16 MULTIMEDIA INFORMATION RETRIEVAL FRAMEWORK: FROM THEORY TO PRACTICE

F. Moelaert EL-Hadidy, H.J.G. de Poot, and
D.D. Velthausz

Telematica Instituut
P.O. Box 589, 7500 AN, Enschede, the Netherlands,
Phone: +31-53-4850485, Fax: +31-53-4850400,
E-mail: {moelaert,poot,velthausz}@telin.nl

Abstract: Easy access to multimedia information is important as the amount of information is exponentially growing. Content-based retrieval is a viable approach. Extensibility and flexibility of data models is another important issue as the field is constantly evolving. The ADMIRE framework applied in this paper satisfies both criteria. This we show in two practical cases, one in the field of clinical assessments of measurement data, the other in the field of video directory services. ADMIRE offers a uniform solution for structuring these data. Content disclosure is supported by labelling of these data structures. This is done by automatic, semi-automatic or manual labelling.

16.1 INTRODUCTION

The aim of multimedia IR systems is to handle general queries such as "find outdoor pictures or videos of an interview with James Cameron discussing the making of the Titanic film". Answering such queries requires intelligent exploitation of both speech and visual content. For multimedia retrieval, the combination of multiple integrated media types increases the performance of content-based retrieval, to overcome mismatches and missing matches. Available content analysis and retrieval techniques tailored to a specific media are therefore not adequate for queries as the one mentioned above. From the above example it is clear that multimedia IR is a very broad area covering both infrastructure issues (e.g. framework models, efficient storage criteria, networking, client-server models) and intelligent content analysis

and retrieval. This all needs to be integrated as a seamless whole. That involves expertise from a wide variety of fields and lots of research.

The infrastructure issues are closely related to information supply. The way information is structured has a direct impact on the retrieval strategies. So for example if a slide show presentation and a corresponding audio track are stored as two loose information objects, it is difficult to retrieve the sound related to a certain slide in the presentation. On the other hand if the system allows the definition of relationships between different media then retrieval becomes more easy. Multimedia models have thus to allow for a uniform way to represent different media and the relationships between media.

Here the need emerges for a heterogeneous framework for representing all types of media in a uniform manner. This paper contributes to the infrastructure issues with a general framework for representing multimodal information.

Existing multimedia models focus mostly on a single aspect of multimedia information, like presentation (e.g. PREMO [ISO94]), or exchange of documents, or on a particular format (e.g. HyTime [ISO97]). Models that do facilitate content-based information retrieval in general are for example MORE [TYH+91], VODAK [GuNe93], CORE [WNM+95] and AIR [GuRV96]. These models either do not support a layered definition of information objects (e.g. MORE and VODAK) or can only represent the content of specific unstructured media types (e.g. AIR). MPEG7 which is currently being developed as a general multimedia standard for content representation for information disclosure [MPEG98] is promising, but it is still in an early stage. The model in [CMF96] for representing digitised document and the Amsterdam Hypermedia Model (AHM) [Hard98] is directed to modelling and authoring of Hypermedia documents.

This paper presents a general framework called ADMIRE [VeBE96]. It resembles the CORE model but offers more flexibility in modelling object relationships. ADMIRE emphasises on the disclosure of all kinds of forms and types of existing digital information [Velt98]. It uses an object-oriented modelling technique together with a layered definition of information objects. It is thus suitable for representing multimedia information. The basic difference between these models and ADMIRE lies in the ability to represent features and concepts at every level in the information hierarchy. Features are context independent functions operating on the data (e.g. shape). Concepts are context dependent interpretations of the data (e.g. goal).

The intelligent content analysis and retrieval of multimedia IR are related to information demand. Unlike traditional databases, an exact matching between demand and supply in multimedia databases is frequently not possible since the information is differently structured. This requires, besides an accurate and comprehensive model of the available information, specific functionality to handle the queries and the presentation of the (intermediate) results.

In content-based retrieval, information should be extracted from features, concepts or a combination of both. This should be possible across all media. To retrieve a statement that James Cameron made in the interview, a combination of pattern recognition (feature) in video material for lip movement and the concept “Oscar” in the audio material can be combined. Such queries are strongly dependent on the context where they are used. Therefore, the choice of features and concepts and extraction rules depend on the context. This paper shows how the ADMIRE framework supports extraction of multimedia information in general. Further more, two cases are presented to demonstrate context specific matters one in the medical field and the other in the field of video directory services (VDS).

In the clinical information case patients’ movement disorders are recorded and monitored using multiple techniques. This results in huge files with multimedia data. To support data analysis data disclosure techniques are used. The VDS case covering the soccer domain offers an interactive way to retrieve (fragments of) content that the user is interested in. To support queries in that domain we need to disclose soccer related information.

In section 16.2 the ADMIRE framework is briefly explained. A general guideline for content-based retrieval using the ADMIRE framework is presented in section 16.3. This has been applied in two cases that are presented in sections 16.4 and 16.5. Finally, some conclusions are given in section 16.6.

