The Role of High-level and Low-level Features in Style-based Retrieval and Generation of Multimedia Presentations

Frank Nack, Menzo Windhouwer, Lynda Hardman, Eric Pauwels,
Michèle Huijberts,
CWI Amsterdam
Kruislaan 413, P.O. Box 94079
1090 GB AMSTERDAM, The Netherlands
{Frank.Nack,Menzo.Windhouwer,Lynda.Hardman, Eric.Pauwels, Michele.Huijberts}@cwi.nl

ABSTRACT

In this article we argue that the automatic generation of dynamic multimedia presentation requires both low-level collections of objective measurements for media units representing prototypical style elements, and high-level conceptual descriptions supporting contextual and presentational requirements. Only the combination of both facilitates the retrieval of adequate material and its user-centred presentation. We discuss the problems of visual signification for images in dynamic systems and explain how a combined approach can help overcome such problems. We then propose an architecture for such a system and present its applicability for a museum-oriented multimedia system with a working example.

Keywords

Multimedia semantics, feature grammars, style-based multimedia presentation generation, multimedia retrieval

1. INTRODUCTION

Over the last decade, museums have been contributing to the information revolution by digitising their collections and providing access to them for the general public. Currently, many sophisticated and visually pleasing environments, such as exhibitions in virtual galleries, art collections and presentations of different cultural artefacts, are available. A problem with these environments is that they are handcrafted. In nearly all cases, they lack adaptive [8] or adaptable qualities [28], which could otherwise facilitate the adjustment of the multimedia presentation to the specific context of an individual user. For overcoming these problems, various attempts to explore and develop innovative presentation techniques are described by [1, 2, 6, 20, 36]. These approaches facilitate the synthesis of multimedia documents and plan how this material is presented to various users. The underlying assumption of these systems, however, is that all material and its combinatorial possibilities are known.

In dynamic environments, such as web-based museums, where neither the individual user requirements nor the requested material can be predicted in advance a top-down planning approach is not sufficient. Instead, we claim that a system must be provided with knowledge of simple codes, i.e. collections of objective measurements for media units [10, 19] representing prototypical style elements, which are combined with high-level conceptual descriptions [31] supporting contextual and presentational requirements. Using such a combinatorial approach it is possible to establish conceptual presentations that support a better understanding of art, so that the system can find satisfactory solutions for upcoming questions (e.g. based on the content of an image), misunderstandings (rearrangement of the material) or non-understanding (creation of a new sequence).

The mapping of high-level conceptual structures with low-level feature descriptions as an essential mechanism to enhance the automatic generation of dynamic style-oriented multimedia environments is part of the methodology followed in the Cuypers project [34]. The aim of this work is to improve existing search techniques on category and image features to facilitate user-centred recall even if relevant semantic information in the form of keywords or schemata is only partially available, and to combine these with presentation independent as well as presentation related knowledge in order to improve the quality of the generated presentations.

In this article we first present a brief analysis of the presentation problem for dynamically arranged visual artefacts. We then present the framework of our prototype multimedia generation environment, which facilitates the combination of high-level concepts and low-level feature descriptions to improve the retrieval of user-centred material and enhance its arrangement based on style-oriented presentation techniques. We describe the underlying architecture and provide a working example. The article concludes with an evaluation of the presented approach and an outline of further research.

2. THE COMPLEXITY OF VISUAL PRESENTATION

The aim of museums is to exhibit, preserve and support the study of historical, artistic or scientific artefacts. As far as the exhibition is concerned, usually the space is limited and thus mere fractions of the collection can be presented at a time. Hence, most museums use the Web for presenting large parts of their collections to the general public. On the whole, the interfaces are visually pleasing but the common way of providing information is a static combination of text and different sorts of visual data, where the text conveys the main information. Most museums also perceive themselves as educational institutions – offering the visitor some initial information on their collections and thus enabling them to consult more specialized sources.

For example, the Rijksmuseum (http://www.rijksmuseum.nl/) in Amsterdam provides information on 1,200 of its top exhibits, in the form of texts, photos, video and animation. Links enable objects to be connected from different departments of the museum's vast collection, for example associating a painting of a 19th-century landscape with its 17th-century forerunner. Access is via four categories: artists' names, themes, encyclopaedic terms and a systematic catalogue.

The current system depends on the designer's vision of how the different media items should be presented so that a typical user can understand the relationship between the different units and the ideas they represent, as well as the differing meanings that can be attributed to the objects within the media units. The challenge for the designer of such a media-based information environment is to foresee the circumstances and presuppositions of the user at the time of accessing the information.

From the point of view of visual media this process of human perception is the core of the presentation problem. If we look at visual material, such as single images or video, as an abstract element, we can experience it on two levels: optically (objectively, realistically) and mentally (subjectively). On the optical level we try to identify as many objects as we can in the available time of perception [3]. On the mental level we then transform the identified object into a meaningful unit, or a sign. A sign usually consists of two distinguishable components: the signifier (which carries the meaning) and the signified (which is the concept or idea signified) [14]. The relation between the signifier and the signified is arbitrary, which enables the creation of higher order sign systems. It should be stressed here that the diversity of such semiotic-based cultural-rooted sign systems [14, 16] their combinatorial potential [7, 13], provide a continuum of meaning that forms the basis for a subjective interpretation by each viewer.

Consider FIG. 1 on the next page, which shows Rembrandt's ‘Philemon and Baucis’. In this image Rembrandt made use of the technique *clair-obscur* or *chiaroscuro*. Both terms mean 'light-dark' and are used to describe strong contrast of light and dark shading in paintings, drawings and prints. On a perceptual level this technique indicates spatial orientation and, due to its strengthening of realistic presentation, a better identification of objects.

Rembrandt, however, uses the presentation of light not only for technical means but also as a way to portray meaning on a metaphorical level. Conceptually speaking, light represents for him either a religious symbol or the means of portraying the centre and the range of a narrow world. Light used as a religious symbol has typically its source outside the portrayed scene and the objects within are merely reflections of it. The observer is thus led to believe that the world is dependent on the existence of an unknown and invisible power [3]. Light used as the centre of an image demonstrates the limitations of our perception of the world. Here, the viewer is confronted with the insignificance of the individual within the world, if not the universe.

FIG. 1: Rembrandt's Philemon and Baucis. (1658), The National Gallery of Art, Washington, DC, USA, url: http://www.nga.gov/cgi-bin/pinfo?Object=1207+0+none

Thus, while the characterisation of visual information on a perceptual level using objective measurements, such as those based on image processing or pattern recognition, facilitate the accessibility of the image on content level, it is the myriad of cognitive and cultural codes that allow the access of its meaning.

