SPONSORSHIP TRACKING USING DISTRIBUTED MULTI-MODAL ANALYSIS (DIRECT-INFO)

G. Kienast*, H. Stiegler*, W. Bailer*, H. Rehatschek*, S. Busemann†, T. Declerck†

* JOANNEUM RESEARCH Forschungsgesellschaft mbH
Institute for Information Systems & Information Management
Steyrergasse 17, A-8010 Graz, Austria
{gert.kienast, harald.stiegler, werner.bailer, herwig.rehatschek}@joanneum.at

† DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{stephan.busemann, thierry.declerck}@dfki.de

Keywords: Media monitoring, logo recognition, natural language processing, MPEG-7, sponsorship tracking, multi-modal analysis.

Abstract

DIRECT-INFO is a system for media monitoring currently 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. Basic features are automatically extracted from TV and press and thereafter fused to semantically meaningful results to support executive decision makers. Extracted features include detected logos, positive & negative mentions of a brand or product, multimodal video segmentation, speech-to-text transcripts and teletext subtitles, detected topics and genre classification. We first describe the technical workflow and architecture of the DIRECT-INFO system and then present its main innovations in the areas of logo detection, text analysis and fusion of results.

1 Introduction

The EU advertising sector is a huge market where more than 70 bn Euro are spent every year. The measurement of this advertisement expenditure is a main task of so-called media monitoring companies, which collect daily information on how much a company spent on advertising a specific product. Having this information at hand executives and policy makers have a basis to make strategic marketing decisions. As an example, imagine Renault wants to introduce the new Laguna on the market and the general manager wants to know, how much money he has to spend on advertising in order to be successful. One important fact for such a decision is to know how much a competitor spent for introducing a similar product.

In the EC-funded R&D project DIRECT-INFO we target the concrete business case of sponsorship tracking. A sponsor wants to know how often his brand has been mentioned in connection with the sponsored company. The visual detection of a brand (e.g. logo) alone is not sufficient to meet above requirements. Multi-modal analysis and fusion of low level analysis results – as implemented within DIRECT-INFO – is needed in order to fulfil these requirements. Within this paper we briefly introduce the overall system and discuss some of the most innovative components in more detail.

2 Related Research

Logo recognition falls under the well-studied and broad area of object recognition, which has been approached from numerous directions over the past several decades. A short introduction to some of the various approaches is given in [9]. There are two primary types of features that have been followed: geometric object features and photometric object features. Geometric methods rely on 3D properties such as lines and vertices for use as salient features for matching. An in-depth survey of these types of methods can be found in [10]. Photometric methods in turn rely on the actual pixel values of the imaged object to build a model for matching. Robust color histograms or histograms built from steerable filter responses or Gabor filters all belong to this category [11, 12, 13, 14]. Much of the recent research in this area has focused on photometric features as they are capable of dealing with partial visibility (when computed locally), and are better at differentiating between large groups of similar objects [11].

There have been a number of systems designed particularly for the task of logo recognition [13, 15, 16, 17, 18] as well as a plethora of approaches toward robust object recognition in cluttered environments [12, 14, 19, 20, 21, 22, 23, 24, 25]. Whether an application driven or general approach is taken, the most important factor in a successful recognition system is the selection of salient visual features. A number of invariant features and descriptors have been developed over the years. Some of the references to visual features include: [11, 12, 14, 16, 17, 18, 24, 26, 27, 28, 29].

Spikenet Technology [6, 7] is performing research in the area of logo and object recognition. The initial target of the application was tracking and generating statistics for logos within Formula 1 race broadcasts. An initial evaluation of the

on-line demo 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%. However in tests based on the demo these numbers could not be achieved. It may also be that careful parameter tweaking needs to be carried out for each logo class and broadcast to get these optimal results.

Concerning the text-based automatic detection of positive or negative mentioning of an entity, we consider this to be a task related to a broader one called "opinion mining" (or trend mining)[1].

There are approaches that tackle opinion mining as a classification task, where documents have to be sorted for example in two (or more) classes, like "for" and "against" something or someone. For this task any known text classification algorithm could apply, on the base of appropriate feature selection. [31] reports 59% accuracy for this classification task. But [32] mentions that performance will decrease if the documents are in fact short texts, which are the kind of linguistic objects we deal with in DIRECT-INFO.

