DIRECT-INFO
A Media Monitoring System for Sponsorship Tracking

G. Kienast, H. Rehatschek, A. Horti
JOANNEUM RESEARCH
Institute of Information Systems & Information Management
Steyrergasse 17, A-8010 Graz
+43(316)876-1182
gert.kienast@joanneum.at

S. Busemann, T. Declerck
DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
+49(681)302-5282
stephan.busemann@dfki.de

V. Hahn, R. Cavet
Fraunhofer Institut für Graphische Datenverarbeitung
Fraunhoferstraße 5
D-64283 Darmstadt
+49(6151)155-612
volker.hahn@igd.fhg.de

ABSTRACT
DIRECT-INFO is a system for media monitoring applied to the field of sponsorship tracking. Significant parts of TV streams and electronic press feeds are automatically selected and subsequently monitored to find appearances of the name or logo of a sponsoring company in connection with the sponsored party. For this purpose basic features are fully automatically extracted from TV and press and thereafter fused to semantically meaningful reports to support executive decision makers. Extracted features include logos, positive & negative mentions of a brand or product, multimodal video segmentation, speech-to-text transcripts, detected topics and genre classification. In this paper we first describe the technical workflow and architecture of the DIRECT-INFO system and then present the main innovations in four key areas.

Categories and Subject Descriptors
J.0 [**General Computer Applications**]

General Terms
Algorithms, Economics, Management, Measurement

Keywords
Media monitoring, sponsorship tracking, logo recognition, object detection, natural language processing, multimodal video segmentation, information fusion.

1. INTRODUCTION
Sport sponsors are spending significant amounts of money to be publicly associated with a team or an athlete. Knowledge about how often a sponsor is mentioned in connection with the sponsored party is a direct indicator for executive managers to estimate whether to continue sponsorship or not. The sponsored party can use this information in order to further motivate the sponsor to invest. Such information is gathered through global advertisement expenditure measurement, which is performed by media monitoring companies.

This type of business intelligence is a very complex task, which is currently performed manually in the sector of sponsorship tracking, therefore it is very expensive. These companies employ a huge staff to continuously monitor a set of media channels ranging from TV and radio to press and internet. Technology support on this work is currently mainly on the data management side after the information from the media monitored have been extracted by humans.

The focus of the DIRECT-INFO project is to create an application for semi-automatic sponsorship tracking in the area of media monitoring which shall speed up the manual process. This means a sponsor wants to know how often his brand is mentioned in connection with the sponsored company. The simple detection of a brand in one modality (e.g. video) is not sufficient to meet the requirements of this business case. Multimodal analysis and fusion – as implemented within DIRECT-INFO – is needed in order to fulfill these requirements.

1.1 Benefits for service providers & end users
Reducing labor-intensive monitoring work of TV & newspapers by providing a semi-automatic system can greatly reduce costs for media information companies and their customers. Human intervention is mainly limited to verification and validation of results. Both end users and service providers benefit from having the desired information available more quickly.

Information extracted from heterogeneous media sources (TV, newspaper, internet…) can be accessed through a common interface. Examples for possible end user requests are:

- The overall time a company logo was visible in a given broadcast
- Whether a company (their representatives, sponsors or products) appeared in positive or negative contexts in a given period in the media.

In addition to sponsorship tracking also other application scenarios are possible, e.g. the monitoring of political election campaigns. Questions arising may include: How much coverage did a party or a politician get in the news or talk shows? Was s/he

mentioned in a positive or negative way? What about competitive candidates?

2. RELATED WORK

Currently there exist only semi-automatic commercial solutions and ongoing research projects for dedicated business cases in the media monitoring domain. According to our knowledge the only multimodal analysis system which targets sponsorship tracking has been developed by the R&D project DIRECT-INFO. For the very specific (unimodal) task of brand detection the R&D project DETECT and commercial products from SPORTSi, Spikenet Technology, and Omniperception can be identified. Furthermore it has to be mentioned that there are many (manual) service providers in the market, however, very few technological solution providers.

DETECT is a completed EC project which aimed to automatically extract features from digital video streams. In particular logo appearances were detected and based on these appearances statistics for the sponsoring party were generated [5].

SPORTSi™ is a commercial software marketed by tnssport [6]. It tracks television exposure and value of brands within sport events. The software can measure exposure from advertising boards. The SPORTSi™ system is being developed with BrandTraX Inc., based on software developed by Bell Labs, the R&D department of Lucent Technologies. tnssport offers media monitoring as a service for the business cases: coverage analysis, audience data interpretation and quantity/quality of sponsor exposure.

Recently Margaux Matrix [9] began marketing a system developed by Spikenet Technology [10]. The initial target of this application is to track and to generate statistics for logos within Formula 1 race broadcasts. An initial evaluation of the demo system seems to indicate that the method is accurate, but extremely sensitive to a variety of threshold values. They claim a detection accuracy of 97.3% with a false positive rate of just 0.1. Careful parameter adjusting needs to be carried out for each logo class and broadcast to get these optimal results.

