Semantic Multimedia

Steffen Staab[1], Ansgar Scherp[1], Richard Arndt[1], Raphael Troncy[2], Marcin Grzegorzek[1], Carsten Saathoff[1], Simon Schenk[1], and Lynda Hardman[2]

[1] ISWeb Research Group, University of Koblenz-Landau
http://isweb.uni-koblenz.de
[2] Semantic Media Interfaces, CWI Amsterdam
http://www.cwi.nl

Abstract. Multimedia constitutes an interesting field of application for Semantic Web and Semantic Web reasoning, as the access and management of multimedia content and context depends strongly on the semantic descriptions of both. At the same time, multimedia resources constitute complex objects, the descriptions of which are involved and require the foundation on sound modeling practice in order to represent findings of low- and high level multimedia analysis and to make them accessible via Semantic Web querying of resources. This tutorial aims to provide a red thread through these different issues and to give an outline of where Semantic Web modeling and reasoning needs to further contribute to the area of semantic multimedia for the fruitful interaction between these two fields of computer science.

1 Semantics for Multimedia

Multimedia objects are ubiquitous, whether found via web search (e.g., Google[1] or Yahoo![2] images), or via dedicated sites (e.g., Flickr[3] or YouTube[4]) or in the repositories of private users or commercial organizations (film archives, broadcasters, photo agencies, etc.). The media objects are produced and consumed by professionals and amateurs alike. Unlike textual assets, whose content can be searched for using text strings, media search is dependent on, *(i)*, complex analysis processes, *(ii)*, manual descriptions of multimedia resources, *(iii)*, representation of these results and contributions in a widely understandable format for, *(iv)* later retrieval and/or querying by the consumer of this data.

In the past, this process has not been supported by an interoperable and easily extensible machinery of processing tools, applications and data formats, but only by idiosyncratic combinations of system components into sealed off applications such that effective sharing of their semantic metadata remained impossible and the linkage to semantic data and ontologies found on the Semantic Web remained far off.

[1] http://images.google.com/

[2] http://images.search.yahoo.com/

[3] http://www.flickr.com/

[4] http://www.youtube.com/

C. Baroglio et al. (Eds.): Reasoning Web 2008, LNCS 5224, pp. 125–170, 2008.

MPEG-7 [52, 57] is an international standard defined by the Moving Picture Experts Group (MPEG) that specifies how to connect descriptions to parts of a media asset. The standard includes descriptors representing low-level media-specific features that can often be automatically extracted from media types. Unfortunately, MPEG-7 is not fully suitable for describing multimedia content, because *i)* it is not open to standards that represent knowledge and make use of existing controlled vocabularies for describing the subject matter and *(ii)* its XML Schema[5] based nature has led to design decisions that leave the annotations conceptually ambiguous and therefore prevent direct machine processing of semantic content descriptions.

In order to avoid such problems, we advocate the use of Semantic Web languages and a core ontology for multimedia annotations throughout the manual and automatic processing of multimedia content and its retrieval. For this purpose, we build on rich ontological foundations provided by an ontology such as the Descriptive Ontology for Linguistic and Cognitive Engineering[6] (DOLCE) and sound ontology engineering principles. The result presented in this tutorial is COMM, a core ontology for multimedia, which is able to accommodate results from manual annotation of data (cf. Section 6) as well as from automated processing (cf. Section 4).