16.2 ADMIRE FRAMEWORK

As the field of multimedia is constantly developing, we advocate a generic approach to multimedia modelling. Our ADMIRE framework, presented in [VeBE96], provides an information object hierarchy on the one hand and an abstraction hierarchy on the other hand. The information object hierarchy makes sure that every piece of multimedia information, from loose pictures or soundbites to elaborated reports, coverages etc. can be described. The abstraction hierarchy distinguishes respectively raw data, computable features of this data and context and interpretation dependent concepts. See Figure 16.1. All properties in the data layer (format and attribute properties) are stored since they convey information that cannot be determined differently. The feature and concept layers contain information that can be determined using the properties of the lower layers, and extraction algorithms for features and (domain) inference rules for concepts. As opposed to features, concepts are context dependent: Different concepts can be inferred from the same data and features, Figure 16.1.

We go through some simple examples of multimedia information to demonstrate how the ADMIRE framework works. Figure 16.2 is an image containing three objects. This image can be modelled in ADMIRE as an information object (IO) as shown in Figure 16.1. At the data level this IO

consists of an array of pixels or, depending on the format, a set of vectors and areas. At the feature level functions can be defined that operate on this data, e.g., the percentage green, the presence of sharp edges, and the such. At the concept level interpretations and meaning is given to the data, e.g., the fact that there is a tree, a sun and a partially visible car.

Figure 16.1: Illustration of an IO **Figure 16.2:** an example of an IO.

We distinguish between retrievable and non-retrievable IOs. The smallest retrievable IO we call a *basic* IO e.g. a frame in a video and is defined by the designer. A collection of basic IOs we call composite IO. In this paper when we mention IO we mean retrievable IO. A non-retrievable IOs where further decomposition of an IO is unfeasible, e.g. individual words or pixels, but still useful for extraction, we call *pseudo IOs*. Pseudo IOs are part of the information hierarchy. They are also characterised by features and concepts, but the data possesses a function that specifies the subset of the raw data and format of an accompanying IO. It is comparable to the non-materialised nature of ‘scripted objects’ [ScWy95] and ‘anchor values’ used in hypermedia models [HaSc94]. Examples of pseudo IOs are: sequence of characters within a textual fragment; a specific region in a pixel based image (frame), e.g. rectangle 1 indicating a ‘car’ as illustrated in frame 1 in Figure 16.3 or a person’s voice within an audio fragment.

In pseudo IOs, features and concepts are strongly linked. E.g. in pseudo video IO B, colours and shapes are features from which the concept “tree” can be inferred. Also the sound feature of an audio IO, if present, of a high frequency white noise could reflect the wind blowing through the leaves, adding to the concept “tree”. Therefore, any feature from any modality can contribute to the concepts that are being inferred.

Figure 16.3: Four sample frames of a video shot with the corresponding nine pseudo frame IOs and three pseudo video shot IOs

Figure 16.4: The information hierarchy of our sample movie shot

In the ADMIRE framework, this integration of media is modelled in hierarchical relations which exist for IOs as well as their pseudo IOs. The example of Figure 16.2 is frame *j* in the video shot IO in Figure 16.3. In the centre nine corresponding pseudo *frame* IOs are shown. Related pseudo *frame* IOs in successive frames are linked forming a pseudo video shot IO. See A,B and C in Figure 16.3. The information hierarchy for this example for a movie shot is given in Figure 16.4. It includes images (frames) from the video shot and sound tracks from audio. The data layer of the pseudo video shot object allows new features to be extracted, e.g., concerning a more completed shape of an object as some occlusions in individual frames are resolved or concerning rotation (viz. of the wheels). These features may then support concepts like “wheels”, “fast”, “car” etc.

Motion between objects, is part of the feature layer of the video shot IO. At the concept layer, these motions can be interpreted, e.g. using world knowledge that the camera rather than a tree moves, and that cars do not fly. From this the motion of the car may be inferred. Each feature and concept has to take its suitable place in the information hierarchy.

In Figure 16.5 the ADMIRE information structure for the example is given. The frames are *basic IOs*. The shot is a *composite IO* as it contains sub IOs

of type frame. The contents of the (pseudo)IOs can be read as follows: At the pseudo frame IO level colour and shape are extractable features and objects are extractable concepts. At the frame level new features and concepts arise. Positions of and relations between pseudo-objects are features and states are concepts. At the pseudo video shot level similar features can be combined to new features and concepts, either to extract static aspects (shape, paint) or to extract dynamic aspects (wheels turning). At the video shot level motion is an extractable feature and action verbs and changes are extractable concepts.

A formal approach to do concept inference is presented in section 16.3, while practical examples of concept inference presented in section 16.4 and section 16.5.