Magellan™ [11] Automatic Brand Analysis and Logo Detection software provides the tools necessary to automate the chronometry process for Sports programs. Magellan™ enables significant improvements in manual workflow throughput. The Automatic Brand Analysis and Logo Detection are based upon purpose-built proprietary image recognition algorithms. Magellan™ searches through frames of video to automatically detect the exposure of a brand image or a logo. The Magellan system by OmniPerception claims to be a robust solution to logo tracking, which offers real-time multiple brand/logo detection in video streams. It handles occlusion, overlap, rotation, and can localize the logo within the frame. It is currently not possible to tell any details of how the algorithm works or how accurate it truly is.

3. THE DIRECT-INFO SYSTEM

This section gives a short introduction on the main innovation of DIRECT-INFO (achieved both on the level of basic analysis as well as in fusion of heterogeneous low-level information to high-level semantic results) and then describes the components and their roles in the application.

3.1 Main Innovations

3.1.1 Logo Detection

An algorithm for logo recognition based on receptive fields [6], [7] has been developed by the DETECT project. This approach for logo recognition has been shown to be invariant against affine transformations, perspective transformations, occlusions and motion blur. Matching is performed by histogram intersection, which compares the histogram of computed receptive fields in a detected region of interest within a frame versus the histograms loaded from the model database.

Although the receptive fields approach is very fast the number of false positive detections is rather high depending on the logo. Furthermore the receptive fields approach is unable to recognize multiple adjacent logo occurrences as such. Adjacent occurrences are merged to one object recognition.

To overcome these drawbacks the “Scale Invariant Feature Transform” (SIFT) [1], [3] is now used and has been enhanced for logo recognition in DIRECT-INFO. SIFT is able to recognize adjacent logos as independent occurrences and application of this algorithm results in a higher precision rate. Additionally the SIFT algorithm has been extended to improve object recognition rates (see section 4 for details). The first improvement handles recognition of an object if just one interest point has matched. In this case the surrounding of the candidate object location is investigated in detail. The second improvement applies the same strategy in the temporal domain: If an object is recognized with less confidence in the next frame at the same location, the SIFT implementation examines the candidate region in more detail.

3.1.2 Multimodal Segmentation

The aim of the multimodal segmentation is the unsupervised extraction of meaningful semantic information from broadcasted video taking advantage of the media's multimodality. Since semantic is not independent of context, the goal is to detect and extract logical entities (scenes) from the video stream on top of results coming from the classification of basic event sequences.

In contrast to most solutions for video analysis, which are still focusing on one modality, the multi modal scene classification approach is based on the analysis of different kinds of information channels:

- Visual modality including artificial (graphics) and natural content (video)
- Audio modality, which includes environmental sounds as well as music, jingles etc.
- Text modality: text overlays, which provide semantic information.

The usage of multimodal analysis in video raises the question about what should be analyzed in the video stream and how could it be done. Regarding human beings the process of perception is a pre-conscious level of cognition (“signal level”); it organizes the incoming sensoric signals into information instances such as objects and events. This perceptual organization is then taken over by higher cognitive levels in order to be enriched by knowledge, so that we can become aware of what is present in the world around us. Because object recognition is still a hard task, event detection and modeling is the more promising way towards

automatic semantic annotation and description of multimodal broadcasts [14].

3.1.3 Text Analysis

In order to detect positive or negative mentions of brand names in the news, DIRECT-INFO applies linguistics-based text analysis tools. Going that far beyond keyword matching is new and highly promising, given a solid coverage of the languages in question (English and Italian). Text is available from different media:

- Speech input transformed to corresponding written text
- Newspaper transformed from PDF to plain text
- Text from on-line tickers reporting on a sports event

All sources are unrestricted and error-prone. Both these facts call for fallback strategies to more shallow analysis techniques. Clearly the performance depends on the correct recognition of the key brand names. The alternative use of statistical approaches is hardly possible, as annotated data supporting the classification process are not available and expensive to provide. In chapter 5 we concentrate on the linguistic analysis of newspaper text.

3.1.4 Information Fusion

After basic analysis is finished all results are stored in a common MPEG-7 document, which can be accessed by all components through the MPEG-7 document server.

Heterogeneous multi-media information in MPEG-7 does usually not directly answer customer needs (section 1.1). However, automatically relating pieces of information in appropriate ways can indeed answer such needs. A novel type of human-machine interface has been designed and implemented to support the media analyst in interpreting fused *Appearances* of company information and making this information available to the end user. Starting from basic MPEG-7 data, rule-based fusion operators can be used, modified and parameterized by the media analyst in order to create the complex appearances the end user is interested in. Using a comfortable use case management system combined with extensive search and retrieval functionality, the media analyst can view, interpret and publish the fused appearances for presentation to the end user. The system operates semi-automatically – the media analyst finally decides which appearances are relevant and will be published for a particular end user.