[30] presents another approach to opinion mining, which is closer to the one we pursue in DIRECT-INFO for the detection of pos./neg. mentioning: learning of linguistic patterns that match opinion on text, as described in [33]. The difference with our approach is that our system is not learning such patterns, but is applying rules that contain such patterns for the detection of the quality of mentioning of an entity.

The conclusion of [30] is that the first approach described above provides only for a raw binary classification of documents, which is not enough for the goals of DIRECT-INFO. We need to identify relevant entities in text and to report on the kind of mentioning associated with them, even in the context of very small textual documents. Therefore we have to go for more linguistic analysis, using the kind of linguistic patterns described in the second approach, for which [30] reports promising results, and specializing those to the specific tasks of DIRECT-INFO, as described in sections 4 and 6 of this paper.

3 The DIRECT-INFO System

The DIRECT-INFO system is designed as a 24/7 monitoring solution. However, the system is not a real-time system but will work in "near real-time" which is defined as follows: analysis results are available with a constant delay. Only the acquisition of material and a follow-up filtering step work in real-time, while the other components will work in near real-time. Only selected Semantic Blocks (broadcasts that contain relevant data) will be analyzed. Hence if analysis of one semantic block is finished before the next relevant block, the system never builds up a backlog of analysis jobs.

[1] A good introduction to this topic is given by Jamie Callan in [31], which we summarize here.

The technical workflow consists of media acquisition, multi-modal analysis, data fusion and result delivery. The analysis system (composed of several subsystems) is controlled by the Content Analysis Controller (CAC). All metadata produced by the subsystems are stored in a central metadata server which is fully MPEG-7 compliant.

The acquisition component is responsible for capturing digital content (video, PDF newspapers and teletext subtitles). The analysis subsystems automatically extract all relevant metadata from the digital content and store them in one MPEG-7 document per semantic block in the central metadata server. The analysis subsystems include: logo detection, speech-to-text, text analysis & topic detection, multi-modal event modelling and heuristic genre classification.

After analysis the data is fused based on rules defined per use case. The extracted data are correlated and finally presented to the operator. After a final selection phase suitable data for the final end-user (typically executive managers) is selected by the operator and passed to the delivery system. The final presentation of the results is in a standard web browser showing statistics requested by the end-user.

The DIRECT-INFO system is highly flexible: It can be easily reconfigured through the CAC in order to fulfil new use cases or to fine-tune existing ones. The workflow of the analysis system can be configured, and new analysis subsystems can be easily added. Using Web Service technology for communication between the system components makes the analysis subsystems platform independent.

4 Main Innovation of DIRECT-INFO

The SIFT algorithm [22, 23, 24], which is used in the logo recognition module has been extended to improve recognition rates and speed.

Since the original SIFT algorithm was optimized for single images, we have developed enhancements exploiting the time dependencies between subsequent frames of a video to adapt to video requirements. To overcome detection drop-outs in individual frames a tracker is used to follow logo appearances over time.

The Text Analysis (TA) component aims at interpreting whether a company or some other entity was mentioned in a positive or negative context. Not only typical words are being handled here for detecting a positive or negative mention, but also a larger linguistic context is taken into consideration in order to ensure a high-quality detection of such mentions. For this task state-of-the-art linguistic annotation tools had to be enriched with a procedure that computes the type of mention of entities on the basis of the lexical semantic properties of relevant words, taking into account the various syntactic properties of the linguistic fragments in which those words are occurring as well as the role those fragments are playing in the sentences being analysed.

Having available analysis results from various sub-systems is interesting yet insufficient to satisfy end-user needs. DIRECT-INFO can relate this information according to the

time line and to configurable semantic rules. A major innovation consists in a flexible, scalable data fusion system that is limited in principle only by the kinds of metadata encoded in MPEG-7. The Fusion Component provides a web-based user interface called "Setup Application". Through this interface the media analyst configures the fusion process according to the end users' needs and controls the publication of results through the Delivery System.

5 Logo Detection

5.1 Functionality

The system described in this section is designed to provide number, duration and size of appearances of a set of commercial logos within a video stream. The logo detection analysis component is composed of three parts: shot boundary detection, SIFT and logo tracking. These analysis units are executed in the context of a content analysis framework and have been established to achieve a high degree of flexibility. This high degree of flexibility was a necessary design criterion in order to meet user requirements. It is possible to integrate the logo detection into a slim standalone application dedicated merely to logo recognition. This application is called "BrandDetector" [36] and can be distributed on its own without requiring the entire DIRECT-INFO infrastructure.

