topic. These approaches can be categorized based

on the interaction design including devices and on-screen interfaces and based on the genres or formats which are transported. As an example, the interaction of the consumer from BBC interactive is accomplished by using the "Red Button" on the remote control, to interact with the program. Various formats like quiz games or voting opportunities are already available.

A lot of work and research has been undertaken to optimize the on-screen consumer interface by e.g. using well-known metaphors from personal computers (like menus, buttons etc.) or to enable the viewer to have a multi-perspective view of an event (e.g. by using split screens, showing different camera views of an event). These approaches work well with almost any content, because they are created to work irrespective of the actual semantic essence of the video feeds.

But: there are currently almost no design principles available to connect different video streams in such a way that consumers have the possibility to switch between these channels *and* to have a consistent dramaturgical concept over all these channels. Thus for the most part current iTV formats fail to generate two consumer experiences: flow and immersion.

To explain this a little further, one example from another medium. In movies, beneath the underlying story, the camera work and the editing of the scenes and shots are crucial for the dramaturgical arc over the whole film. The editor is responsible to collect and compose the raw film material into a consistent entity depending on the story elements been told and the camera representation of these elements. As a result, good editing and storytelling lead to an *immersive* experience for the consumer. The semantic meaning and the representation are crucial for success. This has already been known since the beginning of beginning of the past century. In the famous Kuleshov experiment from 1919 two unrelated images have been shown in juxtaposition to the viewer. Kuleshov discovered, that the viewers interpreted both images as a part of a related sequence. In the experiment a face with neutral facial expression was shown after showing images with different moods. Depending on the preceding image, the same facial expression led to a completely different interpretation by the viewers.

2.2 Levels of Meaning

Thus as a consequence, iTV formats could be enhanced, if different live feeds are composed carefully depending on the *meaning* of content which is presented in this channel. Instead of allowing the consumer to switch randomly between different channels which are not related at all, switching should be stimulated at those points where it makes sense from a dramaturgical point of view. One has to be aware, that "dramaturgical" does not only mean "thrilling" or "exciting", also formats with funny elements need a dramaturgy.

Another important fact is, that one and the same video stream has different meanings for one and the same consumer depending on the context in which it is presented to the consumer. Thus by allowing the consumer to switch between different live channels, the meaning of each channel depends on the other channels. This makes it possible to stitch one live stream into different live formats, each time with at different meaning to the viewer.

In the LIVE project we distinguish between different levels of meaning, reaching from particle to universal statements. The information of the different levels of meaning is extracted from and assigned to the AV material relative to an abstract viewpoint (i.e. an ontology). The execution of this task will be done by an "Intelligent Media Framework" (cf. Section 3.3).

As a starting point we defined the different levels of meaning of an AV object relative to an observer perspective. The four levels are termed "Particle", "Local", "Global" and "Universal".

Example 1. Consider a clip where a ball in a football match crosses the goal line. We could assign the following levels of meaning to this clip:

- **Particle:** Ball crosses line, player kicks ball, goalkeeper misses ball
- **Local:** Team A wins football match, end of career of coach from team B
- **Global:** Sportive competition between opposing teams
- **Universal:** Success depends on skills

One could easily construct similar examples for all kind of situations of live event situations such as e.g. a folded piece of paper put into a box (at elections).

2.3 Dramaturgical Enhanced Interface

An important question is how the interface for such new content for live iTV may look like. It is well known, that the use of a classical "WIMP" (window, icon, menu, pointing device) interactions techniques easily destroys the immersion of a consumer. Thus, although indispensable for many iTV formats, these kinds of interfaces are not suitable for keeping a dramaturgical arc over a longer period of time. In computer games, a similar problem arises because good games should not pull out gamers from the cognitional state of "flow". Current solutions use in-game controls, which connote objects from the game world with a known in-game meaning with an additional meaning of the (real) players world. This technique can not be transferred directly to iTV, because in most cases it will not be possible to add real objects to the physical world of the filmed live events, which could be interpreted by the consumer as a meta object. But the content of a live stream could be used as is to provoke the consumer to switch to another channel. To give a very simple (but drastic) example: Showing an empty pool over minutes during a swimming championship most likely provokes the consumer to switch to another channel – or to switch off.

3 Extraction of Meaning and Content Recommendation

The creation of future interactive content needs (as stated in Section 2) new workflows and new tools. In this section a new workflow satisfying the demands of future interactive content creation is proposed. It is based on the extraction of meaning of content and content recommendations for the video conductor. Also a first prototypical interface for the recommendation system is shown.

