DIRECT-INFO: A DISTRIBUTED MULTIMODAL ANALYSIS SYSTEM FOR MEDIA MONITORING APPLICATIONS*

H. REHATSCHEK, N. DIAKOPOULOS AND G. KIENAST

JOANNEUM RESEARCH Forschungsgesellschaft mbH
Insititute for Information Systems & Information Management
Steyrergasse 17, 8010 Graz, Austria
E-mail: herwig.rehatschek@joanneum.at

V. HAHN

Fraunhofer, Institut für Graphische Datenverarbeitung
Fraunhoferstraße 5, D-64283 Darmstadt, Germany
E-mail: volker.hahn@igd.fhg.de

T. DECLERK

Deutsches Forschungszentrum für Künstliche Intelligenz
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
E-mail: declerck@dfki.de

DIRECT-INFO aims to create a basic system for semi-automatic sponsorship tracking in the area of media monitoring. Its main goal is to offer an integrated system combining the output of basic media analysis modules to semantically meaningful trend analysis results, which shall give executive managers and policy makers a solid basis for their strategic decisions. In this paper we put a special emphasis on the subsystems logo recognition, the multimodal scene classification and the text analysis subsystem.

1. Introduction

Main goal of media monitoring companies is to calculate global advertisement expenditure on all kind of products (e.g. cars, toothpaste, cameras, etc.) and deliver to their customers numbers on specific products. Customers of media monitoring companies are executives, policy and decision makers, who have different motivations why to receive data on how much money one company spent on a specific advertisement campaign for one product. As an example a car producer who wants to introduce a new product on the market wants to know, how much money he has to spend in order to

* R&D work 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).

make the market introduction with a high probability successful. A first important basis for this decision is the amount of money his competitors spent on campaigns in order to introduce successfully a compare-able competing car into the market. That is where media monitoring companies are contacted in order to prepare these numbers.

Another important business case in this connection, which is also the main target of the DIRECT-INFO system, is sponsorship tracking. The main goal of this business case is to find out how often the brand of the sponsor is mentioned in connection with the sponsored company. So the simple detection of a brand in one modality (e.g. video) is not sufficient in order to meet the requirements of this business case. A multi-modal analysis and fusion, which relates information from different modalities will be needed in order to fulfill these requirements. Such a multi-modal analysis is implemented within DIRECT-INFO.

Some examples in order to show the complexity of the task: Juventus has Nike as a sponsor. If the brand Nike was detected in a Juventus soccer game it is a relevant detection and shall be recorded. If the brand Nike was detected in an AC Milan soccer game it is a non-relevant detection because Nike is also sponsor of AC Milan. If the brand Nike was detected in an advertisement during a Juventus soccer game it is still a non-relevant detection because this is a different business case (advertisement monitoring).

Currently media monitoring is still a mostly manually performed task, meaning for each modality (TV, radio, newspaper, cinema, etc.) a number of people are hired which continuously monitor this modality and annotate appearances of advertisements according to the targeted business case. There are some automatic solutions offered by technology providers. However, they primarily support only specific pieces within the area of media monitoring and decision support (e.g. speech-to-text, text-based knowledge extraction) and mostly only one modality, such as only audio or only video. According to our knowledge no technology provider offers a comprehensive solution that addresses a combined solution for multimedia monitoring over different information channels.

Within the DIRECT-INFO project we target a solution capable of monitoring and relating different modalities for the specific business case sponsorship tracking. This paper first introduces the system architecture and the general workflow. Then we concentrate on three concrete analysis subsystems: static object / logo recognition, multi-modal scene classification and text analysis.

2. Related Work

In this section we introduce related work in connection with the three subsystems, which are described within this paper in more detail.

2.1. *Static Object / Logo Recognition*

The task of logo recognition is essentially that of detecting and recognizing known planar objects in both still and moving imagery. Geometric and photometric affine invariance are thus obvious components to the recognition algorithm since logos can appear at different locations, scales, and rotations as well as under different lighting conditions in a 3D imaged scene. Logos may also be mirrored or placed on non-rigid deforming surfaces such as clothing, which adds additional non-affine transformations that also need to be recognizable. For the particular application being considered, the algorithm must also not rely exclusively on color information as logos may vary in color, and should be tunable for speed vs. quality of recognition. The overall speed should allow for timely processing of video blocks, though real-time is the ultimate goal.