While a single image is able to trigger multiple interpretational levels, the use of numerous visuals simultaneously in one presentation is even more problematic [13]. Humans tend to search for cues as support for the creation of meaning and for associative cues that facilitate the combination of indicators.

Assume a user read about Rembrandt on one of the museum's web pages and is now interested to learn more about *clair-obscur*. A presentational option is to display a number of paintings in *clair-obscur* style. Cues to guide the user in understanding the style are, for example, a title stating the main purpose of the presented images. However, we could also present the images in the style they are drawn in. For this, we have to judge the order of the paintings depending on the way the dark and light parts are distributed in each painting for coming up with a reasonable arrangement of images. In other words, we use, among other indicators, space as an associative cue. In fact, the conceptualisation of space is, therefore, an elementary principle of the analysis and organisation of material in the presentation of multimedia presentations. In addition, aspects of the style in the underlying content can also be carried over into the temporal dimension, e.g. in the *clair-obscur* example by fading each image completely to black before displaying the following image.

The countless presentational strategies mainly depend on the ability of the user to comprehend or at least appreciate the content, contextual and stylistic level of the presentation. Given the unpredictable nature of the mirroring relationship between user and presentation in a dynamic environment, it is infeasible to create all relevant media documents in advance. However, a presentational system based on knowledge on how to collect objective measurements for media units [10, 19] and interpret and manipulate them using high-level conceptual descriptions [31] and style-oriented presentation mechanisms [27, 34], it is possible to react adequately to emerging information needs satisfying various levels of the complex conceptual understanding of visuals. For the above problem of presenting images in the clair-obscur style this means that a free interaction with the user on a rhetorical, e.g. educational basis, is possible. This means, a fairly vague information request, i.e. about a particular style, can be answered in such a way that not only the appropriate material can be displayed but that also the conceptual core element of the request, i.e. style understanding, can be reinforced as a substantial part of the presentation by applying the particular style as presentational technique.

Moreover, we are now in the position to increase the amount of relevant metadata automatically and thus improve the recall. For example, it might be possible that the user would like to see illustrations of all different styles Rembrandt used during his life. Usually, a number of high-level conceptual representations will be associated with an image, such as painter, date of painting or title of painting. However, the amount of this useful data is usually limited because it requires manual labour – an expensive endeavour normally not covered by the archival budget. If images have no further annotation attached than 'Artist = Rembrandt' we would not be able to classify the retrieval results according to the

query. Having access to the specific representation of intrinsic features of a style, we can now analyse an image during the retrieval process and indicate which style is the most appropriate for this particular image.

In the rest of the paper we describe an approach that addresses the initial requirements for dynamic style-centred multimedia presentations, as outlined above. The aim of this work is to improve existing search techniques on category and image features, and to combine these with presentation independent as well as presentation related knowledge in order to improve the quality of the generated presentations.

3. A FRAMEWORK FOR COMBINING HIGH-LEVEL CONCEPTS AND LOW-LEVEL FEATURES IN THE AUTOMATED GENERATION OF MULTIMEDIA PRESENTATIONS

We incorporate automatically extracted image features in our existing framework for generating hypermedia presentations [34]. The architecture as described in FIG. 2, consists of three major units:

- the style repository, which embodies style schemata, style grammars and rule-bases for different presentation styles
- the data repository, containing the images and related metadata, and the retrieval engine
- the presentation environment, including a presentation generator and a hypermedia browser.

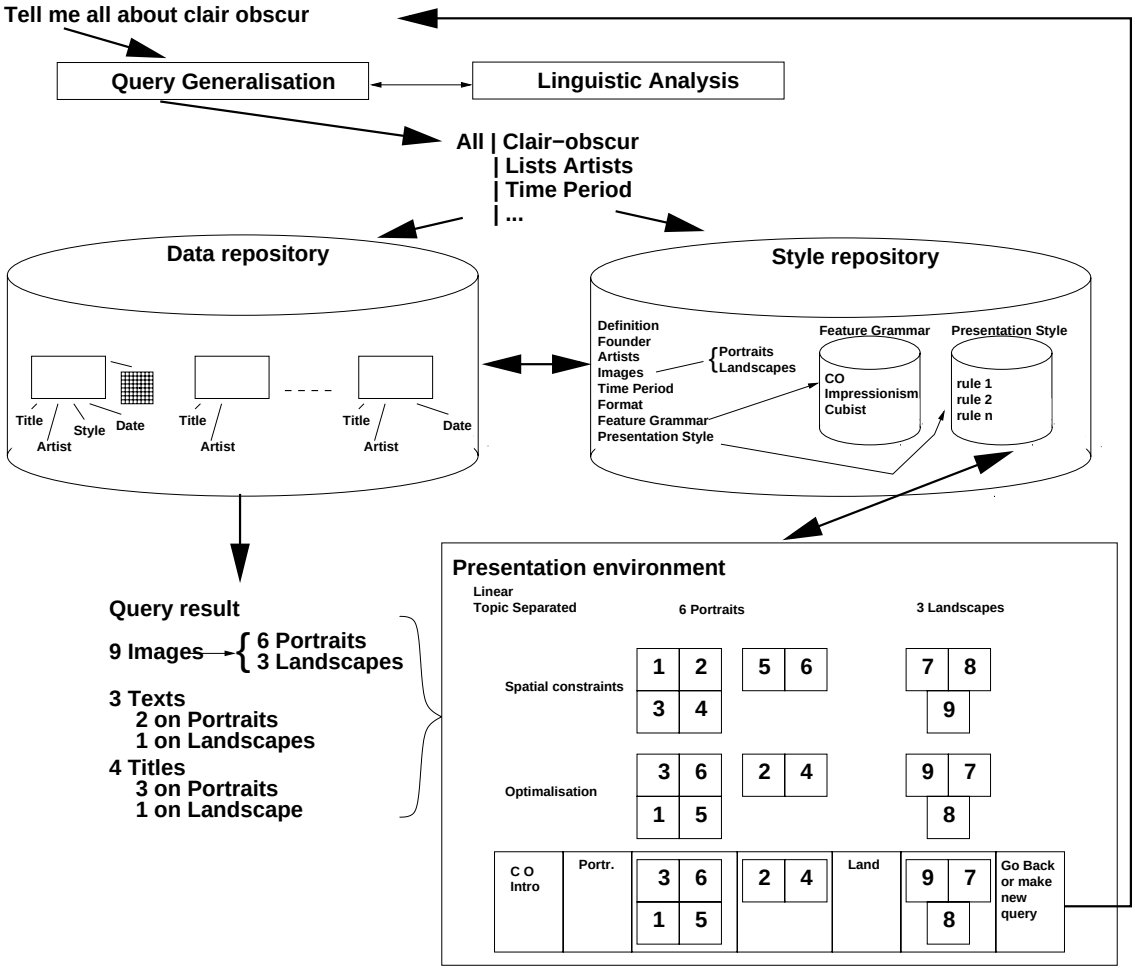

FIG. 2: General scenario framework

We give an overview of the style repository, the data repository and the presentation environment, and the ways in which they interact. In section 4 we discuss an example of how the described units behave with material from the Rijksmuseum in Amsterdam.