3.1 The Concept of Content Recommendations in TV Production

The introduction of interactive digital TV [2], [3] brings new challenges to the existing world of TV production. One of the important new features of digital and interactive technologies is the possibility to produce and transmit programs which are composed of several sub-channels. The exploitation of this feature is a big challenge, because this means that some proven concepts of classical TV production have to be transformed. It requires new tools and new workflows to support the parallel production of several multi-channels. A vital requirement to the tools and workflows is that it should be possible to combine live audio-visual material and also to enable the reuse of already existing content in video archives for live production.

To reuse archived TV material a new content workflow for iTV production has been developed in the LIVE project (cf. Figure 3). This workflow addresses the issue of on-the-fly content selection from broadcaster archives and its reuse during live production. For implementing the content workflow, several items have to be realized:

- Adapt the production workflow so that the professional user (the editor) will be able to search and retrieve material on-the-fly.
- Properly annotate the TV content so that an intelligent search at the fine level of content granularity will be possible (for example, at the level of shots).
- Establishing a digital TV archive in which the content is instantly available for play-out.
- The development of a suitable personalized content selection algorithm.

The recommender system prototype of the LIVE project is aiming at providing a first implementation of the production support system. This system will enable live personalized content selection for the TV production. We have to stress the fact, that the system will recommend content for a professional user (the video conductor) at the production site. Still (and presumably also in the long run) we believe, that humans will have to be responsible for the *creative* assembly of the different streams.

We will now explain the workflow in more detail (cf. Figure 3).

First, the AV content is processed to extract semantic information. This can be done in two ways: either automatically or by human annotation. With human annotation higher levels of meaning (e.g. at universal or global level) can be generated. The results of both approaches are "semantic content annotations" of the audio-visual material. The extracted information enables the generation of "content recommendations". The video conductor receives content recommendations in the form of a list of audio-visual segments matching the search query. Items from this list of AV material are included into the final program by the video conductor. The resulting TV streams are sent to the consumers.

The crucial requirement to produce content recommendation is that the meaning is extracted from raw audio-visual material first and is represented within the predefined ontology.

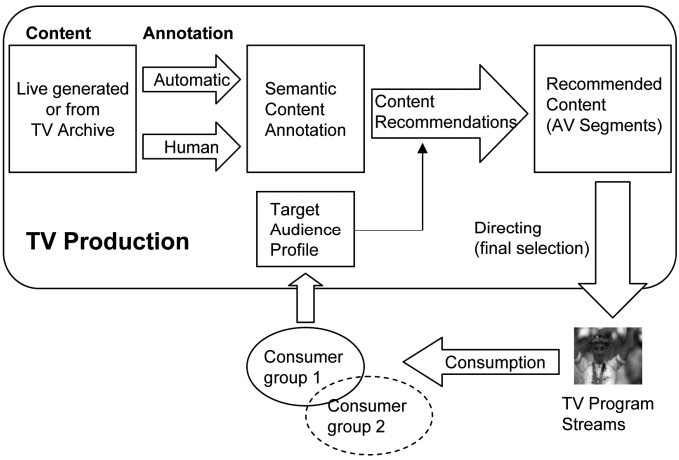

Fig. 3. The content workflow

3.2 Extraction of Meaning from Videos

Extraction of meaning is the process of analysis of videos and in turn the semantic enrichment of the metadata that accompanies its binary representation. Semantics and the interpretation of the content is important to make content machine-processable. This semantic enrichment of content can be done manually, which is expensive, or automatically, which is error-prone. In particular, automatic semantic enrichment must be aware of the gap between the semantics that are directly retrievable from the content and those which can be inferred within a given interpretative context (the "Semantic Gap"). Semantic machine readable metadata of content is beneficial in many ways. Based on earlier investigations [4], semantic descriptions of content can enhance fast and easy navigation through audio-visual repositories for example. It can be used to create content recommendations for a video conductor, who can make decisions based on these recommendations which AV-stream will be played out and which of the streams will be interlinked.

Some of the recent research projects in the area of semantic (or symbolic) video annotation try to derive the semantics from the low level features of the audiovisual material or from other available basic metadata, e.g. by audio-classification of classification of camera movement. Some of the projects aim at highly automated indexing using the results of automatic speech recognition however error-prone they may be. Most of these approaches are - as also pointed out in [5] - not capable to derive the semantics of multimedia content because in many cases the results of the analysis cannot be related to the media context [6]. For humans the construction of meaning is an act of interpretation that has much more to do with pre-existing knowledge (the "context") than with the recognition of low-level-features of the content. This situation is commonly referred to as the "semantic gap" [7].