Figure 16.5: The hierarchy between (pseudo) IOs (arrows) and within IOs (layers)

16.3 PROPERTY DERIVATION

As we have seen, information in the ADMIRE framework can be modelled at multiple levels of granularity. Not just data, but also features and concepts of basic IOs can be aggregated forming composite IO. Figure 16.6 is an extended version of Figure 16.1. It shows that a concept associated to a composite IO can be inferred in multiple ways: entirely on the basis of features and concepts of the same composite IO, i.e., within the composite IO, or entirely on the basis of concepts of its sub-IOs, or anywhere between. This depends on the quality of features and concepts.

Two alternatives are illustrated in Figure 16.7. On the left all necessary features are aggregated to a feature within the composite IO which we call

feature aggregation. On the right concepts of sub IOs are aggregated to infer a new concept. This we call concept aggregation.

Figure 16.6: Property aggregation operations

Figure 16.7: Composite IOs concept derivation using feature aggregation (left) and concept aggregation.(right)

As the two examples show, concept inference initially requires features. Concepts can also be labelled manually. But that depends on the interpretation of the labeller. When concepts are automatically inferred, the features must be explicit. Features are mathematical functions. E.g. the feature greenness is a function of an object's RGB value. How high need a greenness be, to call something "green"? This is the process of concept inference: In what range must function values lie to infer a concept? And how certain are we about this concept? Therefore, many researchers have associated *belief values* to concept labels and so do we. The concept inference is twofold: How are belief values estimated when features are evaluated and concepts are inferred and How are belief values combined? (multiple knowledge rules applying to one concept, multiple concepts being aggregated, concept as premises in a knowledge rule). The estimation of belief values is not addressed here. We apply the definition of *precision* as in [Rijs79]. High precision as a search key justifies high belief.

The combination of belief values has been extensively addressed in literature. Some well known methods to do automatic inference of computer supported evidence were considered, namely rule based systems ([BuSh84]), Bayesian Belief networks ([LaSp88], [Pear90]), and Dempster-Shafer sets

([Shaf78]). Below we characterise these methods and the method we propose by their pros and contras (for an extensive review, see [KrCl93]):

method	pros	contras
rule based	simple, allowing belief & disbelief	causality & its inverse mixed up [KrCl93, p63], order effect & no semantics [Heck86]
BBN	causality made explicit, anything is computable	closed world assumption, lot of input needed [WiEd88]
DS sets	notion of ignorance, semantic soundness, no order effect, idempotence	strange renormalisations due to closed world assumption [Zade84], [Clar88]
our approach	the above pros	none of the above contras

As a basic building block we adopt knowledge rules of the kind $P[c/e]=p$, i.e., The belief (elicited probability) that the concept c applies, given evidence e is p , where $0 \leq p \leq 1$. When in an IO the criterion e is met, the triple $\langle c, e, p \rangle$ is added to the “bunch” of evidence, i.e., the set of instantiated knowledge rules, e.g. $\{\langle c, e, p \rangle\}$ when $\langle c, e, p \rangle$ is the first piece of evidence. Working with sets of evidence commutativity (unimportance of order of pieces of evidence) and idempotence (unimportance of double instances of pieces of evidence) are assured.

Suppose that concept c has a priori belief c_0 , and that there would be k independent pieces of evidence, $B = \{ \langle c, e_1, p_1 \rangle, \dots, \langle c, e_k, p_k \rangle \}$, then the belief in concept c given bunch of evidence B would be $P^+[c/B] = 1 - (1 - c_0)^{1-k} \prod (1 - p_i)$. The a priori correction factor is necessary because it is implicitly included in each of the p_i 's. The $+$ in P^+ denotes that only positive evidence is combined here. This is comparable to DS theory, but with an open world assumption and without re-normalising belief values.

Similarly counter evidence may be combined where $P(-c) \equiv P^+(-c)$, where we propose a conservative combination $P(c) = P^+(c) \times (1 - P(-c))$.

If pieces of evidence for a concept are not independent, the joint conditional belief values must be known e.g., $\langle c, e_m, \cdot \rangle, \dots, \langle c, e_n, \cdot \rangle$ being reduced to $\langle c, \{e_m, \dots, e_n\}, P[c/e_m, \dots, e_n] \rangle$.

When concepts are evidence to other concepts, their uncertainty is propagated into the knowledge rule: E.g., concept X , with belief value $P[X]=p$ is used to infer concept Y with conditional belief value $P[Y|X]=q$, this is equivalent to a piece of evidence $\langle Y, X, r \rangle$ with belief value $r = P[Y|X] \times P[X] = p \times q$, i.e., the belief in both X and Y given X .

With Bayesian belief networks our approach has in common that knowledge rules are chosen with care: We use feature and concept information only in a bottom-up fashion, and attribute information only in a top-down fashion. For example a concept of an IO can be inferred on the basis of concepts of its sub-IOs, but not on the basis of concepts of its containing super IO. On the other hand, an attribute of an IO is inherited by its sub-IOs. Because of this strictness, a sequence of inferred concepts cannot be dependent on itself across different IOs. Circular or *mutual* dependencies between concepts within a single IO are easily avoided by checking the knowledge rules operating on that IO.