The MPEG-7 parser, the database, the fusion component and the facts management interface to the media analyst, which we also call "set-up application", are implemented in Python using the open source Zope/Plone framework [14], [22].

3.2 Pilot Use Case

For the use case targeted in the evaluation phase of the project we are interested in the *Juventus* soccer team and its sponsors *Nike*, *Tamoil* and *Sky Sports*. Appearances and mentions of the sponsors in connection with Juventus will be monitored.

The material to be monitored includes TV recordings from the Italian TV channels *Sky Sports* and *Italia 1* of December 17-19, 2004 and Italian daily newspapers *Gazzetta dello Sport* and *Corriere della Sera* of December 18-20, 2004 covering a soccer match of Juventus against Inter Milan including pre- and post-game coverage.

3.3 Workflow

In order to meet the requirements of sponsorship tracking the following workflow has been identified. This workflow can be easily configured by the user and adapted to other usage scenarios.

1. The Acquisition Component records video chunks of constant length (a few minutes) & EPG information and notifies the central content analysis controller (CAC) on their availability. Based on EPG information the Content Essence Repository (CER) and the Content Analysis Controller (CAC) prepare "Semantic Blocks", each represented by an MPEG-2 essence file and an MPEG-7 document for metadata. A semantic block is a closed TV broadcast, e.g. a soccer game or a talk show. A semantic block may be assembled from one or more recorded video chunks.
2. For each semantic block the CAC starts an analysis subsystem that performs automatic genre classification on this semantic block in order to get another indicator – next to the EPG information – if the current semantic block is to be considered "relevant" for the given use case. Based on a condensed result of the genre classification and the EPG information the CAC decides whether the semantic block is analyzed or not, e.g. a soccer game is relevant for sponsorship tracking while a soap opera is not.
3. If the semantic block is relevant for analysis, the CAC passes it to the further analysis subsystems according to the user defined workflow. For sponsorship tracking these subsystems perform logo detection, video segmentation, text analysis and topic detection based on a speech-to-text transcript. All analysis results are stored in the MPEG-7 document of the semantic block.
4. After analysis the user performs a "quality check" via the MPEG-7 Result Editor. The user may directly modify and/or approve the results or restart analysis with changed parameters or start and end time of semantic blocks.
5. After the quality check results are passed to the Fusion Component. It automatically reduces the complexity of the various low-level results by creating rule-based "Appearances" of a sponsor, team or brand. Based on user interaction appearances may be classified and edited manually. The fused results are stored in a local database.
6. When a specific customer request comes in the user queries the database and prepares the data relevant for the specific customer.
7. The customer accesses the data via a web-based interface providing access to the relevant appearances of his brand, as well as to statistics.

In parallel there is a simplified workflow for processing PDF editions of newspapers. Via a (web-) subscription interface PDF files are downloaded automatically once per day. Semantic analysis is restricted to text analysis and logo detection in images.

3.4 System Architecture

The system has been designed in a modular way to be easily reconfigurable. Between the individual components web service interfaces have been defined and implemented. Figure 1 shows the

individual components as well as the data- and control-flow between them.

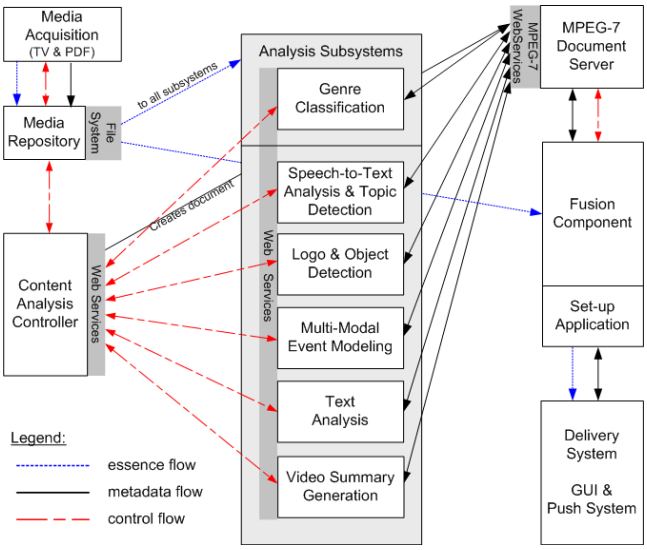

Figure 1. DIRECT-INFO system architecture.

The *Acquisition & Content Essence Repository (CER)* captures TV streams, Electronic Programming Guide (EPG) information and downloads digital newspaper subscriptions from the internet. The analysis subsystems access essence data via a shared network folder.

The *Content Analysis Controller (CAC)* is the central control logic in the DIRECT-INFO system which manages the analysis workflow. It is notified by the CER about newly available semantic blocks and distributes the analysis jobs according to a pre-defined workflow. After analysis the results can be verified via the integrated MPEG-7 Result Editor.