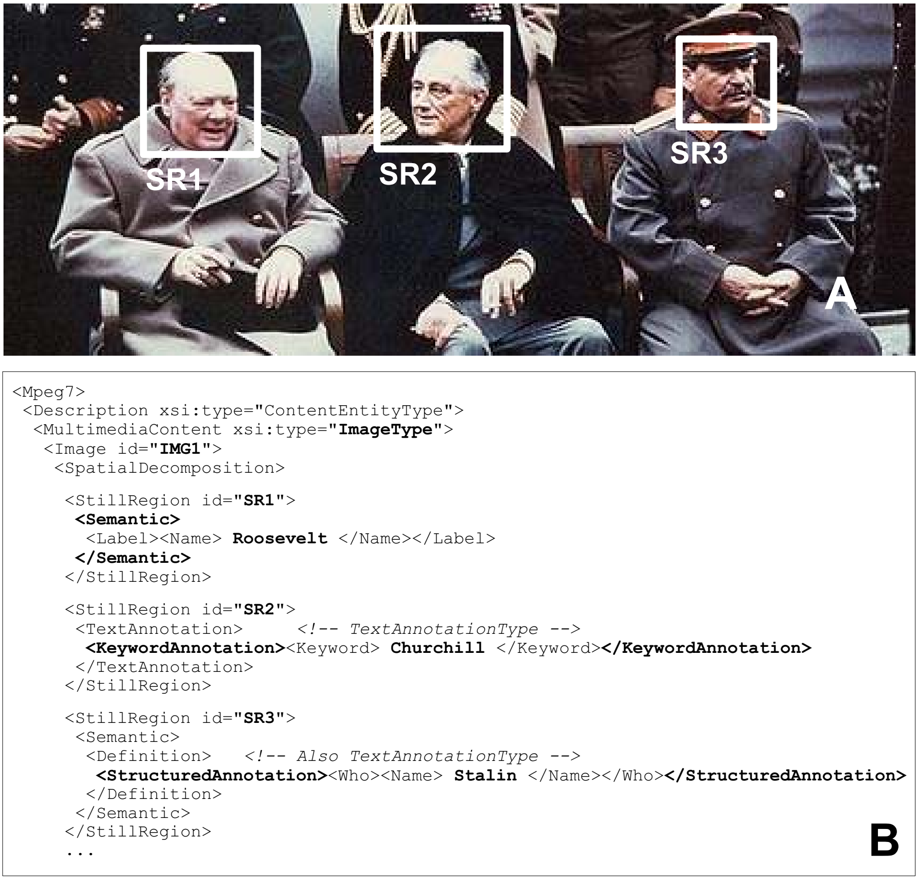
The remainder of this document is organized as follows: In the next Section 2, we illustrate by an example scenario the main problems when using MPEG-7 for describing multimedia resources. Subsequently, we define in Section 3 the requirements that a multimedia ontology should meet. We review work in image and video processing in Section 4, before we present COMM, an MPEG-7 based ontology, in Section 5 and discuss our design decisions based on our requirements. In Section 6, we illustrate how to use COMM in a manual annotation tool. In Section 7, we demonstrate the use of the ontology with the scenario from Section 2 and in Section 8 we indicate challenges and solutions for querying metadata based on COMM. Further and future issues of semantic multimedia are considered in Section 9, before we summarize and conclude the paper.

2 Annotating Multimedia Assets

For annotating multimedia assets, let us imagine Nathalie, a student in history, who wants to create a multimedia presentation of the major international conferences and summits held in the last 60 years. Her starting point is the famous "Big Three" picture, taken at the Yalta (Crimea) Conference, showing the heads of government of the United States, the United Kingdom, and the Soviet Union during World War II. Nathalie uses an MPEG-7 compliant authoring tool for detecting and labeling relevant multimedia objects automatically. On the Internet, she finds three different face recognition web services that provide very good results for detecting Winston Churchill, Franklin D. Roosevelt, and Josef Stalin, respectively. Having these tools, she would like to run the face recognition web

[5] `http://www.w3.org/XML/Schema`

[6] `http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf`



```
<Mpeg7>
 <Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="ImageType">
   <Image id="IMG1">
    <SpatialDecomposition>

     <StillRegion id="SR1">
      <Semantic>
       <Label><Name> Roosevelt </Name></Label>
      </Semantic>
     </StillRegion>

     <StillRegion id="SR2">
      <TextAnnotation>     <!-- TextAnnotationType -->
       <KeywordAnnotation><Keyword> Churchill </Keyword></KeywordAnnotation>
      </TextAnnotation>
     </StillRegion>

     <StillRegion id="SR3">
      <Semantic>
       <Definition>   <!-- Also TextAnnotationType -->
        <StructuredAnnotation><Who><Name> Stalin </Name></Who></StructuredAnnotation>
       </Definition>
      </Semantic>
     </StillRegion>
     ...
```