Two solution paths have emerged for this problem: The first one is to provide rich annotations created by humans as training data for the system to learn features of videos for future automatic content-based analysis. The second approach does not rely on training, but purely on analysis of the raw multimedia content. The training

approach is not well suited for scenarios in which a great amount of content has to be annotated before any training and automation can be done or in which the application domain is very broad. The second approach usually only works well in settings where the relevant concepts can easily be recognized. However, most content based services demand richer semantics. As pointed out in section 4, popular examples on the Web show that there are currently many service-based platforms that make use of their users' knowledge to understand the meaning of multimedia content.

We differentiate between three levels of knowledge to refer to this semantic enriched information about content: This distinction is based on the organization of so-called Knowledge Content Objects [8], a semantic model for the trade of information goods: (1) There is the resource or the video file itself which can uniquely be identified by URIs, then (2) traditional metadata attached to the content like frame rate, compression type or colour coding scheme, and (3) semantic knowledge about the topics (subject) of the content as interpreted by an actor which is realized by the video. The semantic knowledge is further divided into particle, local, global, and universal knowledge (cf. Section 2). This distinction refers to the validity of the knowledge.

In the following section we explain how knowledge comprising to these different levels can be extracted from the raw content with the Intelligent Media Framework.

3.3 Real-Time Extraction of Meaning with the Intelligent Media Framework for the Staging of Live Media Events

In the terminology of the project, “staging live media events” is a notion for the creation of a non-linear multi-stream video show in real-time, which changes due to the interests of the consumer (end user). From a technical viewpoint, this requires a transformation of raw audiovisual content into “Intelligent Media Assets”, which are based on the Knowledge Content Object that were already introduced above in Section 3.2. To extract knowledge on all levels (e.g. particle, local, global, universal) the development of a knowledge kit and a toolkit for an intelligent live content production process including manual annotation and automated real-time annotation is needed (cf. Figure 3).

To design this knowledge kit we applied lessons learnt from automatic approaches to overcome the weaknesses of automatic metadata extraction. We started to design an "Intelligent Media Framework" (IMF) that is taking into account the requirements of real-time video indexing to combine several automatic and manual annotation steps in order to enrich content with knowledge on different levels. The Intelligent Media Framework thereby integrates the following sub-components into one consistent system:

The “*Intelligent Media Asset Information System*” (IMAIS) provides access to services for the storage of media, knowledge models and metadata relevant for the live staging process. It also provides services for the creation and management and delivery of intelligent media assets. The IMAIS will be the central component of the Intelligent Media Framework and will semantically enrich incoming metadata streams with help of incoming manual and automatically derived annotations. The semantic enrichment process in LIVE is twofold: An application for automatic analysis delivers typical video metadata like close-ups, shots, faces, camera-motion, colour schemes,

scenes and artists. This information is enriched in a manual step done by an human agent in the Intelligent Media Framework that has knowledge about the context of the analysed media item, and in the Recommender System (cf. below) which has the knowledge about the user preferences.

The "*Recommender System*" provides content recommendations for the video conductor based on previous user feedback, making use of knowledge. In Section 3.5 the functionalities of the Recommender System for the Video Conductor are presented in more detail.

In the next section (Section 3.4) we will explain how the components of the Intelligent Media Framework work together.

3.4 Content Recommendations Within the TV Production

Content selection within LIVE will primarily focus on the selection of audio-visual materials which may come from the TV archives or from a loop server where the live created material we be stored directly after been annotated by humans. The content selection is made according to the preferences of the target audience.