For the purpose of logo recognition the SIFT (Scale Invariant Feature Transform) algorithm [22, 23, 24] is applied to each frame, regardless of the enabled/disabled state of the subsequent tracking module. If the subsequent tracking module is enabled, the output of the SIFT gives the tracker a hint to realign the tracked logo region. If the tracking module is disabled, the output of the SIFT algorithm gives a direct correspondence of a recognized logo to a logo occurrence.

5.2 Logo Recognition (SIFT Extensions)

The SIFT algorithm is described in detail in [22]. This paper only describes its significant enhancements, which were developed within the DIRECT-INFO project.

In [24] an extremum differs only by an arbitrary small epsilon greater zero from any neighbour. To improve results in our implementation an adaptive value has been used for this epsilon. The most dominant keypoints are extracted first, which are determined by a high epsilon. If a higher number of keypoints are required, epsilon is made smaller and more keypoints will be extracted. This way of adaptive epsilon's value change allows controlling of the number of extracted keypoints.

The algorithm has been additionally enhanced to use temporal information between frames of video content. Once the algorithm has detected a logo in a frame, it stores the logo's position. In the next frame, the algorithm will look in the surrounding of this position more precisely by extracting more keypoints in this image region (by making epsilon smaller as described above in order to get more keypoints).

A similar temporal dependency analysis is applied to the matching step. At first only those keypoints are matched, which have been extracted by usage of a high epsilon. If a match with such a dominant keypoint has been found, more keypoints in this image area will be extracted and matched in order to give more support to a logo appearance hypothesis.

5.3 Logo Tracking

Logo tracking has been used in order to compensate false negative logo detections in a sequence of true positive detections. Due to variations of image quality and lighting the SIFT algorithm may not detect a logo appearance in every single frame of the video, while it is detected in the next or previous frames. In order to outweigh these blackouts, a logo is tracked by a more reliable algorithm once it has been detected by the SIFT algorithm. Nevertheless logos may be still missed, if they are not immediately recognized when they enter the image. This means that tracking is performed only forward. Tracking backward to overcome these omissions is left for future research.

The logo tracking is based on the LK point tracking algorithm described in [30]. The tracker is initialized using the logo position and size output by the SIFT algorithm. Those LK points, which are located within the specified logo region, are considered as logo points, whereas the other points are considered as foreign points. The basic underlying LK point tracking algorithm provides additionally each point with the previous position in the frame.

In order to compute the logo's motion, a computation of all possible three logo point combinations is performed. Every three point logo combination is used to compute the affine transformation consisting of six equations (each point provides two equations: One for the horizontal and one for the vertical movement from the previous to the current position). These affine transformation parameters are inserted in a histogram. The most dominant histogram entry defines the logo motion.

The previously described approach to determine the logo motion is based on the assumptions that the logo's movement can be approximated by an affine transformation from frame to frame and that the logo consists of a uniform movement. The assumption of an affine transformation is often true, especially for logos, which are placed on static panels. But the assumption may fail if the logo is placed on an object that rotates in space depth, e.g. a Formula 1 race car moving around a curve. The assumption of uniform logo movement is true most of the time, prominent exceptions being animated logos.

The LK point tracking algorithm itself is not capable of detecting shot boundaries. In order to prevent logo tracking beyond a shot boundary, additional shot information is extracted and is then provided to the tracking algorithm.

A planned enhancement to improve the speed of the algorithm is to use SIFT keypoints for tracking, since these keypoints are extracted in previous steps of the algorithm.

6 Semantic Text Analysis

For the time being it is established that multimedia analysis alone cannot provide for the kind of semantic annotation/extraction that the DIRECT-INFO scenario does need. So some high-level semantic analysis should be applied to text material accompanying multimedia material, and the results of this analysis can then be merged with the results of the multimedia analysis procedures. Such accompanying document can consist either of transcripts from audio files, captured teletext subtitles or written documents available in various formats, like for example PDF.