We emphasise that our method is pragmatic allowing easy integration in the case studies. Experience with this approach will show its usability.

16.4 CASE 1: CLINICAL ASSESSMENT

At the Roessingh Research & Development rehabilitation clinic (The Netherlands) large amounts of data are collected of patients suffering from movement disorders (e.g., cerebral palsy). Data include muscle activity (EMG), motion, force (viz. of the feet on the ground), and video recordings, as well as ordinary patient data. In Figure 16.8 the processes leading to body movement is drawn: First there must be an intention for body movement, then the central nervous system executes a motor program, activating the muscles, then the muscles start to contract changing the angles between joints. This leads to body movement as we observe it. The EMG, force and motion recordings give information about the last three phases of this process. With the advent of multimedia it is now possible to view these data on a computer and share data with colleagues, cf. Figure 16.9. Furthermore automated searches are possible, e.g. finding patients with similar movement disorders.

Figure 16.8: Different measurements at a functional motion chain

The ADMIRE framework is used to structure all patient information. The information hierarchy is given in Figure 16.10. It facilitates content-based retrieval of folders of patients with similar disorders: From the data recordings, features can be defined like the phase lag between EMG and kinetic extremes. These features enable limited patient similarity searching. For sophisticated searches concepts are needed, as we will illustrate using a ‘stiff legged knee’ case.

Figure 16.9: The Sybar system [Haut97] is used for quantitative monitoring and teleconsultation between Roessingh and AZVU in the MESH project [HoLu98].

Figure 16.10: Information objects hierarchy structure.

Property extraction

Matching features is not just matching graphs of measurement data. The suitability of a method depends on its domain. For example, for a ‘stiff-

legged knee', the knee motion range during the swing phase is an important criterion. Looking at the sagittal plane, the marked dots of Figure 16.11 indicate where abnormal knee function is most significant [Perr92]. So the *knee motion range feature* (max - min, during swing phase) can be used as a degree of 'stiffness'. Such choices should be made by a domain expert. These features can be used to find similar 'stiff legged' patients, but also to extract the concept 'stiff legged knee', e.g. using the rule that the absolute motion range is less than 20 degrees. Features from different modalities have been used to classify patient's gait, e.g. using EMG [BCP+79], and EMG in combination with photographed marker trajectories in [KnRi79] or kinematic data [WiGH87]. The ADMIRE framework applied to the Roessingh case enables a multitude of feature combinations that can be used for feature similarity searching and even more important concept extraction.

Figure 16.11: Knee motion, the marked dots indicate where most abnormal knee function can be observed.

Property aggregation

As explained in section 16.3, two types of aggregation can be applied to infer concepts. Feature aggregation results in the inference of a single concept while concept aggregation combines already inferred concepts. Now let us look at the subobjects of the gait IO. During a single gait measurement multiple cycles are measured. A cycle within a gait, i.e. the period between two successive initial heel floor contacts of the same foot, can be automatically identified. So gait cycles form important pseudo IOs. On the other hand the sub IOs of the gait IO, as Figure 16.10, could lead to concepts. As timing relationships between different measurements, like joint angles and EMG signals, are important for gait analyses, the gait pseudo IO are most important.

As explained before the concept 'stiff legged knee' can be inferred for individual cycles, hence pseudo gait IOs. Aggregating these concepts will result in concepts for the gait IO (cf. Figure 16.7). In addition, features from

multiple successive cycles (pseudo gait IO) would typically result in an aggregated feature from which a concept for the gait IO is inferred (cf. Figure 16.7). Thus the identification and extraction of properties for pseudo IOs is important for the inference of concepts of IOs. Furthermore this enables a better and finer-grained similarity search.

Following the query by example paradigm, the measured patient's gait is used as the similarity search key. Searching for similar patients can be done through by comparing properties of any of the IOs in the object hierarchy. For example using concepts to eliminate the non relevant gait IOs while using features of e.g. the pseudo gait IOs to determine a more finer quantitative similarity. The above indicates that using concepts are more appropriate to search in the information hierarchy in a top-down fashion while features are suitable at lower levels in the information hierarchy.

There is an ongoing research effort between the Telematica Instituut and Roessingh Research Center to apply the methods described above. From the Mesh project [HoLu98], which is a general framework for computer supported co-operative work, a telemedicine pilot is currently running. Here a need has been identified for a multimodal patient database. Also a query language and a feature database are necessary.

Collaborative aspects for multimedia databases are also being addressed. The advantage of using the ADMIRE model in the above research lies mainly in two aspects, the possibility to query concepts as well as features, and being able to combine the evidence of different media to obtain higher accuracy in queries.