A set of *Analysis Subsystems* is responsible for the extraction of low- and mid-level features from the content. The subsystems expose a web service interface towards the CAC to accept control commands as to start or stop a job. One job is always performed on one entire semantic block. Status (progress, success, error) are reported back to the CAC, while the results are stored in the MPEG-7 document server. The analysis subsystems include *Genre Classification*, *Logo & Object Detection*, *Speech-to-Text Engine* including a *Topic Detection*, *Text Analysis*, *Multimodal Segmentation* and *Video Summary Generation*.

The *MPEG-7 Document Server* is the central repository for all metadata generated within the DIRECT-INFO system. It handles concurrent access to the MPEG-7 documents by the subsystems. The CAC creates an MPEG-7 document per semantic block which is then filled up with the results provided by the analysis subsystems.

The *Fusion Component* analyzes the MPEG-7 document to identify the appearances of brand names and computes complex appearances involving results of multiple subsystems. The fusion rules can be parameterized. The results are verified and published for end user delivery. These activities are carried out by the media analyst using an interface we call the *Set-up Application*. The set-up application also includes the management of user data and user requests.

The *Delivery System* is the interface for final users including an easy access to all information that is relevant for them at different levels of detail. It presents the analysis results as bulletin showing graphical illustrations of appearances including graphs, trend charts, etc. for different time periods. Furthermore it gives users the ability to browse single appearances providing related information about time and volume of appearance, positive/negative classification, keywords describing the content of the video and videos representing the appearance itself. Additionally, the delivery system provides an alert system to inform customers immediately via SMS, MMS or email in case that an important event has been detected.

The analysis workflow can easily be customized to fit other use cases by adding or replacing the analysis subsystems. The analysis workflow (which subsystems run in which order) and the analysis parameters (e.g. logos to search for) are configured via the CAC.

4. OBJECT AND LOGO DETECTION

The task of logo detection is closely related to detecting known planar objects in still and moving images, with some special requirements. Since logos vary in size , can be rotated and are subject to different lightning conditions the algorithm shall be invariant against all these factors. Logos may be partly occluded or on non-rigid surfaces (a player's shirt) so a logo shall still be detected, even if only parts of it are visible/matched. Logos shall be found in still images, nevertheless additional information gained from an image sequence may be used to improve its results. Some logos (e.g. the Nike "checkmark") are only defined by their shape and may appear in any color. Hence the algorithm must not rely only on color features. Ideally the algorithm should be configurable to be able to trade off quality vs. speed.

Based on performing a (paper-based) evaluation, using the results from [4] and some practical tests the SIFT algorithm (Scale Invariant Feature Transform) by David Lowe [1], [2], [3] was chosen to be used as it meets these requirements best.

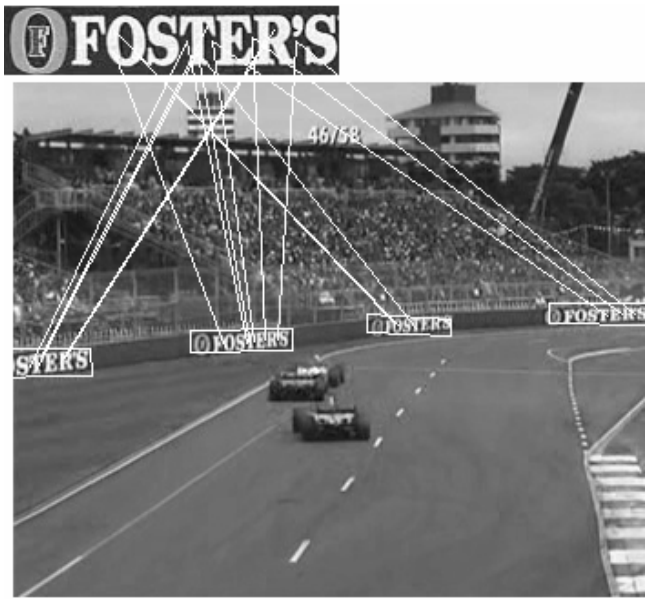

Figure 2. Logo template (top), matched key points (white lines) and detected logo appearances (white boxes).

During the DIRECT-INFO project the SIFT algorithm has been enhanced. SIFT key points are tracked to circumvent variations of logo recognition hits between frames. If an object has been recognized with high confidence in one frame, the surrounding of the estimated object position in the following frame is primarily investigated. If the object is recognized with less confidence (fewer hits) in the next frame, it can be assumed to be the same object occurrence. A two-pass approach has been introduced by splitting the object recognition in a "rough" and "fine" step. If only one single descriptor hit has been detected in the rough step, the fine step looks with more sensitive parameters in the surrounding of the hit in order to detect more matching descriptors supporting or rejecting the thesis of an object appearance.