6.1 Dependency Analysis

DIRECT-INFO uses state-of-the-art text analysis technology, which has already proven helpful for Semantic Web applications. A description of this state-of-the-art is given in [1], which describes the role of linguistic annotation in the context of the Semantic Web. Linguistic annotation can support both so-called Knowledge Markup (annotation of instances of ontology classes in text) or Ontology Learning and Extraction from text (see here [2] for more details).

The main linguistic structure we have been describing and using in [1] and [2] is the *dependency* structure. This kind of structure, as opposed to the sole *constituency* structure of linguistic fragments, is analyzing natural language sentence (or utterance) in a the form of dependency tree, which is encoding the role played by the various linguistic units in the textual environment. So the units playing the central syntactic role in a linguistic unit are called "head", whereas other linguistic units are said to depend on the head, and either complement or modify it. In linguistic units consisting in a nominal phrase (NP), the head of the phrase is typically the main noun, whereas the typical modifier is realized as an adjective or as a prepositional phrase.

So take as an example the following sentence: "I would definitely pay £15 million to get Owen, a decent striker, instead…". In the case of the nominal phrase (NP) "a decent striker", the linguistic annotation states that "striker" is the *head noun* whereas the word "decent" is the adjective *modifying* the head noun. Furthermore, the text analysis tool is able to propose a linguistic annotation that states that "Owen" is a decent striker, by virtue of the application of a linguistic property called "apposition"[2]. In this case, we know than that the second NP "a decent striker" is qualifying the foregoing NP, here a so-called named entity, "Owen".

6.2 Semantic Analysis

On the base of this structure proposed by the linguistic analysis tools, a semantic annotation can take place. In the case of our example, our tools can access a high-level domain specific semantic resource: a soccer ontology. In this ontology, it is stated that the concept "striker" is a sub-type of the more general concept "player". This concept can be applied to the head noun of the NP "a decent striker" and by transitivity to the NP "Owen", which is linguistically qualified by the NP "a decent striker". As the result of this process, "Owen" is being semantically annotated as a "soccer-player", "Owen" building thus an instance of the class "player" in the ontology.

Going further we can also infer from the ontology structure that "Owen" is a person, allowing thus the classification of the named-entity into the "person" type. We see here an example on how high-level semantic resources can help back in resolving tasks that are commonly considered a low-level (or shallow) analysis tasks, like the task of named-entities detection.

6.3 Computing the Polarity of Mentions

The classification of an utterance about an entity as being positive or negative goes beyond the kind of Semantic web applications we have been very briefly describing above. One reason is that to our knowledge no ontology is considering the case of positive/negative mention or interpretation. Another reason being that semantic markup is basically mapping from ontologies (or from text to ontologies in the case of *Ontology Population*). We need to define here some rules for guiding the interpretation of (semantically) annotated text in term of positive or negative mention of relevant entities.

A first step in our work was dedicated in creating specialized lexicons for various types of lexical categories (like nouns, adjectives and verbs) that can bear the property of being intrinsically positive or negative in a specific domain, as can be seen just below:

```
command => {POS => Noun, INT => "positive"}
dominate => {POS => Verb, INT => "positive"}
weak => {POS => Adj, INT => "negative"}
```

So in the case of a sentence like "ManU takes the command in the game against the weak Spanish team", the head-noun of the direct object (linguistically speaking) "the command" gets from the access to the specialized DIRECT-INFO lexicon a tag "INTERPRETATION" with value "positive". Whereas the adjective "weak" in the so-called PP adjunct "in the game against the weak Spanish team", which is modifying the verb "takes", gets an "INTERPRETATION" tag with value "negative".

Once the words in the sentence have been lexically tagged with respect to their interpretation, the computing of the pos./neg. interpretation at the level of linguistic fragments and at the level of the sentences can start on the base of heuristics we have been defining along the lines of the dependency structures delivered by the linguistic analysis. So in the case of the NP "the weak Spanish team", the head noun "team" is getting the "INTERPRETATION" tag with the value "negative", since it is modified by a "negative" adjective. In case the reference resolution algorithm of the linguistic tools

[2] The term "apposition" describes the fact that in text two nominal phrases are joined, without the intervention of a verb between them, and one NP is actually modifying (or specifying) the other one.

has been able to specify that the "Spanish team" is in fact "Real Madrid", than this entity gets a negative mention tag.