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 by Schmid in [13]. 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 [24]. 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 [13, 11, 12, 15]. 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 [13]. The simplest type of photometric matching scheme is template matching, however, this approach is ineffective for all but the most carefully engineered environments.

There have been a number of systems designed particularly for the task of logo recognition [2, 8, 10, 12, 9], as well as a plethora of approaches toward robust object recognition in cluttered environments [11, 5, 3, 15, 17, 18, 20, 19, 22]. 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, not all of which can be included here due to time constraints. Some of the references to visual features include: [6, 8, 9, 10, 11,

13, 14, 15, 16, 19, 23]. The rest of this review will be devoted to discussing the references for the specific approaches, general approaches, and visual features in general, as well as some conclusions on what a good solution for our needs is.

2.2. *Multi-Modal Scene Classification*

The aim of the multimodal scene classification is the unsupervised extraction of consistent and meaningful semantic information based on event modeling of 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 broadcasted 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 four different kinds of information channels:

- Visual modality: this includes everything that is visible in the video scene, including artificial (graphics) and natural content (video)
- Audio modality: this includes environmental sounds as well as music, jingles etc.
- Speech modality: this includes the spoken language in the video, which could already provide semantic information about the content of the video
- Text modality: this includes text overlays, which also provide semantic information already.

The usage of multimodal analysis in video raises the question about what should be analysed 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 (for instance, visual light waves or auditory sound waves) 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 modelling is the more promising way towards automatic semantic annotation and description of multimodal broadcasts [30].

2.3. *Text Analysis for pos. / neg. Mentioning*

Parts of the multimedia material analyzed for the sponsorship tracking consist in textual documents that are either a transcripts from audio files or texts delivered within PDF files. There is a need thus to apply here text analysis techniques for supporting the detection of positive or negative mentioning of a sponsor, related to teams or other entities related to the sponsors.

DIRECT-INFO uses state-of-the-art text analysis technology, which was developed in other projects, and adapted to the objectives of the project. A description of state-of-the-art is given in [31], which describes the role of linguistic annotation in the context of the Semantic Web. The main concept here is the one of the so-called *dependency analysis.* This kind of analysis structures a natural language sentence (or utterance) in a 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 the main noun, whereas the typical modifier is realized as an adjective or as a prepositional phrase, as can be seen in Figure 1, showing an example of a complex NP.

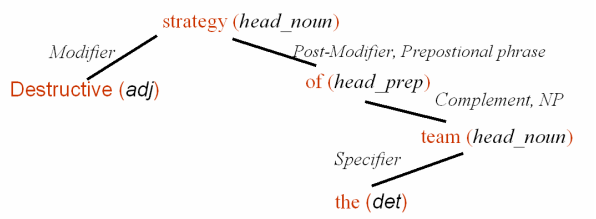

Figure 1: A complex NP

The noun strategy is the head of the whole NP. It is modified by the adjective “destructive” and by the PP “of the team”. In this PP the syntactic head is the preposition “of”, having an obligatory completion, realized as the NP “the team”. In this NP again the head is a substantive, the noun “team”.

A lexical semantic analysis can then be applied on the top of the dependency structure, which allows to precise the relations existing between the different linguistic units present in the dependency tree resulting from the syntactic analysis. So for example the semantic interpretation of the preposition “of”, as a kind of possessive, allows to semantically relating the head noun “team” to the head noun “strategy” in a specific “has_a” relation. The head noun “team” also inherits the adjectival modification “destructive”, with its negative connotation.

For sure not only adjectives are carrying information about “positive/negative” features of a utterance, also nouns and verbs do carry such information, and in all cases the consideration of the whole dependency structure is a necessary step in providing for the right interpretation of the nouns or verbs. How we want to proceed in Direct-Info is explained in more details in the section "text analysis subsystem". Important to note here, is that the output of the text analysis component can be combined with the results of

visual or audio analysis of related multimedia material in order to support better multimedia analysis results, as shown in Figure 2.