3.1. The style repository

The aim of the style repository is twofold. Firstly, it provides a collection of representations describing styles in fine art, such as *clair-obscur*, impressionism or cubism, in a structured way. These collections are designed to improve the retrieval of images or other metadata in the data repository (discussed in the next section). Secondly, it provides a set of rhetorical structures designing the overall organization of the presentation and a style-based rule set determining how relevant material can be presented.

The collection of representations of fine art styles provides schemata for each style. A schema holds information about the definition, the main period, the inventor of this style, other artists using or improving it, etc. The advantage of these style ontologies is that they allow an enlargement of the search-space, if the style plays a prominent part in the query. This state-of-the-art technology within the retrieval community [31] is well supported [17].

This information, while important, is insufficient, because it does not usually cover compositional data, such as colour distribution or shape analysis, that characterise a style or a painter [16]. Such low-level feature descriptions are also required. Rather than a random choice of features such as a style description, features that represent the intrinsic characteristics of a particular style need to be collected. In *clair-obscur* images, for example, we find a clear distinction of light and dark areas. Usually there is one dominant light source, predominantly filled with high luminance colours, alongside dark areas with a high proportion of brown colours, which can be blended with other objects [3].

Thus, a collection of features such as colour, shape, brightness, either in the form of their extraction algorithms or as threshold values for a particular style, facilitate the automatic identification of relevant material. Such a collection we call a feature grammar, and its functionality will be explained in more detail in section 4.

Note that the development of feature-based representations also requires human effort, in particular by specialized experts who have an understanding of the compositional structures of an image, as outlined in FIG. 1. The collection of these features is, on the other hand, manageable, since tools for this particular task already exist [15, 32].

The rhetoric rule set describes high-level presentation structures, as addressed in the Rhetorical Structure Theory [22] or Cognitive Coherence Relations [21], which vary between general and specialised levels. If, for example, the presentation environment is educationally oriented, the system might decide on a linear presentation in the form of a slide show, or a more interactive presentation in the form of additional buttons for individual traversal. Another set of rules can then determine for a sequential organization of material that might result in a structure of the form: Introduction Topic; Introduction Subtopic 1; Details Subtopic 1; Introduction Subtopic 2; Details Subtopic 2; Introduction Next Topic.

The presentational rule set addresses the spatial, temporal and stylistic organisation of the material. In particular the stylistic oriented rules relate to the established rhetorical structure by transforming the available media units into the appropriate form. For example if the applied rhetoric structure requires a reinforcement pattern, a stylistic rule on repetition might be utilized, such as

Repetition: use the image of the topic introduction as background for subtopics,
make it single coloured,
where
the colour is the main colour required by the style or
provided by the colour histogram and
does not interfere with the spectrum of background colours for the font of the text

Development environments, which support the design of rhetoric and design rule sets are described in [5, 24].

3.2. The data repository

The repository, as described in FIG. 2, stores annotation schemata in the form of XML-based documents and media-based data, such as images in various formats (pic, gif, tiff, etc.). The repository itself can be realized using federated database technology.

The annotation documents for high-level descriptions are created by experts, using ontology-based technology [33] for task-specific controlled vocabulary/subject indexing schemata for in-depth semantic-based indexing of various media. The ICONOCLASS system [18] as an example for the arts domain exemplifies how in-depth indexing can be achieved. Note that annotation schemata are different from the style representations. Annotation schemata provide information about one particular image or artist. For example, they capture information about the title of an image, its painter, production date, a list of

exhibitions where it was presented, reviews, and so forth. The annotation process follows a strata-oriented approach, which allows a fine-granulated space-oriented description of media content, where particular areas within an image can be annotated. The connection between description and media unit can be based on linking mechanisms as described in XML path and pointer [9, 11] or MPEG-4 [19].

The metadata schemata describing media-based data are predominantly generated automatically. Having access to the specific representation of intrinsic features of a style, as defined in the feature grammar, it is now possible to analyse an image during retrieval process and collect relevant information, e.g. the colour distribution in form of a histogram represented in XML. Other representations are also possible, depending on the query. Imagine the query includes the search for particular presentation relevant representations of visual material, such as a grid representing the light distribution of an image. If a source is identified as relevant but it turns out that the grid representation is not available, it can be additionally generated. Thus the query can be successfully answered and the search space for the particular image is increased for future use, if a similar request is issued.

We see the result space, as indicated in FIG. 2 by the area titled “Query result” as part of the data repository. That is why we provide some information about it here. Since the main goal of the proposed framework is to facilitate the automatic generation of user-centred multimedia presentations, we suggest that the result space will not only contain the retrieved data, associated metadata, and the relations between these different units but also information required for their presentation. Moreover, it also returns physical information about the retrieved data, i.e. image size and image file type.

3.3. The presentation environment

The presentation environment, as displayed in FIG. 2, is basically a constraint-based planning system, using the definitions provided in the style representation schemata and the presentation styles [34]. Since the system can access descriptions based on spatial, gradient and colour features, the presentation generator is in the position to analyse the retrieved material based on the relevant presentation design, according to design issues such as graphic direction, scale, volume, depth, shapes (i.e. physical manipulation of the material for better integration into the presentation), temporal synchronisation (interactive or linear presentation), etc. and provides a format that a hypermedia browser can interpret, e.g. SMIL or MPEG-4 [19, 35].

Having introduced the main units within our framework, we are now in the position to explain the actual processes by means of a scenario. The scenario illustrates the current state of our prototype environment at CWI.

4. SCENARIO

Imagine a visitor of the web environment of the Rijksmuseum is reading the associated text of a painting. The visitor is intrigued by the technique *clair-obscur* and would like to know more about it and poses the following query: Tell me all about *clair-obscur*.