The head noun of the NP realizing the subject of the sentence, "ManU" gets a positive mention tag, since it is the subject of a positive verb and direct object combination (the NP "the command" having the tag "INTERPRETATION" with value "positive", whereas the verb "takes" is not tagged with respect to "INTERPRETATION", having a neutral reading).

Thus we have presented two examples of the computation of positive and negative readings of linguistic fragments at distinct levels of the dependency tree. In the first case this happens within a linguistic fragment, where a "negative" adjective is modifying a noun, and in the second case this happens between fragments, where a subject NP is specifying a verb-object fragment.

A last aspect to be mentioned here concern the treatment of the so-called polarity phenomenon. Specific words in natural language carry intrinsically a negation or position force. So the words *not*, *none* or *no* have an intrinsic negation force and negate the words and fragments in the context in which those specific words are occurring. The context that is negated by such words can be also called the "scopus" of the negation (or the range). Consider for example a reformulation of our first example: I would definitely pay £15 million to get Owen, *not even* a decent striker, instead…" Our tools are able to detect that the NP "decent striker" is negated, and therefore the positive reading is being ruled out. At the actual level of development, the tools still do not decided if the negation of expressions that are potentially positive leads automatically to a negative interpretation.

To summarize, the linguistic tools used in DIRECT-INFO have been enriched with lexical information and heuristics for computing the neg./pos. mention of entities, going thus beyond state-of-the-art technology in language technology coupled with semantic web technologies.

7 Metadata Description

The different content analysis modules extract a number of different types of metadata, ranging from low-level audiovisual feature descriptions to semantic metadata. The metadata description must be rich, as it is the input of the fusion component and must thus include all the information gathered during content analysis.

7.1 Using MPEG-7 for Detailed Description of Audiovisual Content

In DIRECT-INFO the MPEG-7 standard is used for metadata description. It is an excellent choice for describing audiovisual content, mainly because of its comprehensiveness and flexibility. The comprehensiveness results from the fact that the standard has been designed for a broad range of applications and thus employs very general and widely applicable concepts. The standard contains a large set of tools for diverse types of annotations on different semantic levels. The flexibility of MPEG-7 which is provided by a high level of generality makes it usable for a broad application area without imposing strict constraints on the metadata models of these applications. The flexibility is very much based on the structuring tools and allows the description to be modular and on different levels of abstraction. MPEG-7 supports fine grained description, and it is possible to attach descriptors to arbitrary segments on any level of detail of the description.

The comprehensiveness and flexibility of MPEG-7 lead to a large number of description tools and increases the complexity of descriptions. Profiles and levels have been proposed to define subsets of the standard tailored towards certain functionalities and with different levels of complexity [3]. Three profiles have been adopted for standardization [5].

A more general issue is that of interoperability between systems and applications using MPEG-7. The limited functionality of a profile provides an opportunity for better interoperability, as the openness and generality of the full standard can be constrained according to the application area.

The profiles that have been defined in part 9 of the standard [5] are not sufficient for such rich and diverse metadata descriptions as created in DIRECT-INFO. For example, none of the currently defined profiles includes the tools for visual and audio feature description (part 3 and 4 of the MPEG-7 standard). The adopted profiles have been designed with complexity reduction and not with interoperability in mind and thus no semantic constraints have been defined for these profiles. The semantic constraints can be compared to the explanations on description tool semantics in the text of the standard. In a profile, a reduced set of supported functionalities is defined, and the semantic constraints can be as restrictive as this set of functionalities allows. Thus the semantic constraints in a profile definition can be used to resolve ambiguities caused by the openness of the tool semantics in the standard and thus support interoperability.

In DIRECT-INFO, the metadata description needs to support annotations on different semantic levels and several annotations, with some of them defining their own decomposition of the content. Examples for such diverse annotations with different and independent decompositions are shot structure, genre classification and automatic speech recognition results. This requires a profile that includes all the necessary tools and defines sufficient constraints on a semantic level to ensure unambiguous use of the included tools.

7.2 MPEG-7 Detailed Audiovisual Profile (DAVP)

DIRECT-INFO has thus contributed to the development of the Detailed Audiovisual Profile (DAVP) [4]. The profile contains tools for the description of the spatial, temporal and spatiotemporal structure of the types of content listed above, the description of media information, the description of creation and production information, the description of semantic information, the description of visual and audio features and the summarization of image, audio, video and audiovisual content.