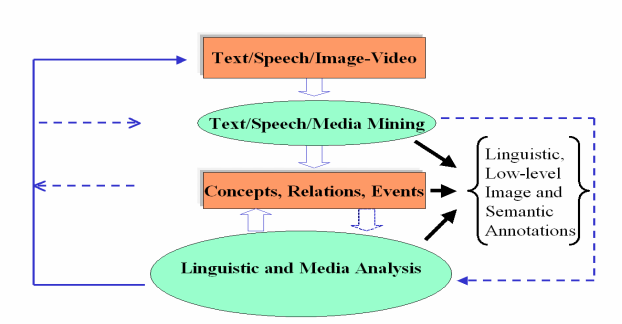

Figure 2: Combination of multiple media for producing semantically relevant annotations about positive/negative mentioning

3. Overview on DIRECT-INFO System

The DIRECT-INFO system architecture is based on the following basic design criteria: to take a proven, feasible and affordable and cost effective approach, to build on reliable technology, to define an open architecture in terms of extensibility, modularity and exchangeability of components and to be scalable of computing power and data throughput. Other important requirements that are more technical are: independency of the main system components, ease of integration and system robustness.

The DIRECT-INFO system shall be able to do 24/7 monitoring. Since DIRECT-INFO will not work in real-time we need to pre-filter the incoming stream in order to detect so called relevant "semantic blocks", which refer to programs that are relevant for analysis. The Content Analysis Controller decides on the relevancy of these blocks by taking into account the EPG information and the genre classification, which is given in Figure 3.

Obviously this approach works only on the assumption, that in the average all the subsystems can perform the analysis steps in between the time the next semantic block is detected. This assumption is a realistic and was agreed with the users and aligned to the use cases.

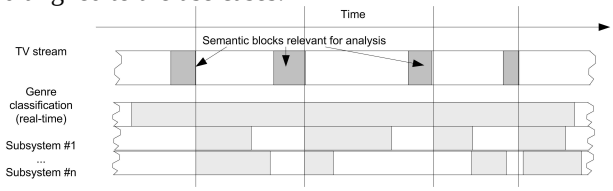

Figure 3: Analysis timeline

3.1. *Technical Workflow*

In order to meet the requirements of sponsorship tracking the workflow given below has been identified which is visualized in Figure 4. As already mentioned above, this workflow can be easily configured by the user and adapted to other scenarios.

1. Acquisition records video chunks of constant length & EPG information and notifies the central content analysis controller (CAC) on their availability.
2. Based on EPG information the CAC prepares semantic blocks (represented as MPEG-7 documents) i.e. per sport event, TV show etc. A semantic block can cover one or more video chunks.
3. The CAC starts an automatic genre classification subsystem on this semantic block in order to get another indicator – next to the EPG information - if the current semantic block is relevant for analysis.
4. Based on a condensed result of the genre classification and the EPG information the CAC decides if the corresponding semantic block shall be analyzed or not.
5. If a semantic block is relevant for analysis, the CAC passes the block now according the user defined workflow to the corresponding analysis subsystems as given in Figure 4.
6. After analysis a “Quality Check” is performed by the user indicated by the MPEG-7 XML result editor/viewer in Figure 4. In case the user realizes that begin and end time codes are wrong or wants to change any parameters for the analysis, he defines new values and the process for this semantic block restarts at step 5.
7. After the quality check the results are passed to the fusion component.
8. The fusion component first automatically reduces the different results of the analysis subsystems according to user defined rules. Then based on user interaction the data will be classified and manually edited. The fused classified semantic blocks are stored in a local database of the fusion component.
9. If a specific customer request comes in the user can now via a set-up application query the database of fused classified semantic blocks and put together for the specific customer relevant data.
10. The GUI / Push system visualizes the output of this set-up application via a web interface and/or immediately alerts (via SMS, MMS or email) the end user in case of important events.