4.1. Query generation

The query of the user is transformed based on simple linguistic syntax decomposition into two main units: a measure unit to describe the amount of data to be retrieved, and a topic unit, to explain what content needs to be searched for. The result for the example query is 'all' as a measure and *clair-obscur* as a topic. The former is interpreted by the system as: full database search on Topic, and results in the relevant qualifiers for the final query. The topic is analysed if it falls into one of the categories: Artist, Style or Art-Type, for which conceptual representations are available. Since *clair-obscur* is a style, the style representation is addressed, which covers the following information:

- Definition, a text describing the style;
- Founder, which provides the name of the artist who came up with the first image of this style;
- Artists, a selection of artists using this style;
- Images, a list of the most important examples of this style;
- Format, presents a list of typical formats, such as landscape, portrait, still life, etc.;
- Grammar, a link to a collection of feature detectors;
- Presentation style, link to a top-level rule in the presentation rule base.

The query generator extracts the inventor, artists, images and format elements and adds them as additional attributes to the topic. Finally, the query generator adds information from the simple user model to the query, mainly the user's preferences on information types (text, audio, video) and level of detail. This collection forms the basis for the query, which is represented in W3C's 'XML query' database format [12]. Definition, grammar and presentation style are marked as existing. This is important because related activities will only be performed if these mechanisms are available.

4.2. Retrieval

In the context of this paper, we are primarily interested in the retrieval of simple concepts and will thus ignore the retrieval built on text-based content descriptions.

If we interpret the database, or a collection of databases, as the complete search space, then we can divide this space in three areas:

- material with annotations that provide evidence that the main topic is identifiable. For our example this means that the style is provided. If the image fulfils the query further processing on this image is not necessary. However, if additional information, such as brightness histogram, segmentation representation or grid abstraction cannot be retrieved the feature grammar will be activated for producing these items;
- material with annotations that allows a relation to the required topic through one of the additional attributes but provides insufficient evidence that the topic is part of the image. An example would be an image that fits the time period and an artist but has no identification about the style. This is when the feature grammar is applied. A positive identification of the style results in an annotation of style and the other representations (histogram, segment, abstraction);
- material with no identification. Here the grammar can also be applied. Since the evidence is merely based on the execution of the grammar, extra validation needs to be collected, e.g. the judgment of a human expert.

Since the functionality of the feature grammar is the key to linking low-level auto extraction with high-level style we will now explain it in more detail.

4.2.1. A style feature grammar

When there are no annotations of a multimedia item available, there is still the raw data of the item itself. In research areas such as computer vision, speech recognition, etc. many algorithms have been devised to extract meaningful features from this raw data [10]. These low-level features can then be aggregated into high-level concepts. Features or concepts determined by one algorithm can, of course, be input to another algorithm. Using these chains of algorithms an annotation of the multimedia item could be created semi-automatically.

However, to create this annotation the algorithms should be executed in the right order, so a high-level description of their dependencies is needed. The Acoi system [37], a toolset also developed at CWI, provides the framework to specify this description and to use it to create the annotations. The description is specified in the feature grammar language, which is at the core of the system.

A feature grammar describes the relationships between annotations (so called features), algorithms (so called feature detectors) and mutual detectors in a set of grammar rules. The grammar itself is basically context-free [30] (using the extended notation), but, additionally, some of the symbols are associated with feature detectors.

To build the annotation, the grammar has to be interpreted using the Feature Detector Engine (FDE). The FDE is a simple recursive-descent parser, which calls the detectors and manages a stack of tokens and a parse tree. The main goal of the FDE is to determine the validity of the start rule of the grammar by executing the feature detectors. These detectors will push new tokens on the token stack. The tokens, when they match the grammar, will move from the stack into the parse tree. When the validity of the start rule has been proven the complete annotation is available in the form of a DOM structure.

The feature grammar for a number of styles including the *clair-obscur* style, is shown in FIG. 3. A partial walkthrough of this grammar will explain how the FDE processes it.

```
%start painting(location);
%atom url;
%atom url location;
%atom flt threshold, coverage, contrast,
                co corr, cu corr, im corr;
%atom int column, row, x, y, width, height,
                regions;
%detector matlab::light_segment(location);
%detector matlab::histo_segment(location);
%detector segment(location,
                light.histo_segment.threshold[0]);
%detector global(global.location);
%detector light(light_segment.segment.location,
                shape);
%detector matlab::contrast(general.location);
%detector grid(general.location);
%detector regions(grid);
%detector co_corr(general.location);
%detector cu_corr(general.location);
%detector im_corr(general.location);

%detector bright
                coverage[0] > coverage[1];
%detector dark
                coverage[0] <= coverage[1];
%detector clair obscur
                co_corr > 0.8;
%detector cubism
                 ((contrast > 17.7) and (cu_corr > 0.5))
                or ((contrast > 19.4) and (cu_corr > 0.4))
                or (contrast > 24.3);
%detector impressionism
                im_corr > 0.5;

painting        →general global local style?;
general         →location light segment;
light segment   →location histo segment segment;
histo segment   →threshold+;
segment         → location;
global          →shape features contrast;
shape           →bbox;
bbox            → x y width height;
features        → light;
light           → coverage[2] (bright | dark);
local           →grid regions;
grid            → cell+;
cell            →column row shape features;
style           → co_corr clair_obscur;
style           → cu_corr cubism;
style           → im_corr impressionism;
```

FIG. 3: Styles feature grammar

The start rule of the grammar is *painting*. This declaration also states that the initial token stack needs to contain the *location* token. To prove that the start rule is valid, the FDE tries to prove the validity of all the symbols in the right-hand side of the *painting* rule. The rules are always interpreted in a top-down and left-right manner, so *general* is the first rule to be evaluated. The first symbol in this rule is *location*, which is an atom. If the name of the first token in the stack matches, the token is popped and moved to the parse tree.