The intended use of this profile is the detailed description of image, audio, video and audiovisual content entities. By "detailed" we mean the support of content structuring down to a fine level of granularity and the capability to describe annotations and features on different abstraction levels. There is a broad range of applications where such a detailed description of audiovisual content is required. This encompasses all kinds of applications that deal with the analysis, description, retrieval, summarization and exchange of audiovisual content. The profile is designed to support the use of automatic and manual annotation tools and content-based query paradigms such as query by example. Possible application areas include audiovisual archives, image and video databases, media monitoring applications, audiovisual content production and educational applications. The XML Schema of the profile and a comprehensive description of the semantic constraints are available on the web [4].

A basic principle of DAVP is to describe only a single audiovisual content entity per MPEG-7 description. The description includes the detailed description of the content, consisting of structural description and associated textual or feature descriptions on arbitrary levels, and a summary of the content description. The profile allows the use of a wide range of spatiotemporal structuring tools. A further design principle is to keep content decompositions and the related annotations as modular as possible, i.e. to separate decompositions based on different modalities, on different levels of abstraction, or created with/without the use of domain knowledge.

In contrast to the three adopted MPEG-7 profiles, part 3 and 4 (visual and audio feature descriptions) have been completely included in DAVP. We are convinced that a comprehensive description of audiovisual content must allow the use of low- and mid-level feature descriptions. For example, audiovisual content analysis techniques are in many cases not capable of producing high level information which is directly semantically meaningful to a user. Semantic information extraction approaches are used to infer high-level information from a set of media descriptions and often additional external sources.

8 Fusion of Basic Analysis Results

Data fusion using different resources is a challenging task. As only high-quality results are acceptable to end-users, DIRECT-INFO opted for an automatic fusion process complemented with a manual assessment and correction phase. Hence quality assurance remains with the human media analyst.

The level of granularity is the *appearance*, representing an occurrence of a logo, an inlay or a mention of interest. Fusing appearances requires a homogeneous representation scheme, which is defined using archetypes in Section 7.4.

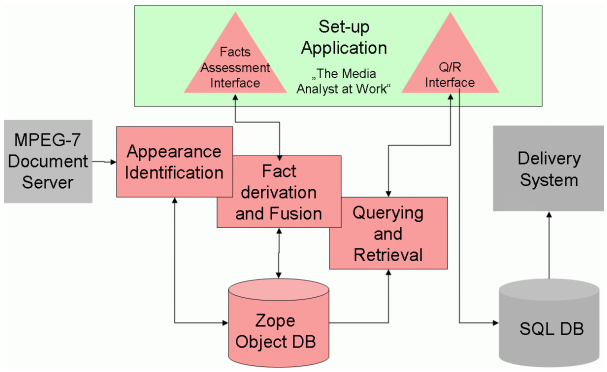

Figure 1. Data Flow in the Fusion Component.

8.1 The Fusion Technology

The technology used in the Fusion component is based on the Zope [34] application server, the Plone [35] content management system (CMS), and the Archetypes package that allows the easy definition of new content types for the Plone CMS, including the automatic generation of pages to edit the content, and abstractions over the way the content is stored and presented. This software is in the public domain.

The Fusion Component is triggered by the Content Analysis Controller after the latter has collected all data for the MPEG-7 document of one Semantic Block. It passes the ID of the MPEG-7 document to the Fusion Component via an HTTP call to a URL that is being administered by the Zope server. Using this ID the Fusion Component requests the actual MPEG-7 document from the MPEG-7 document server via its web service interface.

8.2 Data Flow in the Fusion Component

The knowledge processed in the Fusion Component originates from three sources:

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

Figure 1 explains the dataflow in the Fusion Component within the system. From MPEG-7 content basic appearances are derived and stored. Basic appearances and further MPEG-7 information such as EPG are then used to form complex appearances through a set of fusion rules. These rules are parameterized with respect to sponsor name, company name, or date and time. Results are assessed for correctness by the media analyst through the Facts Assessment Interface and stored in the Zope Object Database. The Setup Application Q/R interface queries and retrieves application-specific appearances according to end-user requirements. The media analyst decides which ones to make available to the end-user and stores them in the database for delivery to the end-user.