16.5 CASE 2: VIDEO DIRECTORY SERVICE

Here we present a video directory service (VDS) for soccer matches. Different from the medical case in section 16.4, the users of a VDS are typically ordinary consumers that do not want to be bothered with the rather technical feature properties. So the challenge is to label these concepts (semi)automatically.

Quite a few studies were centred around analysis of soccer games. There are generally two possible, an elaborated off-line video and live video. The approach presented here is mainly designed for live broadcast, like e.g. [AnHR94]. In live broadcast there is no look ahead.

Although soccer seems easily understandable, with players wearing a number, teams wearing specific colours, and the ball being clearly visible, the images from video material are quite difficult to interpret with a computer algorithm. This is due to limited TV image resolution and the camera focus at one depth at a time.

There are many studies dealing with translation of different media into language, e.g. languages for the description of images [Okad79], video [Badl75], etc. These languages can be used for disclosure of the media. An

example used in the soccer game can be found in [AnHR94]. There a system is built that automatically generates textual comments to soccer videos in real-time. Here one stationary camera was used for recording.

In [InBo95] the source material is a normal recording as opposed to the fixed camera in the previous study. Here a football game is used for disclosure using video data only. The complexity here lies in tracking objects. Domain knowledge is used to solve uncertainties.

The paper [GSC+95] focuses on labelling of video shots. It is assumed that relevant objects are present in the first two frames of the video shot. Here only a subset of soccer related features is supported such as detection of edges to identify the soccer court, colour detection of a team's uniform etc. This allows the identification of concepts as corner kicks.

As presented in section 16.2, our approach distinguishes between the data model and the retrievable properties. This allows the disclosure of concepts across different media. Therefore disclosing information from multimedia rather than mono-media offers a better disclosure. The price is increased complexity. This work has been partly done in co-operation with the Dutch telecommunications provider KPN. A prototype has been built [WVP+98].

Figure 16.12 shows how the ADMIRE model is applied to soccer coverages. Every soccer match, summaries and complete (= 'live-taped') coverages may exist. Within a TV-coverage we distinguish the following IO classes: *scene* (semantic coherence), *shot* (continuous sequence), *video shot* (shot's video part), *audio track* (shot's audio part), *text title* (e.g. actual score, play time, name of player that received a yellow card), and *frame* (a 2-dimensional image sample).

Figure 16.12: Hierarchy of IOs within a TV-coverage composite IO

Property extraction

Like was described in section 16.3, inference percolates through the information hierarchy in a bottom-up manner. We start with the extraction of features from basic IOs, frames, text and audio track. However, many phenomena are not present at the basic IO level. E.g. changes in scores in the text domain and long sentences in the audio domain. So at higher aggregation levels new features are extractable. A user friendly system must links these features, like roundness, to concepts like “ball” or “head”. Some of these concepts may be extracted at a basic IO level. In general, concepts become more pronounced at higher levels of abstraction with an increasing spatial and temporal window. The inference of, e.g. the concept ‘ball’, should thus be delayed. This is illustrated by Figure 16.13, with two candidate pseudo frame IOs, ‘1’ and ‘2’, that share features at the frame level, but differ in their motion features at the video shot level. ‘1’ moves and is actually a ball, while ‘2’ is a spot on the camera lens.

Figure 16.13: A video shot IO consisting of two successive frame IOs. The frame IOs contain multiple pseudo frame IOs. Corresponding pseudo frame IOs, e.g. indicated by ‘1’, form pseudo video-shot IOs.

Property aggregation

Now we show how this process is supported quantitatively as introduced in section 16.3. A shape feature, circularity, is a piece of evidence for the concept ‘ball’ that is equally strong for the four pseudo frame IOs *pfio i.1*, *pfio i.2*, *pfio j.1*, *pfio j.2* belonging to the frames *i* and *j* in Figure 16.13. So their bunches of evidence *B* contain the same piece of evidence e.g., *pfio i.1.B* $\ni$ {‘ball’, circularity, 0.6}.

The motion features differ for the two pseudo video shot IOs, *pvsio 1* and *pvsio 2*, and so do their bunches of evidence, e.g., *pvsio 1.B* = { {‘ball’, frame, 0.6}, {‘ball’, moving, 0.3} } and *pvsio 2.B* = { {‘ball’, frame, 0.6}, {‘spot on lens’, sticking to lens, 0.9}, {¬‘ball’, ‘spot on lens’, 1} }

Suppose a priori belief for an *arbitrary* pseudo video shot IO is $P[\text{ball}] = 0.01$. Then the belief values for the concept ‘ball’ for the pseudo video shot IOs 1 and 2 diverge: $P[\text{‘ball’} \mid pvsio\ 1.B] = 1 - (1-0.6)(1-0.3)/(1-0.01) = 0.72 > P[\text{ball} \mid pfsio\ i.1.B]$ whereas $P[\text{‘ball’} \mid pvsio\ 2.B] = 0.6 \times (1-0.9) = 0.06 \ll P[\text{ball} \mid pfsio\ i.2.B]$ and so do the belief values for the concept ‘spot on lens’: $P[\text{‘spot on lens’} \mid pvsio\ 1.B] = 0$ and $P[\text{‘spot on lens’} \mid pvsio\ 2.B] = 0.9$.