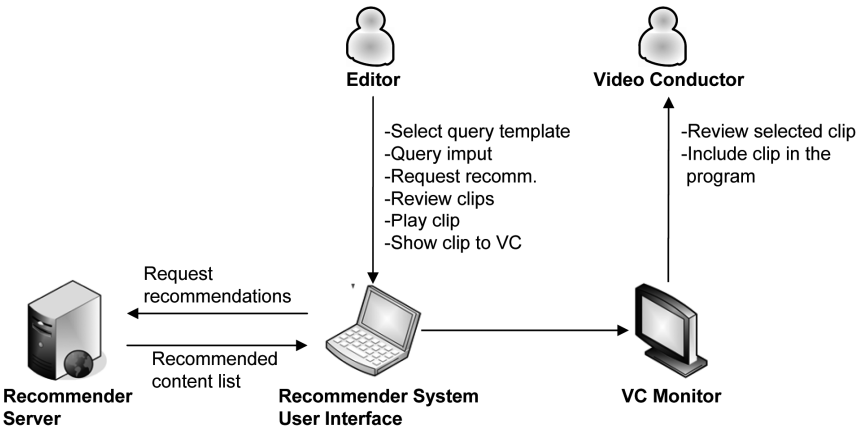

Fig. 4. The Recommender System setup and actions performed by the users

In Figure 4 the proposed components of the content selection workflow are shown. Content selection within LIVE is done in two stages. First, the archived content is processed and recommended within the production workflow. The editor receives "Archive Content Recommendations" in form of a list of audio-visual segments found in the TV archives. Items from this list of AV material are included into the final program by the video conductor. The resulting TV stream is sent to the Consumers.

3.5 Recommender System Functionalities for the Video Conductor

The video conductor and members of his team will be able to benefit from content recommendation functionalities by using the Recommender System user interface. The Recommender System user interface was implemented as a standalone Java

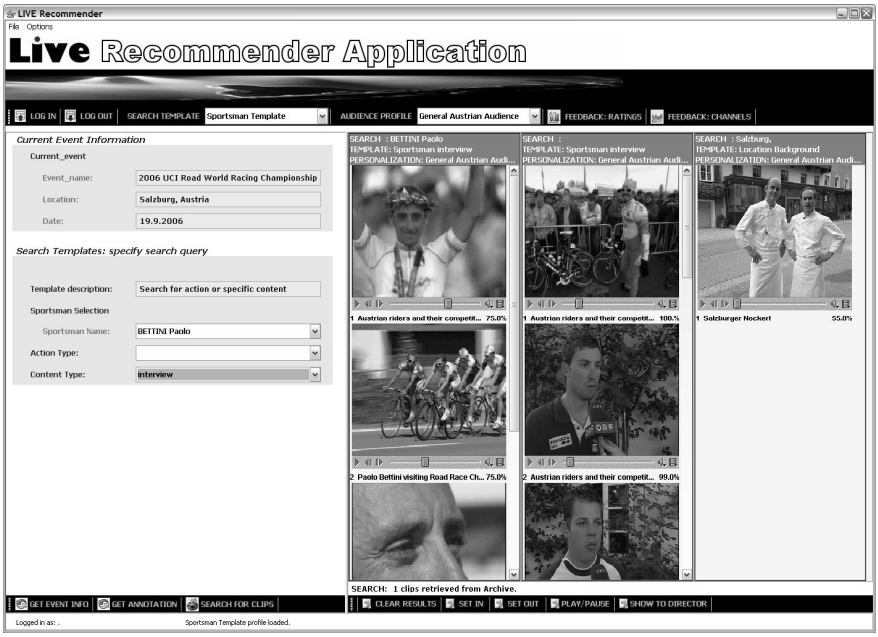

Fig. 5. The Recommender System user interface

application. The user interface connects to the "Recommender Server" and retrieves a recommended content list. This content (AV material) is then displayed in the Recommender System user interface. The editor is able to review the content and prepare a final selection for the video conductor.

The main functionalities of the Recommender System user interface (cf. Figure 5) are the following:

- **Search specification.** The editor is able to select search terms in event-specific fields, for example athlete name, venue, etc. Selected or entered keywords are combined into a query, which is sent to the Recommender Server.
- **Search Templates.** The editor can select a more specific search template which has some predefined knowledge already built in. For example, he can search for interviews, and the single parameter of this search would be the athlete name.
- **Loading of event information.** Pre-existing event information such as athlete lists is retrieved from IMF or loaded from XML file, so that the input fields already contain possible search terms.
- **Display of the recommended clips.** Recommended material can be previewed in low resolution and selected for play-out on the professional equipment.

The editor is using the user interface (cf. Figure 5) to search and retrieve audio-visual materials. The search capabilities apply to all AV material which is properly

annotated and available through the Intelligent Media Asset Information System (IMAIS). This might include TV archive content (if it has been annotated in advance) and also live content which is annotated on-line. We now explain the typical workflow for an editor by example.

Example 2. Suppose that the editor is looking for additional background material on the two leading athletes named Bettini and Hasselbacher. Then the following steps are made:

(1) First, he needs to select a suitable search template which he will use to enter the search terms. Different search templates are available for typical editor's queries. For example, he might select a template named "Sportsman Background":

```
Content Type = {"Portrait", "Interview"}
```

This template already includes pre-specified terms to search for. The only parameter that needs to be specified is the names of the athletes.