8.3 Examples

Complex appearances are formed by rules implemented in Python, the programming language of choice when using the Zope/Plone content management system. We generate a complex appearance, if e.g.

- 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.

8.4 Archetypes

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

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

Moreover, appearances are equipped with a key frame showing recognized information and a link to the relevant part of the broadcast, which can be played in-line upon request. This way the media analyst is supported in her decisions regarding correctness and relevance.

8.5 The Setup Application

The Setup Application is a Python routine that is invoked by the media analyst to drive the fusion process, the identification of fused material, and its shipping to the Delivery System. It implements the data flow described in Section 8.2 and offers at the same time a convenient user interface for the media analyst. The Setup Application consists of the following interdependent modules:

- **User Management**: Maintains end-user profiles.
- **Fusion Use Case Management**: Specifies parameters, selects rule templates, and assigns end-users. Starts the fusion process for a fusion use case definition.
- **Facts Assessment Interface**: Defines the state of basic and complex appearances: This can be *visible*: (initial state of the appearance when created by the fusion process) *approved* (verified by the media analyst), *rejected* (when the media analyst decides the appearance should not be approved), and *published* (the final state for results the end-user will be able to view in the Delivery System)
- **Query and Search interface**: combines date, company, brand and other criteria in order to retrieve appearances of interest.

9 Results & Outlook

First tests of the logo recognition subsystem have shown promising results. On planar logos without major perspective distortions recall values of above 90% can be achieved, with a precision of about 75%. Test material containing logos with significant distortions (perspective or due to non-rigid surfaces) yield lower recall values.

A first evaluation of the text analysis module has been provided on the base of a small scale corpus of Italian PDF newspaper articles dedicated to soccer. This corpus has been annotated with respect to pos./neg. mentions of Juventus manually by an Italian native speaker on the base of her subjective interpretation. This study was biased by the fact that the linguistic parser (the same parser as the one described in [2]) used for generating the pos./neg. mentions had just been adapted for covering Italian language as well. Performances of this parser for Italian was still poor so that a relevant number of linguistic patterns that we consider relevant for the extraction of the quality of mentioning were missing in the output of the parser, a fact which explains that the recall was very low (below 20%), much lower as we can expect from the linguistic-based approach to the detection of qualifying mentioning. But interesting enough, the precision achieved was very high (above 80%).

Detailed tests and evaluations of the DIRECT-INFO system are being performed in Q4/2005 by the end user partner Nielsen Media Research Italy by tracking sponsorship for the Juventus soccer club as a first use case. For the end-users a high recall rate will be of special importance. False detections can be easily deleted in a manual post-processing step while missed detections have to be considered lost for the customer.

Adding additional languages to the text analysis system, currently Italian and English only, is foreseen. Future work will explore the possibility to define fusion rules with through a user interface, thus avoiding the media analyst having to program rules in Python.

Acknowledgements

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 Work Programme 2003 (IST FP6-506898).

More information about the DIRECT-INFO project can be found on the website http://www.direct-info.net.

References

[1] P. Buitelaar, T. Declerck, "Linguistic Annotation for the Semantic web", in: Siegfried Handschuh, Steffen Staab

(eds.) Annotation for the Semantic Web, IOS Press, (2003).

[2] T. Declerck, M. Vela, "Linguistic Dependencies as a Basis for the Extraction of Semantic Relations", in Proceedings of the *ECCB'05 Workshop on Biomedical Ontologies and Text Processing*, Madrid (2005)

[3] Definition of MPEG-7 Description Profiling. ISO/IEC JTC 1/SC 29/ WG 11 N6079, Oct. 2003.

[4] MPEG-7 Detailed Audiovisual Profile (DAVP). http://mpeg-7.joanneum.at.

[5] Study of MPEG-7 Profiles Part 9 Committee Draft. ISO/IEC JTC1/SC29/WG11 N6263, Dec. 2003.

[6] Spikenet Technology, Homepage 2004. http://www.spikenet-technology.com

[7] M. Cutler. "Bringing Science to Research." http://www.spikenet-technology.com/ download/BusinessF1%20March%202004.pdf

[8] Envisional, Homepage 2004. http://www.envisional.com/technology_logo.html

[9] C. Schmid, R. Mohr. "Local Greyvalue Invariants for Image Retrieval." IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 1997. Vol. 19, no. 5, pp. 530-535.