The next symbol is a detector, *light_segment*. The *matlab::* prefix indicates that the implementation of the detector is done in Matlab [23]. The default implementation language is *C*. As the FDE knows only

the in- and output of these detectors these kinds of detectors are also called black-box detectors. Contrary detectors like *bright* or *cubism* post a query on the parse tree, which is executed by the FDE itself, and are thus called white-box detectors. The *light_segment* declaration specifies one parameter: the path to the *location* symbol. Paths always refer to symbols, i.e., nodes, in the parse tree, and a cursor to the proper position in the tree is passed on to the detector. The *light_segment* detector takes the original image of the painting and calculates a feature value for each pixel, *i.e.*, the intensity value, and stores this in a new image. The location of this new image is pushed as a token on the token stack, and, because it fits the grammar, will be moved by the FDE to the parse tree. In this way the FDE will continue to encounter atoms and detectors, and move tokens from the stack to the parse tree. When finally the whole annotation is available it can be dumped as an XML document. This document is then used as input for the next step in the presentation generation, as described below.
The feature grammar has a generic setup. For example it is easy to add new features, another localizing schema or other styles to the grammar. In these cases not the whole annotation has to be regenerated, but the FDE can do an incremental parse taking the old parse tree and adding new branches, therefore only executing the detectors for the new or updated rules in the grammar.
In the next section the global flow of execution embedded in this feature grammar is explained.

4.2.2. Style feature detection

Since we aim to design an approach that is data-driven and can therefore operate unsupervised, it is important to incorporate adaptive decision-making algorithms. For instance, in the case of the *clair-obscur* a vague high-level description of the style could be 'a *brightly lit* object or person surrounded by a *dark* background'. To translate this vague conceptual description into an operational low-level feature extractor, we need to assign precise values to fuzzy concepts such as *bright* and *dark*. However, we cannot fix them in advance because these values depend on the context, *dark* and *bright* being defined relative to the rest of the painting.

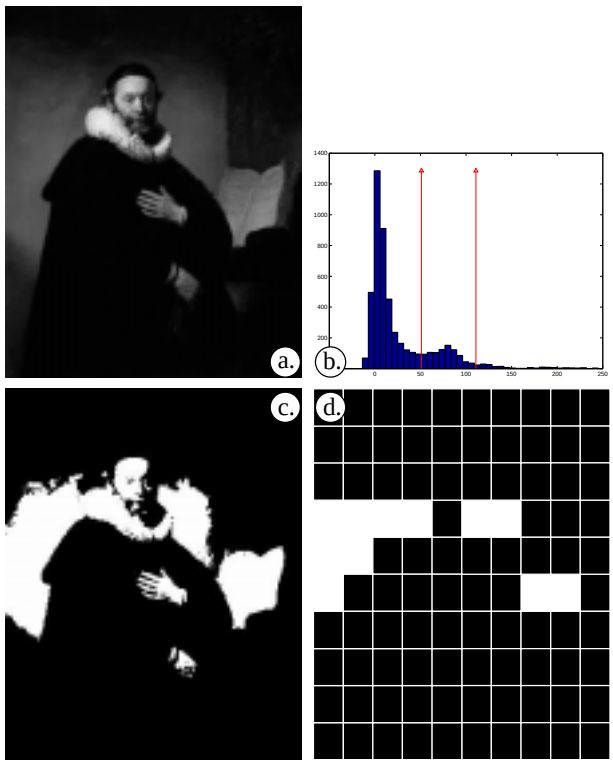

FIG. 4: Feature detection steps

The approach we propose is data-driven in that it inspects the data in search for natural thresholds, i.e. thresholds that are dictated by the structure apparent in the data [25]. To be more precise, assume that we have a numerical image feature x that can be computed at each of the n pixel in the image (e.g. hue, or brightness, see FIG. 4.a). This gives rise to a numerical dataset $x_1, x_2, \ldots, x_n$. To get an idea of how these values are distributed over the image, we can look at the histogram. If, in terms of this feature, the image has a clear structure then we expect to see a multi-modal histogram, with peaks over the most-frequently occurring feature values.

For instance, in the case of *clair-obscur*, computing the brightness histogram we expect to observe (at least) two peaks: one peak at low values created by the pixels in the dark regions, and one at high values corresponding to bright pixels, see FIG. 4.b.

Locating the grey-value at the minimum in between these peaks determines a threshold that can be used to separate the bright from the dark regions in the image. This seemingly simple task is complicated by the fact that a data-histogram almost never has a clear-cut unimodal or multi modal structure, but exhibits many local maxima and minima due to statistical fluctuations. The challenge therefore is to devise a mathematically sound methodology that allows us to construct a smoothed version of the histogram, suppressing the spurious local extrema that unduly complicate the histogram structure.

To this end we introduce the empirical distribution function $F_n(x)$ which for each feature-value x determines the fraction of observations x_i that are smaller than x. The reason for switching to the empirical distribution is that it allows us to compute the precise probability that the given sample is drawn from a theoretically proposed distribution $F(x)$. The idea is simple: we search for the *smoothest* distribution F that is compatible with the data, i.e. such that there is a high probability that the sample $x_1, \ldots, x_n$ has been obtained by sampling from F.

In mathematical parlance this amounts to solving the following constrained optimisation problem: given $F_n(x)$ find $F(x)$ that minimizes the functional

$$\Psi(F) = \int (F''(x))^2 \, dx \text{ subject to } \sup_x | F_n(x) - F(x)| \leq \varepsilon.$$

The value for ε is fixed in advance by specifying an acceptable level of statistical risk. This optimisation problem can be solved using standard spline-fitting routines. Once the shape of the smoothest compatible distribution F is determined, its inflection points can be used to determine the genuine local minima in the histogram, thus yielding natural thresholds for the image-segmentation extractor.

The lowest of these thresholds is then used to segment the image into dark and light areas, see FIG. 4.c. As a next step, information about the areas is localized by overlaying the image with a grid, see FIG. 4.d.

In the feature grammar of FIG. 3 these steps are distributed over several detectors and their dependencies are described. For example, the *light_segment* detector calculates the brightness value of each pixel, this set of values is then taken as input by the *histo_segment* detector to determine the segmentation thresholds.

The grammar defines several other (global) features. For example, the *c_corr* detector computes the normalized correlation between the colour histogram of the painting and two average normalized histograms, respectively for *clair-obscur* and *non-clair-obscur* paintings (see for a similar approach [4]).
All these features form the input for the final step: determinating the style type of the image. For this step a decision tree is derived using C4.5 [26], resulting in the white-box detectors *clair_obscur*, *cubism* and *impressionism*. Notice also that more than one of the alternative *style* rules can be valid, which means that a painting can be annotated with multiple styles, i.e., different views on the same painting. Future versions of the feature grammar tools will provide mechanisms to annotate each alternative view with a belief value, i.e., support for the rule. These belief values form the basis for the ranking of query results.
Furthermore, this grammar is mainly constructed to recognize paintings in the *clair-obscur* style. More features may be needed to fine-tune the decision rules for the other styles, e.g., impressionism. The use of a feature grammar is well suited for this evolutionary approach as it supports incremental maintenance of the annotations.
Based on the above discussion we are now in the position to describe the result space of a query.