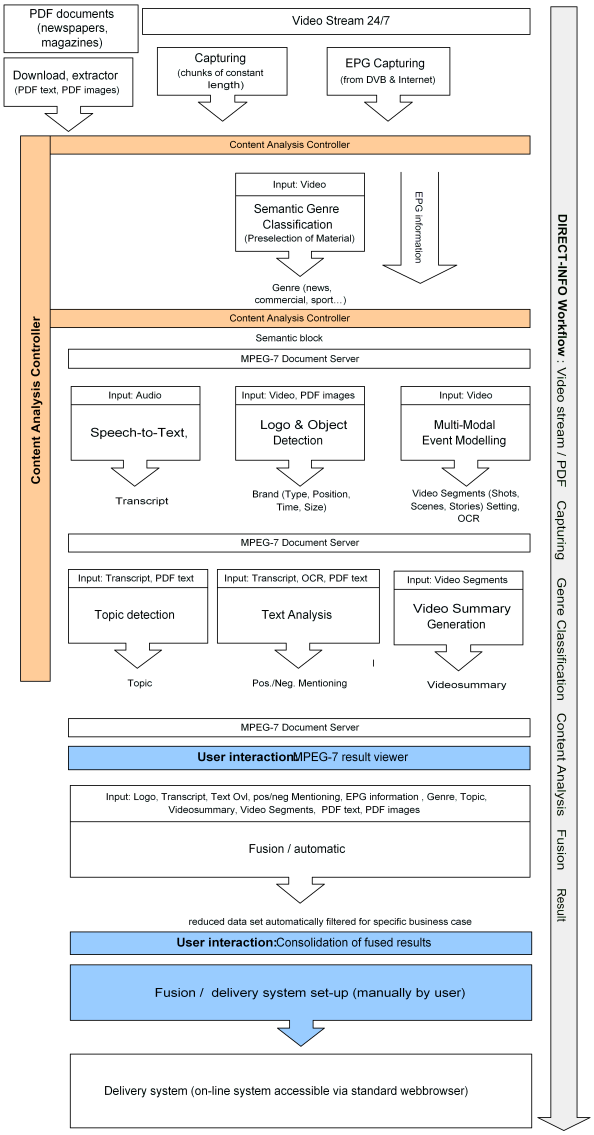

Figure 4: workflow for business case sponsorship tracking

Even though we target within DIRECT-INFO the specific business case "sponsorship tracking", the system architecture is kept flexible in terms that the system can also be reconfigured to other business cases on demand by e.g. defining a different workflow or by integrating different analysis subsystems.

3.2. *System Architecture*

For realization of this workflow the system architecture as visualized in Figure 5 was defined. In order to easily enable further integration of new analysis subsystems all communication between components is based on well defined interfaces via web services. Metadata are stored centrally in the international standardized MPEG-7 format within the MPEG-7 document server. Essence data is stored centrally in the media repository and transferred on demand to the analysis subsystems.

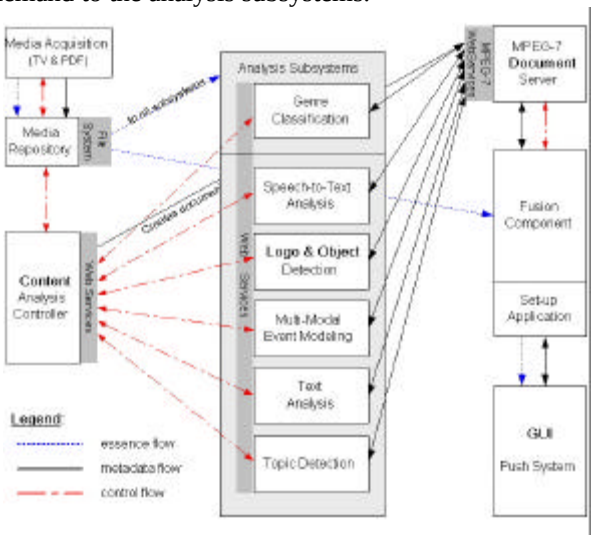

Figure 5: System architecture of DIRECT-INFO system

As shown in Figure 5 there are three different kinds of flows within the DIRECT-INFO system. The essence flow describes where the content (meaning: text, audio, video, images) has to be distributed within the system. The metadata flow describes the distribution of all descriptive data belonging to the content, e.g. acquisition time of MPEG-2 essence, or the results generated by the logo detection module. The control flow denotes which components of the system control (start, stop…) other components.