[10] P. Besl, R. Jain, "Three-dimensional Object Recognition." ACM Computing Surveys. 1985. Vol. 17, no. 1, pp. 75-145.

[11] C. Schmid, R. Mohr. "Local Greyvalue Invariants for Image Retrieval." IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 1997. Vol. 19, no. 5, pp. 530-535.

[12] D. Hall, V.C. de Verdère, James Crowley. "Object Recognition Using Coloured Receptive Fields." In Proceedings of the *European Conference on Computer Vision (ECCV),* Dublin, Ireland. 2000.

[13] F. Pelisson, D. Hall, O. Riff, J. Crowley. "Brand Identification Using Gaussian Derivative Histograms." *International Conference on Vision Systems*, Graz, Austria. 2003.

[14] D. Slater, G. Healey. "The Illumination-Invariant Recognition of 3D Objects Using Local Color Invariants." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. 1996. Vol. 18, no. 2, pp. 206-210.

[15] G.G. Medioni, G. Guy, H. Rom and A.R.J. François. "Real-Time Billboard Substitution in a Video Stream." *Proceedings of the 10th Tyrrhenian International Workshop on Digital Communications*, "Multimedia Communications", Ischia, Italy, 1998, pp. 71-84.

[16] F. Aldershoff, T. Gevers. "Visual Tracking and Localization of Billboards in Streamed Soccer Matches." *SPIE Electronic Imaging 2004*, San Jose, CA., USA, 2004. Vol. 5307, pp. 408-416.

[17] R. den Hollander, A. Hanjalic. "Logo Recognition in Video Still By String Matching." In *Proceedings of International Conference on Image Processing.* 2003. Vol. III, pp. 517-520.

[18] B. Kovar, A. Hanjalic. "Logo Appearance Detection and Classification in a Sport Video." 2002.

[19] F. Schaffalitzky, A. Zisserman. "Automated Scene Matching in Movies." In Proceedings of *Conference on Image and Video Retrieval*. 2002, pp. 186-197.

[20] J. Sivic, A. Zisserman. "Video Google: A Text Retrieval Approach to Object Matching in Videos." Proceedings of the *International Conference on Computer Vision (ICCV).* 2003, pp. 1-8.

[21] V. Ferrari, T. Tuytelaars, L.V Gool. "Simultaneous Object Recognition and Segmentation by Image Exploration." In *Proceedings of European Conference on Computer Vision* (ECCV). 2004.

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

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

[24] D. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision.* 2004 (to appear).

[25] B. Funt, G. Finlayson. "Color Constant Color Indexing." IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995. Vol. 17, no. 5, pp. 522-529.

[26] M. Bertini, C. Columbo, A. Del Bimbo. "Automatic Caption Localization Using Salient Points." *International Conference on Multimedia and Expo (ICME)*. 2001, pp. 69-72.

[27] K. Mikolajczyk and C. Schmid. "An Affine Invariant Interest Point Detector." In *Proceedings of the European Conference on Computer Vision (ECCV)*. 2002.

[28] J. Matas, O. Chum, M. Urban, T. Pajdla. "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions." In *Proceedings of the British Machine Vision Conference, London. 2002.* pp. 384-393.

[29] C. Harris, M. Stephens. "A Combined Corner and Edge Detector". In *Fourth Alvey Vision Conference*. 1988, pp. 147-151.

[30] J.-Y. Bouguet. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm, Intel Corporation, Microprocessor Research Labs, *OpenCV documents*.

[31] P.D. Turney. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." In *Proceedings 40th Annual Meeting of the Association for Computational Linguistics (ACL'02),* 2002, pp. 417-424.

[32] J. Callan. A lecture on Opinion Mining, download from http://hartford.lti.cs.cmu.edu/classes/95-779/Lectures/07-OpinionsB.pdf.

[33] J. Wiebe, R.F. Bruce, T.P. O'Hara. "Development and use of a gold-standard data set for subjectivity classifications." In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99).* 1999.

[34] Zope application server. www.zope.org

[35] Plone content management system. www.plone.org

[36] BrandDetector, Product sheet, download from http://www.joanneum.at/cms_img/img2289.pdf