4.2.3. Result space

The result space provides a structured description of the retrieved material, represented in the form of an XML document. The structure reflects the types of retrieved data (text, images), the relation between them in the form of triplets (e.g. name-of, author, image, definition-of, style, text, etc.) and additional information regarding valuable presentation information, such as sizes of data units, attached grid abstractions, and so forth. Note, using links refers to the actual data, i.e. the file of an image.
For the sake of presentational clarity, suppose that for the original query 'Tell me all about *clair-obscur*' the retrieval space consists of 6 images of the type portrait, 3 landscapes, and a text containing a general description of the style. For some of the paintings the name of the artist is also available. The final step is the transformation of the retrieval results into a presentation according to the layout functions, described in the rule set of presentation styles.

4.3. Generating the presentation

Part of our prototype implementation is a generation engine that is able to transform a high-level description of a presentation [27] to a final-form encoding that is readily playable on the end-user's system, in our case SMIL [35].

The presentation generation engine of our system is a constraint-based planning system. The constraint system is used for solving the design-based constraints, such as:

- the overall presentation dynamics (e.g. linear or interactive) and the resulting subdivision of information blocks;
- organising material for each information block, e.g. number of elements on a block and their spatial outline based on the actual size of each information unit;
- optimisation of ordered material based on additional style criteria, such as colour or brightness distribution, in particular to emphasize a particular style.

In this paper we focus on the last point. The inner details of the other parts of the system itself, especially the transformation of the presentation structures generated by the constraint engine into a SMIL presentation, have been discussed in previous work [34].

For our purposes, assume that the system constructs a linear presentation consisting of 3 topic blocks to present the material: the introduction, the portraits and the landscapes. Finally, based on spatial constraints, it calculates how many media items for each topic block can be presented on a page.

At this stage the generator tries to arrange the items in each topic block. For the introduction this means for example, that the heading should be at the top right corner, because users expect it to be there. Depending on the user preferences for a medium (visuals versus text) it shows either first the textual definition and right to it the image, or the other way round. For the presentation of the set of paintings the generator tries to fulfil the style criteria for *clair-obscur* because the overall presentation goal requires intensification (see section 3.1.the style repository). A typical decision rule is presented in FIG 5.

```
style_order(Image_Style, List_Of_Images, Images_Per_Page, Presentation_List):
        gradient-match (Image_Style,List_Of_Images, Result__List),
        border-match (Result_List,Images_Per_Page, Presentation__List).
```

FIG. 5: A decision rule determining the order of images based on style.

With this rule the system analyses not the image itself but rather its grid abstraction, as described in FIG. 4d. The system first tries for a particular style (Image_Style) to order the images of one topic (List_of_Images) based on the pattern provided by those cells of the grid that represent light values. The analysis of these patterns is based on graphical shapes, such as triangles or rectangles. The direction of the light is derived from a number of criteria, such as solidness of a pattern (main light centre), position in the grid (at the border indicates that the light source is outside the image), and the direction of the dissolve of this shape (direction of light beam). For FIG. 4d, the result is that the light source is outside the image, that light is coming from the left side and dissolves towards the right side in a rectangular way. The Result_List groups images in lists, where light follows similar directions, such as left, up-left, up, up-right, right, down-right, down, down-left, circular.

Once this ordering is achieved, the system tries to align the images based on similar border pattern. If we take once again FIG. 4d as the example, the system would try to find an image which shares a similar distribution of light and dark cells (up or down by one grid cell) but only on its mirrored side. The combination of images is performed on the previous calculated maximum size of images per page. FIG. 6 and FIG. 7 demonstrate how the unordered presentation looks before and after the transformation.

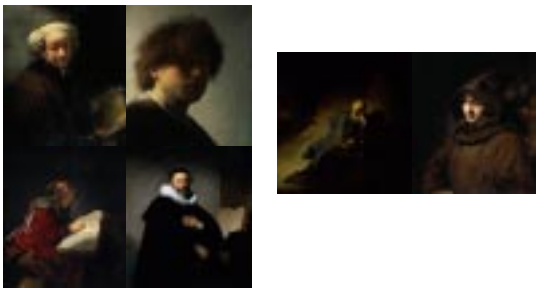

FIG. 6: Ordered retrieval results before optimistion

FIG. 7: Ordered retrieval results after optimisation

Once the order of images is identified, the system can apply rules for placing relevant captions alongside the images. The relevant rule for *clair-obscur* requires that texts for images in the upper row be placed above the paintings, whereas those associated with images in the lower row are placed below. If the system followed the default rule for captions, it would always place them below. This would, however, destroy the clair-obscur effect of the presentation.
After the order of the various blocks has been established, the system finalises the colour for the background and the text. Since *clair-obscur* paintings are usually dark around the edges, the default suggestion for this style is a black background. The default style criteria for text colour associated with a black background is white. The final presentation might look as presented in FIG. 8.

FIG.8: The presentation

The numbers of presented objects and their graphical complexity, the number of words for text elements, or temporal presentation qualifiers such as fade-in or out times, form the basis for calculating the temporal duration for each display. The last screen offers choices for the next step.

4.4. Evaluation of the example and the approach

The current database for our prototype environment contains about 190 images, of which 18 can be identified as *clair-obscur*, 56 as impressionism, and 25 as cubist. Furthermore there is textual information for every image available, covering names of painters, image titles and smaller definitions of styles. Additionally we have an extra set of 83 paintings from the time period between 1600 and 1700 where no extra annotation is available.
Regarding the collection of features in the grammar we do understand that *clair-obscur* is a misleadingly simple style. The grammar is simple even though it covers a mix of brightness, colour level and shape detection on the grid level. Thus, for this particular style we can provide a rather impressive precision ratio, as exemplified in the performance analysis described in FIG. 9.

↓ Number and style of paintings in the database	clair-obscur	cubism	impressionism	unknown
18 clair-obscur	17	-	-	1
25 cubist	6	21	3	1
56 impressionist	1	5	31	20
83 unclassified	70	1	26	4
182 paintings	94	27	60	22

FIG. 9: Results for the style feature grammar

The results for impressionism and cubism are promising but still far from satisfactory. The last column corresponds with the fact that there is no matching style found, i.e., the validity of the *style* rule is optional. Finally, some paintings were annotated with multiple styles, in particular those from the category cubist. We are currently improving the grammar for impressionism and cubism with additional information on colour and shape level, to adapt it to the particular style features.