Eventually queries like ‘show all corners that lead to a goal’ must be addressed. Such a query can be formulated in SQL-like expression in terms of high level properties:

```
SELECT <scene> FROM <TV-coverage> WHERE
<scene.‘Approved_Goal’ AND shot.‘Corner_Event’> ORDERED BY <
Function( $P[\text{‘Approved_Goal’} \mid \text{scene}.B]$ ,  $P[\text{‘Corner_Event’} \mid \text{scene}.B]$ )>
```

The wanted scenes in this query have an approved goal and a shot with a corner event. We show how the concept ‘Approved_Goal’ is inferred. First of all evidence for the concept ‘Approved_Goal’ comes from shot y1 in scene-x where this goal shot is visible:

$\text{scene-x}.B \ni \langle \text{‘Approved_Goal’}, \text{shot-y1.‘Approved_Goal’}, 1 \rangle$

But also other evidence might add to the belief in the concept ‘Approved_Goal’ e.g. ‘replay’, ‘close-up’, ‘cheering cluster’, ‘cheering public’, ‘approved text goal’. This is visualised in the Figure 16.14.

Figure 16.14: Example of evidence for extracting the concept “real-goal-opportunity”

Temporal relations between these pieces of evidence are vital. For example a referee’s whistle just before a goal shot makes the goal shot invalid. We follow the ‘Real_Goal_Event’.

shot-y.B $\ni$ { 'Real_Goal_Event', { video shot-p. 'Real_Goal_Opportunity',
audio track-q. 'Public_Real_Goal_Cheering' } AND video shot-p
TEMPORAL WITHIN THRESHOLD AFTER audio track-q, λ }

In this knowledge rule two phenomena, namely a 'Real_Goal_Event' and a 'Public_Real_Goal_Cheering' are possible pieces of evidence from different media. If they occur together, the latter must follow after the first within a certain time, e.g. a few seconds. Note that the conjunction of concepts is accounted for by multiplying the belief values:

$$P[\text{'Real_Goal_Event'} | \text{shot-y.B}] = \lambda \times P[\text{'Real_Goal_Opportunity'} | \text{video shot-p.B}] \times P[\text{'Public_Real_Goal_Cheering'} | \text{audio track-q.B}]$$

We follow the 'Real_Goal_Opportunity' concept:

video shot-p.B $\ni$ { 'Real_Goal_Opportunity', pseudo video shot-x. 'Ball'
AND pseudo video shot-y. 'Goal_Area' AND SPATIO TEMPORAL
OVERLAP DURATION (pseudo video shot-x, pseudo video shot-y)
Threshold, λ_2 }

So two pseudo video shot IOs must be identified as 'Ball' and 'Goal_Area' respectively, and they must share the same spatial position over a certain minimal period of time. So a ball passing by quickly does not make the system infer a 'Real_Goal_Event'.

Finally we arrive at a concept we inferred already, 'ball'. In this brief overview it is evident that information from different media is easily combined in one framework.

The prototype is implemented as a client/server application using the Informix Universal Server (IUS) as the object-relational database platform. The clients are implemented on Windows NT 4.0 machines using Delphi and IUS query tools. All Soccer data is stored at the server side. The clients contain several applications for manipulating the data. The Universal Server is extended with a number of software libraries called DataBlades that provide data storage and management functionality. The prototype uses DataBlades for video (Informix), image (Excalibur), audio information retrieval (AIR by Musclefish) and text (Excalibur) as presented in [WVP+98].

The current prototype implementation supports inference. Automated tracking of corresponding frame IOs is supported. Also, several features have been predefined in the software libraries for instance audio features others have been specially defined for the prototype. At the moment missing algorithms are replaced by manual annotation. Improving this inference is one of the major topics we currently work on.

16.6 CONCLUSIONS

We have shown the importance of structuring multimedia information (information supply) and its effect on content analysis and retrieval (information demand). The ADMIRE framework was presented where all types of media can be modelled in a uniform way. Beside the three level structure of information within an Information Object, the information object itself is part of a hierarchical structure that is imposed by the application domain. There are three types of Information Object, namely basic, composite and pseudo. Further, a set of relations are defined between Information objects. This is explained in section 16.2. There are different ways to label concepts in an Information Object. This may vary anywhere between fully-automated and manual. The ADMIRE framework offers the freedom to choose the appropriate labelling method. Labels generated automatically or semi-automatically should include uncertainty measures as algorithms have finite precision. Our approach using uncertainty was presented in section 16.3.