Figure 2 shows a typical result of the logo detection performed on one video frame. As not all appearances of logos are found in every single frame, intraframe information is used to bridge those gaps to be able to track logo appearance over their entire visibility time. Problems arise due to interlaced TV content, motion blurring and the small size of logos. The size of the logo appearances in Figure 2 are on the lower limit of what currently can be detected.

However, the SIFT algorithm is not invariant against strong perspective distortions of logos. However, this can be overcome by learning synthetically distorted logo models, which are created automatically during the learning step.

Several performance improvements have been made allowing processing of five frames per second using half PAL resolution (360x288 pixels) on a standard (single CPU) PC. If necessary real-time can be achieved by only processing every n^{th} frame (depending on quality requirements) and/or using multi-processor hardware. In our application real-time is not a necessity, since the DIRECT-INFO system performs a pre-filtering step based on the relevancy of a semantic block, reducing the daily feed of 24 hours of material per channel to about six hours to be fed into analysis.

5. LINGUISTIC TEXT ANALYSIS

In this section we describe a set of subsequent processing steps from shallow to deeper linguistic levels, as well as tools to support this analysis. Thanks to pre-existing large-scale linguistic resources this new approach to opinion extraction yields promising results on free newspaper text.

As the newspaper articles are available in PDF format, some additional processes are needed to access the textual source. A tool is currently being implemented that reconstructs the logical units of the PDF document by heuristically interpreting fonts, positions and keywords. In the current system, we just extract the text from the HTML document, accepting any errors this may involve.

The textual elements are then morpho-syntactically processed, so that they can be associated with their part of speech (POS) and morphological properties (like number, gender). This is achieved by combining a huge lexical resource taken from the multi-lingual indexing software IDX with rule-based POS disambiguation tools.

We decided to use IDX – a further development of the system described in [13] – instead of a classical morphological analyzer since IDX offers the possibility to encode additional information with the tokens, such as translations in various languages and synonyms. For instance, the synonym set for the name of the soccer team "Juventus Turin" also includes "Juve" and "bianconeri", which allows us to relate these surface forms to the said team. In an application that detects opinions about an entity, this feature is for sure crucial, since it allows covering all known synonyms of the official naming of the entity.

The information delivered by IDX is refined by a POS disambiguation tool. A word like "comando" may be either a noun or a verb. Which one is correct in a given context can be told quite reliably by investigating the POS of the left and of the right neighbor.

All morpho-syntactic information is then passed to the syntactic SCHUG analyzer (Shallow and CHunk-based Unification Grammar, [12]) that provides for a dependency analysis of the textual input up to the detection of grammatical functions (e.g. subject, direct object). The following SCHUG sample analysis shows the annotation provided for the sentential level, not going into the details of the linguistic analysis of the individual noun phrases and prepositional phrases (NP, PP).

```
<LING_INFO BOS="0" EOS="17" STRING="Per ora
comanda la Juventus di Fabio Capello, che
affronta con quattro punti di vantaggio sul
Milan la volata per il titolo inverno.">
<CLAUSE id="1" BOC="0" EOC="2" MARKER="C"
 POLARITY="positive">
  <PP_ADJUNCT FRAG="0">Per ora</PP_ADJUNCT>
  <PRED">comanda</PRED>
  <SUBJ FRAG="2">la Juventus di Fabio
Capello</SUBJ>
</CLAUSE>
...
```

The sentence analyzed (under `STRING`) has been segmented into "clauses", the first one being "Per ora comanda la Juventus di Fabio Capello". A clause corresponds more or less to a semantic unit of the whole sentence, where we expect one verbal component to be present. In this first clause, SCHUG has identified three main components: the predicate (`PRED`, "comanda"), the subject (`SUBJ`, "la Juventus di Fabio Capello") and a modifying prepositional phrase (`PP_ADJUNCT`).

The `POLARITY` tag included in the annotation above is purely syntactic information: it tells us if a negative particle word ("no", "none", "never", etc.) has been used or not. In our example this is not the case. This information supports the calculus on positive/negative mention described below.

From manual annotation work on newspaper articles on soccer a list of expressions having typically a positive or negative connotation has emerged. These expressions are listed in a special lexicon, modifiable by the media analyst, which is consulted by SCHUG while processing the document:

```
comando => {POS => Noun, INT => "positive"}
dominare => {POS => Verb, INT => "positive"}
stanco => {POS => Adj, INT => "negative"}
```

For instance, all forms of the noun stem "comando" have a positive connotation (in the soccer domain).

In our example we find the verb "comandare", which is encoded in our lexicon, just like the noun "comando". We can now calculate the assessment expressed in the sentential clause on the Juventus team: it is positive, since the `POLARITY` tag is positive

for the whole clause and the subject of the clause has no negative expression in it.