Of particular interest was for us to test the grammar against the set of 83 paintings from the 17th century. As can be seen most of them are identified as *clair-obscur* images. This result is not too astonishing, since this was a widely used style at the time. However, most of these images represent a mixture of styles where *clair-obscur* is one. More interesting is that roughly one third of these images were classified as impressionistic. So far we are not sure what should be derived from this observation and further investigations with our colleagues from the Rijksmuseum are required here. However, it becomes clear from the above results that the automatic classification of paintings without any extra available annotation requires re-examination by human experts.

We have not tested the grammar on material, which is similar to the various styles. For *clair-obscur* images comparisons with key frames of films of the film-noire genre, or extreme black and white images, as made by the artist Paul Starnd, would be appropriate. We assume that the system would reject them based on the colour histogram correlation. This does not mean, however, that the system could detect the difference between a photograph and a painted image. On the contrary, if the photographs would be transferred into sepia colour, we assume that the system most likely would describe them as *clair-obscur* images, which would not be false per se. In general it would only show the importance of light in art. However, in a presentation about Rembrandt those images would be ill suited because of the missing periodic context. That is why we see the need for extra information within the style schema. A mere low-level description cannot provide more than a indication of what style type an image might have. However, we are intrigued by the idea of using the low level features as one basis for the comparison of styles. We have no illusion that even if we can deal with different colours, textures and shapes, there are still a large variety of styles, which we cannot handle properly. The image 'Lavender Mist' by Jackson Pollock, as portrait in FIG. 10, is a good example.

FIG. 10: Jackson Pollock, Number 1, 1950 (Lavender Mist), 1950, National Gallery of Art, Ailsa Mellon Bruce Fund, Washington D.C., USA (http://www.nga.gov/feature/pollock/painting1.html)

Retrieval for cases where no further annotations are available is rather slow (about 3s per image) and the entire image processing needs to be done for style detection, Under these circumstances the generation of a presentation as described in FIG. 8 can take up to 1 minute. Once material is annotated, including the generation of the abstract image representation, the generation of the final presentation takes a number of seconds.

5. CONCLUSIONS

In this article we argued that for the automatic generation of multimedia presentations we are in need of expandable, descriptions of style, which provide both high-level conceptual and low-level feature extraction information. Only the combination of both facilitates the retrieval of adequate material and its user-centred presentation. We discussed the problems of visual signification for images in dynamic

systems and showed how a combined approach can help overcome such problems. We proposed an architecture for such a system and presented its applicability for a museum-oriented multimedia system.
The described architecture and system behaviour and its theoretical foundation are best regarded as a platform that demonstrates the feasibility of automated stylistic-oriented multimedia generation in narrow, yet complex, domains.
The important question we would like to answer, however, is not so much whether we can detect these styles, but what it would mean to use them for presentation. The *clair-obscur* style showed that the match between background colour, border colours of several images and light distribution within an image can provide a pleasing presentation. Whenever we showed the demo with and without optimisation, users mentioned that they liked the optimised version better. On the other hand if they were asked to determine what stimulated their opinion most viewers could not formulate it. Those who could identify the intensification strategy knew already quite a lot about the style. These preliminary results do not state that intensification strategies won't work *per se* but they focus our research on determining which level of subtlety can or should be achieved in automatically generated multimedia presentations. All we can say for the moment is that the automatically generated presentations in our test environment are pleasing, which is a positive achievement. However, further research is needed to determine how valid our hypothesis is that the automated generation of dynamic presentations can be improved by applying art styles on various presentational levels.
We are currently investigating the use of various other styles not only within the context of presenting the style itself but also as general presentation techniques. Cubism, for example, seems to be attractive as a means to present information in a compact way. At the moment we only have design studies, where we use a space-constrained environment where a larger number of images can be displayed in an area for one image. We see the constructivist notion behind the cubist theory as a possibility for deconstructing images, then recombine key fractions of these images as a kind of 'search summary' and present this as the start point for complex presentation sequence. Impressionism might be of interest for stimulating moods, where ancient Egyptian, Greek and Roman as well as medieval art seem to stimulate the feel for easiness and completeness within a presentation. The latter display significance simply by means of size and quantity, whereas images after the Renaissance also use space in the form of perspective. In future research we endeavour to determine when to use which technique and when a mix is most appropriate.

6. ACKNOWLEDGMENTS

The authors wish to thank the Rijksmuseum in Amsterdam for insightful discussion and helpful comments on art styles and for the provided access to their database. Furthermore we are thankful for their permission to use the images from their collection for this paper.
Part of this research was funded under the Dutch research projects Dynamo and ToKeN2000.

7. REFERENCES

[1] ANDRÉ, E., MÜLLER, J., and RIST, T. WIP/PPP: Knowledge-Based Methods for Fully Automated Multimedia Authoring. In: Proceedings of EUROMEDIA'96, London, UK, 1996, pp. 95-102.

[2] ANDRÉ, E., MÜLLER, J., and RIST, T. Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems. In: Proc. of the Second International Conference on Intelligent User Interfaces (IUI 2000), New Orleans, LA USA, 2000, pages 1-8.

[3] ARNHEIM, R. Art and Visual Perception: A Psychology of the creative eye. The new version. London: Faber and Faber, 1974.

[4] ATHITSOS, V., SWAIN, M. J., and FRANKEL, C. Distinguishing photographs and graphics on the world wide web. In: Workshop on Content-Based Access of Image and Video Libraries, Puerto Rico, June 1997.

[5] AUFFRET, G., CARRIVE, J., CHEVET, O., DECHILLY, T., RONFORD, R., and BACHIMONT, B. Audiovisual-based Hypermedia Authoring: using structured representations for efficient access to AV documents. In Proceedings of the 10th ACM conference on Hypertext and Hypermedia, Darmstadt, Germany, February 21-25, 1999, pages 169-178. Edited by Klaus Tochterman, Jorg Westbomke, Uffe K. Will and John J. Leggett.

[6] BOLL, S., KLASS, W., and WANDEL, J. A Cross-Media Adaptation Strategy for Multimedia Presentations. In ACM Multimedia '99 Proceedings, pages 37-46, Orlando, Florida, October 30 - November 5, 1999. ACM, Addison Wesley Longman.

[7] BORDWELL, D. *Making Meaning - Inference and Rhetoric in the Interpretation of Cinema*. Cambridge, Massachusetts: Harvard University Press, 1989.