We have also demonstrated that structuring the information depends on the context. This has been shown in the two cases presented in sections 16.4 and 16.5 where the ADMIRE framework has been applied. In the first case, data from patients with movement disorders was modelled. Here the medical practice was used for defining the hierarchical structure, property extraction and property aggregation. In the second case, a soccer VDS, the hierarchical structure was easier to identify. This structure is a general VDS. Yet property extraction and aggregation needs some insight in the soccer domain.

For this soccer VDS case we wanted to reach two goals, first we wanted to mingle manual labelling and automated labelling. At the moment many algorithms that extract features are still under development while they improve every moment. Scene detection for example is already common in MPEG editing software. For these algorithms manual labelling was used in the prototype.

In the two cases presented in this paper we have gained field experience which is recapitulated below.

Modelling hierarchical structure was relatively straightforward compared to defining features and concepts. For example, in the clinical assessment case domain expertise is necessary to identify the proper features to be extracted from measurement data and to infer the proper concepts combining these features.

Combining multimodal information is important. Uni-modal information is frequently incomplete and sometimes error prone. Combining different modalities may solve these problems. In the clinical case we saw that EMG graph and motion information are both required for generating analysis information of a patient.

Extensibility of the model is an advantage for supporting new forms of content disclosure. For example extending the hierarchical model of the VDS

with Tele-text information should be a straightforward operation. Further, reuse of information for other purposes is supported. For example, methods to recognise a player in a soccer game can be used for volleyball games as well.

Answering queries is a top-down process where concepts are first identified high in the hierarchy in combination with pruning mechanisms. Whereas concepts and feature extraction is a bottom up processes. From our experience it is advisable to infer concept higher-up in the hierarchy to guarantee higher certainty.

Up till now the ADMIRE framework has been shown to be widely applicable. For example it has also been applied in the hypermedia domain. In [VeEe98], and [Velth98] a retrieval service for the web is presented where users can specify time and/or cost constraints over their queries.

Future challenges are to investigate scaleable and generic aspects of the ADMIRE framework. This a topic for future research at our institute. Think for example of applications for large user groups where collaboration between members is supported on very large distributed databases.

Acknowledgements

The authors would like to thank René Bakker and Henk Eertink for reviewing this paper. The KPN is acknowledged for the contribution in the soccer VDS case. The Roessingh and the faculteit der geneeskunde from the AZVU are acknowledged for the contribution in the medical case study.

References

- [AnHR94] André, E., Herzog, G. & Rist, T. Multimedia Presentation of Interpreted Visual Data, *Proceedings of the AAAI Workshop on "Integration of Natural Language and Vision Processing"*, Seattle, WA, 1994
- [Badl75] Badler N.I., Temporal Scene Analysis: Conceptual Description of Object Movements, Technical Report 80, Computer Science Department, University of Toronto, 1975
- [BCP+79] Bekey G.A., Chang C.W., J. Perry, and M.M. Hoffer. Pattern-Recognition of Multiple EMG Signals Applied to Description of Human Gait. In *Proc. Institut. Electrc. And Electron. Eng.* Vol. 65, 1979, pp 674-681
- [BuSh84] Buchanan, B.G. and E.H. Shortliffe,, 1984, *Rule-Based Expert Systems: The MYCIN experiments of the Stanford Heuristic Programming Project*, Reading, Massachusetts, Addison-Wesley
- [Clar88] Clarke, M.R.B., 1988, Discussion to "Belief Functions" by Smets, P., In: Smets, P., et al (eds.) *Non Standard Logics for Automates Reasoning*, London, Academic Press
- [CMF96] Chiamarella, Y. P. Mulhem, F. Fourel, Clips: A Model for Multimedia Retrieval, internal report, IMAG, Grenoble.