(2) The editor now enters an athlete name or selects his name from the start list.

(3) He selects a target audience profile. This means that he is able to select a specific profile of the target group of TV consumers. The audience profile includes preferences of the TV consumers towards different types of content. The effect of the profile is that the resulting list of content will be personalised to the needs of target audience.

(4) The editor requests a content recommendation. The search query is compiled within the user interface and sent to the Recommender Server, which will return a list of matching content (cf. Figure 4). In our example the recommendation query is composed of the following information

```
Recommendation Query = {
         AudienceProfile = "General Austrian audience"
         SearchTemplate = "Sportsman Background"
         SearchTerms = {
            AthleteName = "Bettini"
            AthleteName = "Hasselbacher"
         }
}
```

(5) Finally, the editor will show the selected content to the video conductor, who will preview clip on the professional video monitor (cf. Figure 4). The video conductor will make the final decision if the clip shall be included into the TV program or not.

3.6 Computation of Content Recommendations Within the Recommender System

The goal of the Recommender System is to compile a list of suitable content based on the query and return it to the user interface. This is done in a sequence of steps where different information is added in each step. The proposed method is essentially a content-based filtering approach where content items are selected according to their metadata description.

Example 3. To explain the different steps for the compilation of a list suitable content we explore the search template from Example 2 in more detail.

(1) The search template information is added to the search terms and an `IMF_Query` is generated. This means that the `IMF_Query` now contains the information on the content type, where `ContentType = {Portrait, Interview}`. The final `IMF_Query` is:

```
IMF_Query = {
          { "Bettini" AND "Portrait" }        OR
          { "Bettini" AND "Interview" }       OR
          { "Hasselbacher" AND "Portrait" }   OR
          { "Hasselbacher" AND "Interview" }
    }
```

(2) The `IMF_Query` is sent to the IMAIS, which returns the list of matching AV segments and corresponding annotations.

(3) The list of matching segments is analysed according to the target audience profile. This means that returned segments are ranked according to their metadata annotations and how they fit with the audience preferences. If the selected audience profile has a preference value of 0.8 for Bettini and 0.3 for Hasselbacher, then clips which are annotated as "Bettini" are sorted on the top of the returned list.

(4) The final ranked list of audio-visual segments is returned to the user interface where clips can be previewed by the editor.

4 Conclusion

We depicted how future content for interactive live TV consisting of multi-channel programs may look like. The key element for the production of this content is the exploitation of the semantic meaning of the AV clips in the production process. A new content workflow was proposed in which semantic annotations of AV objects are integrated into an intelligent media framework. To support the VC, a recommender system is described together with a first prototype of a user interface.

Acknowledgments. The LIVE project is a joint research and development effort of the following academic and commercial institutions: Fraunhofer IAIS, Cologne University of Applied Sciences, Salzburg Research Forschungsgesellschaft m.b.H., University of Ljubljana, ORF – Austrian Broadcasting, Atos Origin s.a.e., Academy of Media Arts Cologne, University of Bradford and Pixelpark AG.

This work was partially funded by the European Commission within the 6th Framework of the IST under grant number FP6-27312. All statements in this work reflect the personal ideas and opinions of the authors and not necessarily the opinions of the European Commission.

References

1. LIVE – Live Staging of Media Events (last visit 09.07. 2007), project website: http://www.ist-live.org/
2. Hartmann, A.: Producing Interactive Television, Charles River Media, Inc., Hingham, MA (2002)
3. ITV Production Standards Committee, iTV Standards Links (last visit 09.07.2007), http://itvstandards.org /iTVPublic/standards.aspx
4. Bürger, T., Gams, E., Güntner, G.: Smart Content Factory - Assisting Search for Digital Objects by Generic Linking Concepts to Multimedia Content. In: Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia (HT '05). ACM Press, New York (2005)
5. Bloehdorn, S., et al.: Semantic Annotation of Images and Videos for Multimedia Analysis. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532. Springer, Heidelberg (2005)
6. Bürger, T., Westenthaler, R.: Mind the gap - requirements for the combination of content and knowledge. In: Proceedings of the first international conference on Semantics And digital Media Technology (SAMT), Athens, Greece (December 6-8, 2006)
7. Smeulders, A.W.M., et al.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12) (December 2000)
8. Behrendt, W., Gangemi, A., Maass, W., Westenthaler, R.: Towards an Ontology-Based Distributed Architecture for Paid Content. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 257–271. Springer, Heidelberg (2005)