4. Static Object / Logo Recognition

4.1. *Requirements on the Algorithm*

From a technical point of view the main features to be extracted include recognize logo within each frame, also several appearances of one and the same, the size (length, width in pixels plus ratio in relation to the frame size,

the position (left top, right bottom co-ordinates, orientation) and the start timecode, end timecode where logo appeared. In more detail the algorithm shall meet the following requirements:

Speed versus quality scaling. The algorithm shall be adjustable according to quality (speed low) and speed (quality high). Ideally this can be triggered by a parameter allowing for any value between quality and speed.

Colour invariance. Logos appear in different colours hence an algorithm which relies only on colour information is not sufficient.

Lighting invariance. Logos may appear under different lighting conditions (e.g. in the sun, in the shadow). The algorithm must be robust against different lighting conditions.

Invariance against translation, scaling, rotation and affine projections. Brand can vary in size, it can be rotated (including depth rotations) etc.. The algorithm shall be invariant against all these factors.

Detection in still and moving images. Even though the main content to be analyzed shall be video the algorithm shall be able to work with still images as well. Hence it shall not necessarily rely on the additional information which may be gained out from a sequence of images.

4.2. *Selection of an Algorithm*

The literature survey carried out represents only a small fraction of the research on object recognition which has been done in the last 30 years. It can be seen that there are a multitude of approaches each offering its own set of advantages and disadvantages. The requirements of the desired system are clearly high and the state of art offers no ready made solution to the given set of tasks. However, there is some very sound progress that has been made in the field of object recognition which should allow for the construction of an algorithm fulfilling most of the requirements. In our opinion, the work of Lowe [18, 20, 19] provides the most promising direction for the development of an algorithm for logo detection in connection with our requirements.

Speed versus quality scaling is an algorithm characteristic that is not available in all implementation strategies. There is always the option to scale an image processing algorithm based on the resolution of the input image, however, for many object recognition algorithms there is not a simple architecture allowing for scalability. Some algorithms are based on color characteristics whereas others are not. Ideally a method which is fundamentally based on luminance channel features but which can be supplemented with color information should be used. Much of the most recent research in the field focuses on being invariant to scale and rotation, however the question of mirror invariance is one that hasn’t been carefully considered in the research. As a mirror transform is non-affine it is not so easily integrated into the existing methods. However, a workable approach would be to create two models for

each logo – one normal and one mirrored. All of the algorithms detailed above can be applied equally well to either video imagery or still images extracted from, for example, newspapers.

Based on the above requirements, an implementation strategy along the lines of [19] will be developed. This is due to the fact that this approach promised the best trade-off between speed and accuracy (quality) and also meets the set requirements very well by achieving the highest score. As time permits, additional features can be integrated and efforts undertaken to accelerate the algorithm as much as possible.

5. Multi-Modal Scene Classification

DIRECT-INFO's multimodal scene classification approach aims to segment the broadcasted video stream into logical story units first, and then to annotate them with semantic information derived by the classification of the visual context and with the information that is gathered through an OCR engine. For the description and identification of logical story units (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 e.g. transitions from silence to music, speaker changes, categorisation of audio segments into speech, music, 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 "master events" representing logical units of coherent content. Master events could occur on different stages or levels of the cognitive analysis process as they could be used for the description of compact entities as single news stories, trailers, interview situations, etc. as well as for complete broadcasts, which itself contain several master events.

In contrast to the multimodal scene classification recent approaches for video segmentation and storyboarding are limited mostly to the detection of shot boundaries and clustering of keyframes [29]. The results of these shot detection algorithms are mainly a very syntactic segmentation of the video stream without any logical and semantic structure.

To achieve the goal of multi-modal integration in terms of event detection, classification, or identification, an approach based on Hidden Markov Models is used. 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. The result of the multimodal event analysis is a detailed description of the content structure, which will be used as input for the video summary generation as well as a description of its visual context. For news broadcasts the system differentiates e.g. complete news stories, anchorman, interviews, trailers and advertisements as for sport broadcasts especially football we can differentiate trailers, background reporting, interviews, highlights and the game itself. The multimodal event analysis additionally will be 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.