- [GuRV96] Gudivada V.N., V.V. Raghavan, and K. Vanapipat, A Unified Approach to data Modelling and Retrieval for a Class of Image database applications. In *Multimedia Database Systems*, Springer, pp37-78, 1996
- [GuNe93] Gu J., and E.J. Neuhold, A data model for multimedia information retrieval. *Proc. Multimedia Modeling*, Singapore, 1993, p. 113-127
- [GSC+95] Gong Y., L. T. Sin, C. H. Chuan, H. Zhang, and M Sakauchi, Automatic parsing of TV soccer programs, *Proc. Of IEEE ICMCS'95*, International conference on multimedia computing and systems, pp. 167-174, 1995
- [Hard98] Hardman, L., Modelling and Authoring Hypermedia Documents, Ph.D. Thesis, University of Amsterdam, 1998.
- [HaSc94] Halasz, F. And M. Schwartz, (1994), The Dexter Hypertext Reference Model, *Communications of the ACM*, 37 (2), Feb, pp 30-39.
- [Haut97] Hautus E., Sybar A Human Motion Analysis System for Rehabilitation Medicine. PhD Thesis TU Eindhoven, November 1997
- [Heck86] Heckerman, D., 1986, Probabilistic Interpretation of Mycin's Certainty Factor Model, In Kanal, L. N., & Lemmer, J.F. (eds) *Uncertainty in Artificial Intelligence*, Elsevier Science Publishers (North Holland)
- [HoLu98] Houtsma M.A.W. and H.J. van der Lugt, Applying advanced multimedia concepts in health and education. In *Advances in Information Technologies: In Proc. of EMMSEC97*, Florence, November 3-5, 1997
- [InBo95] Intille, S. and A. Bobick, Visual, Tracking Using Closed-Worlds, *Proc. of the Fifth International Conference on Computer Vision*, MIT, Cambridge, MA, pp. 672-678, 1995.
- [ISO94] International Standards Organisation, *Presentation Environment for Multimedia Objects (PREMO)*, ISO/IEC 14478, 1994
- [ISO97] International Standards Organisation, *HyTime Hypermedia/Timebased Structuring Language*, ISO/IEC 10744, 1997
- [KnRi79] Knutsson E., and C. Richards. Different types of Disturbed Motor Control in Gait of Hemiparetic Patients. In *Brain*, Vol. 102: pp 405-430, 1979
- [KrCl93] Krause, P.J. and Clark, D.A., 1993, *Representing Uncertain knowledge, An Artificial Intelligence Approach*, Dordrecht, Kluwer Academic Publishers
- [LaSp88] Lauritzen, S.L. and Spiegelhalter, D.J., 1988, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Proc. Royal Stat. Society*, B.50, 157-224
- [MPEG98] MPEG Requirements Group, "MPEG-7 Requirements Document", Doc. ISO/MPEG N2083, MPEG San Jose Meeting, February 1998.
- [Okad79] Okada N., SUPP: Understanding Moving Picture Patterns Based on Linguistic Knowledge, In *Proc. IJCAI*, Tokio, Japan, 1979, pp 690-692
- [Pear90] Pearl, J., Bayesian and Belief-Function Formalisms for Evidential Reasoning: a Conceptual Analysis, In: Shafer G. And Pearl, J., *Readings Uncertain Reasoning*. San Mateo, Morgan Kaufmann, 540-574.
- [Perr92] Perry J., Gait Analysis, Normal and Pathological Function. SLACK inc. Thorofare NJ. 1992
- [Rijs79] Rijsbergen, C.J. van, *Information Retrieval*, Butterworth, 2nd edition. London 1979.

- [ScWy95] Schloss G.A., and M.J. Wynblatt, *Providing definition and temporal structure for multimedia data*. Multimedia Systems. vol. 3, no. 5/6, 1995, p. 264-277
- [Shaf78] Shafer, G. *Non-additive probabilities in the work of Bernoulli and Lambert*. Archive for History of Exact Sciences, 19, pp. 309-370, 1987.
- [TYH+91] Tsuda K., and K. Yamamoto, M. Hirakawa, and T. Ichikawa, *MORE: An object-oriented data model with a facility for changing object structures*. IEEE on Knowledge and Data Engineering, vol. 3-4, '91, p. 444-460
- [VeBE96] Velthausz, D.D., C.M.R. Bal, and E.H. Eertink, *A Multimedia Information Object Model for Information Disclosure. MMM'96 proceedings of the Third International Conference on MultiMedia Modelling*, Toulouse, France, 12-15 November 1996, pp 289-304
- [VeEe98] Velthausz D.D., H. Eertink, *Meta-data for resource-limited retrieval*, paper submitted to *ACM the SIGIR'98 workshop on Hypertext Information Retrieval for the web*, Melbourne, Australia, August 28, 1998.
- [Velth98] Velthausz D.D., *Cost-effective networked based multimedia information retrieval*. PhD. thesis to be published in 1998.
- [WiEd88] Winterfeldt, J. von and W. Edwards, 1988, *Decision Analysis and Behavioural Research*, London, Cambridge University Press
- [WiGH87] Winters T. F., J.R. Gage and R. Hicks. *Gait patterns Spastic Hemiplegia in Children and Young Adults*. In *Journal of Bone and Joint Surger* , Vol. 69, No. 3, March 1987 pp. 437-441
- [WNM+95] Wu J.K., A.D. Narasimhalu, B.M. Mehtre, A.P. Lam and Y.J. Gao, *CORE: a content-based retrieval engine for multimedia information systems*. In *Multimedia Systems*, Vol. 3, No. 1, 1995, p. 25-41
- [WVP+98] Woudstra A., D.D. Velthausz , H.J.G. de Poot, F. Moelaert El-Hadidy, W. Jonker, M.A.W. Houtsma, R.G. Heller, J.N.H. Heemskerk, *Modelling and Retrieving Audiovisual Information: A Soccer Video Retrieval System*. *MIS'98*, September 24-26, 1998, Istanbul, Turkey.
- [Zade84] Zadeh, L.A., 1984, Review of Shafer's "A Mathematical Theory of