The heuristics we use for detecting the positive/negative mentions are encoded as a two-level rule system based on the results of the different stages of linguistic analysis:

1. Calculate positive/negative at the level of linguistic fragments, e.g. within noun phrases (NPs), on the base of the dependency structure. For example:

```
If mention(MOD) = positive &
mention(Head_Noun) = negative
=> mention(NP) = negative
```

2. Then calculate positive/negative at the sentential levels, based on the mention properties of the grammatical relations. For example:

```
If mention(SUBJ) = positive &
mention(PRED) = negative &
no_other_arguments_present
=> mention(Sentence) = negative
```

The second heuristic would apply to a sentence like “The team of Fabio Capello lost the game in the last five minutes”.

6. MULTIMODAL SEGMENTATION

The multimodal analysis approach aims to achieve a segmentation of the broadcasted video stream into logical units, which are annotated with semantic information derived by the classification of the visual and audio context (scene modeling) and with the information derived through an OCR engine.

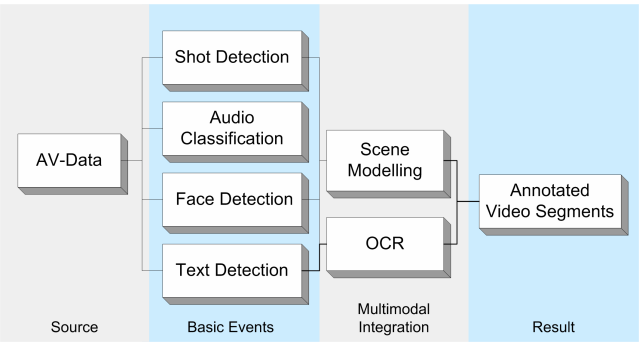

Figure 3. Workflow of the multimodal segmentation process.

For the description and identification of scenes, the detection of several basic events belonging to the different modalities is necessary:

- Transition edits in the visual modality as hard cuts or gradual transitions, which results in the segmentation of the video stream in coherent shots.
- Overlayed text events represented by uninterrupted textual expressions.
- Cuts in the auditory layout, representing changes in the sound signal. This includes transitions from silence to music, speaker changes, classification of audio segments into speech, music and noise.

Starting from basic events in video corresponding to the pure perceptual level as shots, noise, music, text-overlays, etc., the multimodal scene classification approach aims on the identification of logical units of coherent content. The segmented units are annotated using predefined semantic attributes that are derived automatically from the underlying event model describing the context of the extracted scene (e.g. attribute "named face" through combination of a detected face and a textual overlay within the same region). For news broadcasts the system intends to differentiate e.g. complete news stories, anchorman, interviews, trailers and advertisements as for sport broadcasts especially football it is envisaged to differentiate trailers, background reporting, interviews, highlights and the game itself. The multimodal event analysis additionally is accompanied with the results of the OCR engine that recognizes the results of the basic text overlay detection and passes it via the MPEG-7 document server to the text analysis module.

In the following sections the different entities of the multimodal analysis approach will be described.

6.1 Basic Event Detection

For multimodal analysis the main goal of the *Shot Detection* is the identification of visual transitions in a continuous video stream. The transitions form the syntactic structure of the visual modality separating the entire stream into single shots. Shots are generally considered as the elementary units constituting a video from an author’s or director’s point of view. Detecting shot boundaries thus means to recover those elementary video units, which in turn provide the basis, together with all other basic event detectors, for the scene and story segmentation. The shot detection approach used for the DIRECT-INFO system does not rely only on the automatic detection of hard cuts in the visual modality. It is furthermore capable to analyze and identify all kinds of visual transitions including cuts, fades, dissolves and wipes. It therefore consists of two independent parallel loops analyzing simultaneously hard cuts and special effects and consolidating the results of both for the shot segmentation.

The *Face Detection* continuously provides the number of appearing frontal faces and their position in each frame within a color video stream. This information is very valuable for annotating videos with semantic information as the genre of the video (like newscast or commercial), which can in many cases be directly determined if the positions of the appearing faces are known. For example, a video containing a single, almost unmoving face could hint for a newscast with the detected face belonging to the anchorperson. The detection is achieved by combination of a fast initial detection module (skin color filtering) with a verification approach based on SVM classification. The output of the face detection is provided to the scene modeling module for further processing and classification issues.

The *Text Detection* process consists of three different parts: The first part concentrates on the detection of edges in the video image representing the boundaries of overlaid characters. As with the edge based feature, only high contrast characters could be found. The text detection uses a second independent approach that is based on the segmentation of images into coherent regions. In a third part the results coming from both modules are combined for the final determination of text regions.

After the detection and extraction of each individual text region, the resulting sub images are preprocessed in order to create a black on white image representation for further processing through an OCR engine. The outputs of the text detection module

are the identified text regions with spatial extent, display time and recognized characters.