6. Text Analysis Subsystem

The text analysis subsystem, called SCHUG (Shallow and Chunk-based Unification grammars) used in DIRECT-INFO provide for a cascaded linguistic and semantic analysis of (free) text (see [32] for more details, those tools have been further developed in the Esperonto project). It combines a robust syntactic analysis approach with a deeper annotation strategy, relying on high-level semantic information, which can be given in the form of complex domain specific ontologies and/or also lexical semantic networks.

SCHUG processes a sentence in a bottom-up manner, detecting first small linguistic fragments (chunks), and on the base of this first round (or cascade) of annotation detects (and annotates) larger chunks, and so on till the maximal possible level of annotation is reached. Chunks are also annotated with dependency information.

If the system cannot recognize a full sentence (due to the possibly incomplete input or to the capabilities of the system), SCHUG returns the detected chunks so far, ensuring thus the robustness of the analysis. While detecting sentences, SCHUG also provide for a clausal analysis, since sentences can be complex and contain more than just one main verb. The analysis of the clauses of a sentence is an important step in order to provide for an accurate semantic interpretation of the whole sentence.

The system maps then the annotated dependency structures onto available semantic networks and delivers a list of annotations, covering both linguistic and (domain specific) information. Especially for the Direct-Info project, the annotation structure of SCHUG has been augmented with the “polarity” tag and is being currently augmented with “positive/negative” mentioning tags.

The polarity analysis is dedicated to a further linguistic consideration playing a role in the detection of positive and negative mentioning. Some words, normally linguistically categorized as “particles”, bear an intrinsic negative or positive interpretation. So for example: “Unemployment didn’t increase during the last three months”, the verb “increase” has been negated by

the particle “not” (or its contracted form in combination with the verb “do”), and linguists say that the negation particle “not” has scope over the verbal phrase “increase during the last three month.“ The SCHUG tools add in this case a feature “POLARITY” to the linguistically annotated text, and this feature can have two values: “positive” or “negative”.

But one should be aware, that negative/positive mentioning cannot be directly deduced on the base of only the polarity annotation of fragments of the sentence. Context should here also be taken in consideration. So in our example: “Unemployment didn’t increase during the last three month”, we cannot deduce that there is a negative mentioning of the “Unemployment”. The polarity information associated with linguistic units has to be combined with background textual information in order to accurately contribute to the detection of relevant positive or negative mentioning of entities.

7. Conclusions and Outlook

In connection with the static object / logo detection we will finalize first a MatLab implementation of the most promising Lowe algorithm in order to test its accuracy against our requirements. Especially we want to compare it to our already existing algorithm based on the work of [25]. Since the Lowe algorithm is originally designed for still images, a first test will be an application to video content. The original Lowe algorithm does not take color information into consideration. However, for some logos color information may be relevant. So a possible future direction will be the extension of the algorithm by means of color. A first prototype implementation of the algorithm is scheduled for December 2004.

A first prototype of the entire DIRECT-INFO system working on a restricted test data set is expected to be ready in December 2004. Version 1.0 will be available by September 2005. Further up-to-date information on the project can be obtained from the public project website: http://www.direct-info.net.

8. References

1. Tappen M., Freeman W., Adelson E., 2002, Recovering Intrinsic Images from a Single Image. In Proceedings of Conference on Neural Information Processing Systems (NIPS)
2. Medioni G., Guy G., Rom H., François A., 1998, Real-Time Billboard Substitution in a Video Stream, Proceedings of the 10th Tyrrhenian International Workshop on Digital Communications, Multimedia Communications, Ischia, Italy, pp. 71-84.