[8] BRUSILOVSKY, P., KOBSA, A., and VASSILEVA, J. editors. Adaptive Hypertext and Hypermedia. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[9] CLARK, J. and DEROSE, S. XML Path Language (XPath) Version 1.0. W3C Recommendations are available at http://www.w3.org/TR/, 16 November 1999.

[10] DEL BIMBO, A. Visual Information Retrieval. Morgan Kaufmann Ed, San Francisco, USA, 1999.

[11] DEROSE, S., MALER, E., and DANIEL, J. R. XML Pointer Language (XPointer) Version 1.0. W3C Candidate Recommendations are available at http://www.w3.org/TR, 8 January 2001.

[12] DEUTSCH, A., FERNANDEZ, M., FLORESCU, D., LEVY, A., and SUCIU, D. XML-QL: A Query Language for XML. W3C Notes are available at http://www.w3.org/TR, August, 19, 1998.

[13] EISENSTEIN, S. M. *Selected Works: Towards a Theory of Montage*., pp 11-57 and 296-399, London: BFI Publishing, 1991.

[14] ECO, U. A Theory of Semiotic. The Macmillan Press, 1977.

[15] FREDERIX, G., CAENEN, G., and PAUWELS, E. J. PARISS: Panoramic, Adaptive and Reconfigurable Interface for Similarity Search. In Proc. of ICIP 2000 International Conference on Image Processing. WA 07.04, pages 222-225, September 2000.

[16] GOMBRICH, E. H. The story of Art. Phaiadon Press Limited, London, 1999.

[17] GROSSO, W., ERIKSSON, H., FERGERSON, R., GENNARI, J., TU, S., and MUSEN, M. Knowledge Modeling at the Millennium (The Design and Evolution of Protege-2000). Technical Report SMI Report Number: SMI-1999-0801, Stanford Medical Informatics (SMI), 1999.

[18] ICONOCLASS. http://www.iconclass.nl/

[19] International Organization for Standardization/International Electrotechnical Commission. Information technology - Coding of moving pictures and audio, 1999. International Standard ISO/IEC 14496:1999 (MPEG-4).

[20] KAMPS, T. Diagram Design : A Constructive Theory. Springer Verlag, 1999.

[21] KNOTT, A. and DALE, R. Choosing a Set of Rhetorical Relations for Text Generation: A Data-Driven Approach. Trends in Natural Language Generation: an Artificial Intelligence Perspective, 1996.

[22] MANN, W. C., MATTHIESEN, C. M. I. M., and THOMPSON, S. A. Rhetorical Structure Theory and Text Analysis. Technical Report ISI/RR-89-242, Information Sciences Institute, University of Southern California, November 1989.

[23] MATHWORKS INC. The MathWorks Homepage. http://www.mathworks.com/, 2001.

[24] NACK, F. and LINDLEY, C. Production and maintenance environments for interactive audio-visual stories. In Proceedings ACM MM 2000 Workshops - Bridging the Gap: Bringing Together New Media Artists and Multimedia Technologists, pages 21- 24, Los Angeles, CA., October 31, 2000.

[25] PAUWELS, E. and FREDERIX, G. Image Segmentation by Nonparametric Clustering Based on the Kolmogorov-Smirnov Distance. In Proc. of ECCV 2000, 6th European Conference on Computer Vision, Dublin, pages 85{99, June 2000.

[26] QUINLAN, J. R. C4:5: programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[27] RUTLEDGE, L., BAILEY, B., VAN OSSENBRUGGEN J., HARDMAN, L., and GEURTS, J. Generating Presentation Constraints from Rhetorical Structure. In Proceedings of the 11th ACM conference on Hypertext and Hypermedia, pages 19- 28, San Antonio, Texas, USA, May 30 - June 3, 2000. ACM.

[28] RUTLEDGE, L., VAN OSSENBRUGGEN, J., HARDMAN, L., and BULTERMAN, D. C. A. Mix'n'Match: Exchangeable Modules of Hypermedia Style. In Proceedings of the 10th ACM conference on Hypertext and Hypermedia, Darmstadt, Germany, February 21-25, 1999, pages 179-188. Edited by Klaus Tochterman, Jorg Westbomke, Uffe K. Will and John J. Leggett.

[29] SANTINI, S. and RAMESH J. Integrated Browsing and Querying for Image Databases. IEEE MultiMedia, pages 26 - 39, July -September 2000.

[30] SCHMIDT A., WINDHOUWER M., and KERSTEN, M. L. Feature Grammars. In ISAS'99 The 5th. Int'l Conference on Information Systems Analysis and Synthesis, Orlando, Florida, 1999.

[31] SCHREIBER, A. T. G., DUBBELDAM, B., WIELEMAKER, J., & WIELINGA, B. Ontology-based Photo Annotation, IEEE Intelligent Systems, pp 66 – 74, May/June 2001 (Vol. 16, No. 3) http://www.computer.org/intelligent/ex2001/x3066abs.htm

[32] SMEULDERS, A. W. M., WORRING, M., SANTINI, S., GUPTA, A., and JAIN, R. Content-based image retrieval: the end of the early years, 1349 - 1380, 22 - 12, IEEE trans. PAMI, 2000.

[33] VAN HARMELEN, F., and HORROCKS, I. Reference description of the DAML+OIL ontology markup language. http://www.daml.org/2000/12/reference.html, December 2000. Contributors: Tim Berners-Lee, Dan Brickley, Dan Connolly, Mike Dean, Stefan Decker, Pat Hayes, Jeff Hefliin, Jim Hendler, Deb McGuinness, Lynn Andrea Stein.

[34] VAN OSSENBRUGGEN, J., GEURTS, J., CORNELISSEN, F., RUTLEDGE, L., and HARDMAN, L. Towards Second and Third Generation Web-Based Multimedia. In The Tenth International World Wide Web Conference [19], pages 479-488.

[35] W3C. Synchronized Multimedia Integration Language (SMIL 2.0) Specification. W3C Recommendations are available at http://www.w3.org/TR/, August 7, 2001. Edited by Aaron Cohen.

[36] WEITZMAN, L. and WITTENBURG, K. Automatic presentation of multimedia documents using relational grammars. In: Proceedings of the second ACM international conference on Multimedia '94 San Francisco pp. 443-451, October 15 - 20, 1994

[37] WINDHOUWER, M., SCHMIDT, A., and KERSTEN, M. L. Acoi: A system for Indexing Multimedia Objects. In International Workshop on Information Integration and Web-based Applications & Services, Yogyakarta, Indonesia, November 1999.