The *Audio Classification* approach, which was implemented for DIRECT-INFO is based on previous work from [16], [17], [18], [19]. It mainly serves two distinct purposes:

- to identify the classes of sound (speech, music, mixed classes) or the sound-emitting entities (person, printer, car) in the audiovisual content and
- to parse the audio stream into segments during which the acoustic situation keeps a certain homogeneity and at whose boundaries change events take place (e.g. a shot boundary).

The audio classification component focuses on short-time acoustic analysis, but at the same time it sets a framework for temporal analysis of sequential properties of acoustic parameters, which is necessary for the analysis of structured audio. The audio processing can be structured into three consecutive layers:

1. The pre-processing layer: It takes raw audio samples and performs a frame-segmentation into 16 ms or 256 samples using a Kaiser-windowed polyphase filter. The down-sampled frames are DC-compensated and pre-emphasized. The pre-processing results in a spectral estimation, for which a 256-bin FFT with a Hamming window is employed.
2. The feature extraction and filtering layer: This involves different feature extractors, such as Mel cepstrum, zero-crossing rate and frame energy, as well as low-level features that correspond to MPEG-7 low-level descriptors (LLDs), such as spectral flatness, centroid and periodicity.
3. The classification layer: As a final step, the actual decisions on events and/or their classes are taken. The signal change detection uses a weighted Euclidean-distance measure on neighboring frames with moving average smoothing. The signal class detection relies on a support vector classifier. In addition, the framework supports dynamic and sequence-based classification, such as to identify signals according to their segmental structure. The inference layer is designed to support supervised learning by providing an interface to feed pre-classified training data (target class labels). Supported classes are currently music, speech, noise, silence and mixtures of these.

6.2 Scene Modeling

The second stage of the multimodal analysis is merging the segments in a bottom up approach in order to reconstruct the structural layout of the content on the scene level. The scenes are series of consecutive shots connected through transitions (hard cut, fade and panning) that constitutes a logical unit of action in a video. It is defined by "bridging features" such as same visual/audio content, music/speech/noise segments or text overlays, e.g. a goal scene, an overtake event in a motor race, consecutive shots of the same location in a news broadcast.

The scene modeling component does the merging [18] of the segments by usage of the output of all event detectors and the video data itself. The merging approach analyzes consecutive segment similarities and identifies bridging features in order to combine the segments. It is based on a buoy clustering approach as described in [20]. The content of the scene is classified taking into account the information that is derived by all examined modalities using a HMM based classification approach.

Currently, a graph based modeling approach is tested for the extraction of higher level segments residing over the scene level as news stories, interviews, etc.

7. FUSION TO HIGH-LEVEL RESULTS

The knowledge processed in the fusion component originates from three sources:

- Analysis modules embodied in the MPEG-7 meta data
- Media analyst's input in the fusion process (fact assessment)
- End-user personal details, preferences and requests (query and retrieval; Q/R)

Figure 4 explains the dataflow in the fusion component within the system in more detail. From MPEG-7 content basic appearances are derived and stored. Using basic appearances and further MPEG-7 information such as EPG are used to form complex appearances by virtue of a set of fusion rules. The results are assessed for correctness by the media analyst through the facts assessment interface and stored in the Zope Object Database. The set-up application interface queries and retrieves application-specific appearances on the basis of end user requirements. The media analyst assesses which ones to make available to the end user and stores them in the database for delivery to the end user.

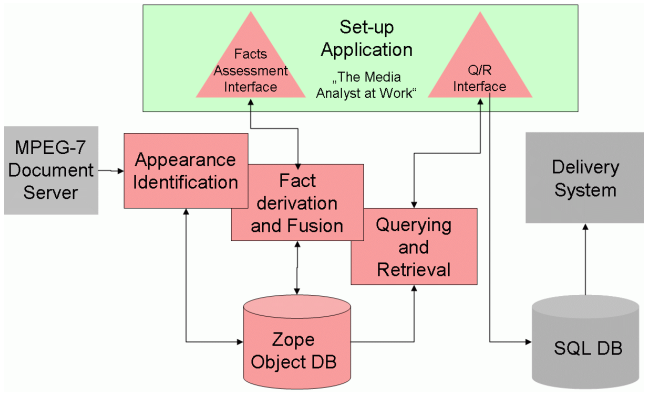

Figure 4. Data Flow in the Fusion Component.

Generating a basic appearance involves rephrasing of relevant information in terms of Plone archetypes. Currently the following types of information are stored for basic and complex appearances:

- Type: logo, speech, PDF, OCR or combinations thereof
- Date, start and end time, duration
- Sponsoring company and sponsored entity
- Text and positive/negative assessment
- Relevance: a numerical value for sorting a set of appearances when being displayed. This value is subject to change by the media analyst.