3. Sivic J. and Zisserman A., 2003, Video Google: A Text Retrieval Approach to Object Matching in Videos., Proceedings of the International Conference on Computer Vision (ICCV), pp. 1-8.
4. Paletta L., Lux A., Crowley J., 2002, ROI Detection Specification", Deliverable DETECT-D4.2.1-ROI Detection Specification.
5. Schaffalitzky F., Zisserman A., 2002, Automated Scene Matching in Movies, In Proceedings of Conference on Image and Video Retrieval, pp. 186-197
6. Bertini M., Columbo C., Del Bimbo A., 2001, Automatic Caption Localization Using Salient Points, International Conference on Multimedia and Expo (ICME), pp. 69-72.
7. Pan H., Li B., Sezan I., 2002. Automatic Detection of Replay Segments in Broadcast Sports Programs by Detection of Logos in Scene Transitions, International Conference on Acoustics Speech and Signal Processing (ICASSP)
8. Aldershoff F., Gevers T., 2004, Visual Tracking and Localization of Billboards in Streamed Soccer Matches, SPIE Electronic Imaging 2004, San Jose, CA., USA, Vol. 5307, pp. 408-416
9. Kovar B. , Hanjalic A, 2002,. Logo Appearance Detection and Classification in a Sport Video
10. Hollander R., Hanjalic A., 2003, Logo Recognition in Video Still By String Matching, In Proceedings of International Conference on Image Processing, Vol. III, pp. 517-520
11. Hall D., Colin de Verdère V., Crowley J., 2000, Object Recognition Using Coloured Receptive Fields, In Proceedings of the European Conference on Computer Vision (ECCV), Dublin, Ireland.
12. Pelisson F., Hall D., Riff O., Crowley J., 2003, Brand Identification Using Gaussian Derivative Histograms." International Conference on Vision Systems, Graz, Austria.
13. Schmid C., Mohr R., 1997,. Local Greyvalue Invariants for Image Retrieval." IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 19, no. 5, pp. 530-535.
14. Mikolajczyk K., Schmid C., 2002, An Affine Invariant Interest Point Detector, In Proceedings of the European Conference on Computer Vision (ECCV).
15. Slater D., Healey G., 1996, 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.
16. Matas J., Chum O., Urban M., Pajdla T., 2002,. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions, In Proceedings of the British Machine Vision Conference, London, pp. 384-393.

17. Ferrari V., Tuytelaars T., Gool L.v.,2004, Simultaneous Object Recognition and Segmentation by Image Exploration, In Proceedings of European Conference on Computer Vision (ECCV).
18. Lowe D., 1999, Object Recognition from Local Scale-Invariant Features." In Proceedings of the International Conference on Computer Vision (ICCV), pp. 1150-1157.
19. Lowe D., 2004, Distinctive Image Features from Scale-Invariant Keypoints." International Journal of Computer Vision. 2004, (to appear).
20. Lowe D., 2001, Local Feature View Clustering for 3D Object Recognition, In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR).
21. Beis J., Lowe D., 1997, Shape Indexing Using Approximate Nearest-Neighbour Search in High-Dimensional Spaces, In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)
22. Funt B., Finlayson G., 1995,.Color Constant Color Indexing." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, no. 5, pp. 522-529.
23. Harris Ch., Stephens M., 1988, A Combined Corner and Edge Detector,. In Fourth Alvey Vision Conference. pp. 147-151.
24. Besl P., Jain R., 1985, Three-dimensional Object Recognition, ACM Computing Surveys, Vol. 17, no. 1, pp. 75-145.
25. Crowley J., Lux A., Paletta L., Hall D., Riff O., 2003, Motion Picture Analysis Tool Specification. Deliverable: DETECT-D4.3.1, Motion Picture Analysis Tool Spec.
26. Thorpe S., 2002, Ultra-rapid Scene Categorization with a Wave of Spikes." In H.H. Bulthoff et al (eds), Biologically Motivated Computer Vision, Lecture Notes in Computer Science, 2525, pp1-15, 2002.
27. Swain M., Ballard D., 1991, Color Indexing, International Journal of Computer Vision, Vol. 7, no. 1, pp.11-32.
28. Sebe N., Lew M., 2001, Comparing Salient Point Detectors, IEEE International Conference on Multimedia and Expo, pp. 65-68.
29. Zhao L., et al., 2001, Video Shot Grouping using Best-first Model Merging, Storage and Retrieval for Media Databases, 4315, pp. 262-269.
30. Snoek C.G.M., Worring M., 2004, Multimodal Video Indexing: A Review of the State-of-the-art, Multimedia Tools and Applications, (to appear)
31. Buitelaar P., Declerck T., 2003, Linguistic Annotation for the Semantic web, In: Siegfried Handschuh, Steffen Staab (eds.) Annotation for the Semantic Web, IOS Press.
32. Declerck T., 2002, A set of tools for integrating linguistic and non-linguistic information, in Proceedings of SAAKM (ECAI Workshop).