Complex appearances are formed by rules implemented in Python, the programming language of choice when using the Zope/Plone content management system. We give two sample rules to illustrate fusion results: We generate a complex appearance if

- a sponsor's logo is detected and EPG data indicate that the sponsored entity is on TV at the same time.
- a sponsor's logo and a speech appearance (positive or negative) of a sponsored entity are detected in the same time interval.

The underlying rules used are general in that they abstract away from MPEG-7 data, which they can be parameterized with. This way the same rules can be reused over different use cases.

This general approach is controlled by the media analyst. In the fact assessment interface he verifies that the basic appearances are correct and generally useful. Appearances not meeting his criteria are discarded from further consideration. The generation of complex appearances is guided by end user requirements. If the end user wishes to be informed about appearances containing a Tamoil logo and speech appearances of the soccer club Juventus at the same time, the media analyst will instantiate the second rule with the respective brand names and create end user specific complex appearances. The media analyst associates end user data with the respective use cases in the fusion component. From the Q/R interface in the set-up application, the instances will be retrieved and published with respective end user data for delivery.

8. CONCLUSIONS & OUTLOOK

The DIRECT-INFO project is entering its last six months where a focus on integration and fine tuning of components will be laid. An in-depth evaluation of the entire system will be performed by end-user partner Nielsen Media Research Italy with several key customers. Detailed tests of the analysis subsystems have been performed on the selected test material (see section 3.2) with promising results. First integrated workflow tests of the DIRECT-INFO system have been performed successfully.

Based on these tests improvements will be made: The logo detection will be extended by using color features instead of the greyscale-only approach of the SIFT Algorithm. Text analysis will greatly benefit from more extensive word annotations in the domain-specific special lexicon. For the multimodal segmentation a graph based modeling approach is tested for the extraction of higher level segments residing over the scene level such as news stories and interviews. User-friendly interfaces for data and use case management have to be implemented before the commercialization of the system by exploitation partner Idioma Ltd.

9. ACKNOWLEDGMENTS

The R&D work presented in this paper was partially funded under the 6th Framework Programme of the European Commission within the strategic objective "Semantic-based knowledge management systems" of the IST Workprogramme 2003 (IST FP6-506898).

10. REFERENCES

[1] Lowe, D., Object Recognition from Local Scale Invariant Keypoints. *Proceedings of the International Conference on Computer Vision (ICCV)*, pp 1150-1157, 1999.

[2] Lowe, D., Local Feature View Clustering for 3D Object Recognition, *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[3] Lowe, D., Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004.

[4] Mikolajczyk, K., Schmid, C., A Performance Evaluation of Local Descriptors, *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[5] Haas, W., Mayer, M., Thallinger, G., Real Time Monitoring of Radio and TV Broadcasts. *CBMI 2003 - Third International Workshop on Content-Based Multimedia Indexing*, IRISA, Rennes, France, 2003.

[6] Crowley, J.L., Riff, O., Piater, J., Fast Computation of Characteristic Scale Using a Half Octave Pyramid. *International Workshop on Cognitive Computing*, 2002.

[7] Crowley, J.L. et al., Brand Identification Using Gaussian Derivative Histograms, ICVS 2003, LNCS 2626, 2003.

[8] TNSSPORT, advertising monitoring company, http://www.tnssport.com.

[9] Margaux Matrix, media information company http://www.margaux-matrix.com.

[10] Spikenet Technology, http://www.spikenet-technology.com.

[11] Omniperception, http://www.omniperception.com/products/magellan.php.

[12] Declerck, T., A Set of Tools for Integrating Linguistic and non-Linguistic Information, *Proceedings of SAAKM 2002, ECAI 2002*, Lyon, 2002.

[13] Zimmermann, H. H., Automatische Indexierung – Entwicklung und Perspektiven, in Dahlberg, I., Schader, M. (eds.), *Automatisierung in der Klassifikation*. Frankfurt/Main, Indeks Verlag, pp 14-32, 1983.

[14] Snoek, C.G.M., Worring, M., Multimodal Video Indexing: A Review of the State-of-the-art, Multimedia Tools and Applications 25, 2005.

[15] Zhao, L. et al, Video Shot Grouping using Best-first Model Merging, Storage and Retrieval for Media Databases, 2001.

[16] Hoyt, J., and Wechsler, H., Detection of Human Speech in Structured Noise, ICASSP 94, Adelaide, Australia, 1994.

[17] Scheirer, E., and Slaney, M., Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, ICASPP 1997, vol. 2 (pp 1331-1334), 1997.

[18] Ostendorf, M., Digalakis, V., and Kimball, O., From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing, 4(5): 360-378,* 1996.

[19] Ge, X., Segmental Semi-Markov Models and Applications to Sequence Analysis. *PhD Thesis, University of California, Irvine,* 2003.

[20] Volmer, S., Buoy Indexing of Metric Feature Spaces for Fast Approximate Image Queries, *Eurographics Workshop on Multimedia,* 2001.

[21] Zope, http://www.zope.org

[22] Plone, http://www.plone.org