Fig. 1. MPEG-7 annotation example of an image adapted from Wikipedia, `http://en.wikipedia.org/wiki/Yalta_Conference`

services on images and import the extraction results into the authoring tool in order to automatically generate links from the detected face regions to detailed textual information about Churchill, Roosevelt, and Stalin (image in Fig. 1-A).

Nathalie would then like to describe a recent video from a G8 summit, such as the retrospective *A history of G8 violence* made by Reuters[7]. She uses again an MPEG-7 compliant segmentation tool for detecting the seven main sequences of this 2'26 minutes report: the various anti-capitalist protests during the Seattle (1999), Melbourne (2000), Prague (2000), Gothenburg (2001), Genoa (2001), St Petersburg (2006), Heiligendamm (2007) World Economic Forums, EU and G8 Summits. Finally, Nathalie plans to deliver her multimedia presentation in an Open Document Format (ODF) document embedding the image and video previously annotated. However, this scenario causes several problems with

[7] `http://www.reuters.com/news/video/summitVideo?videoId=56114`

existing solutions. These problems refer to fragment identification, semantic annotation, web interoperability, and embedding semantic annotations into compound documents.

Fragment identification. Particular regions of the image need to be localized (anchor value in [29]). However, the current web architecture does not provide a means for uniquely identifying sub-parts of media assets, in the same way that the fragment identifier in the URI can refer to a part of an HTML or XML document. Indeed, for almost any other media type such as audio, video, and image, the semantics of the fragment identifier has not been defined or is not commonly accepted. Providing an agreed upon way to localize sub-parts of multimedia objects (e.g., sub-regions of images, temporal sequences of videos, or tracking moving objects in space and in time) is fundamental[8] [25]. For images, one can use either MPEG-7 or SVG snippet code to define the bounding box coordinates of specific regions. For temporal locations, one can use MPEG-7 code or the TemporalURI RFC[9]. MPEG-21 specifies a normative syntax to be used in URIs for addressing parts of any resource but whose media type is restricted to MPEG [51]. The MPEG-7 approach requires an indirection: an annotation is *about* a fragment of an XML document that *refers* to a multimedia document, whereas the MPEG-21 approach does not have this limitation [90].

Semantic annotation. MPEG-7 is a natural candidate for representing the extraction results of multimedia analysis software such as a face recognition web service. The language, standardized in 2001, specifies a rich vocabulary of multimedia descriptors, which can be represented in either XML or a binary format. While it is possible to specify very detailed annotations using these descriptors, it is not possible to guarantee that MPEG-7 metadata generated by different agents will be mutually understood due to the lack of formal semantics of this language [32, 87]. The XML code of Fig. 1-B illustrates the inherent interoperability problems of MPEG-7: several descriptors, semantically equivalent and representing the same information while using different syntax can coexist [88]. As Nathalie used three different face recognition web services, the extraction results of the regions `SR1`, `SR2`, and `SR3` differ from each other even though they are all syntactically correct. While the first service uses the MPEG-7 `SemanticType` for assigning the `<Label>` *Roosevelt* to still region `SR1`, the second one makes use of a `<KeywordAnnotation>` for attaching the keyword *Churchill* to still region `SR2`. Finally the third service uses a `<StructuredAnnotation>` (which can be used within the `SemanticType`) in order to label still region `SR3` with *Stalin*. Consequently, alternative ways for annotating the still regions render almost impossible the retrieval of the face recognition results within the authoring tool since the corresponding XPath[10] query has to deal with these syntactic variations. As a result, the authoring tool will not link occurrences of Churchill in

[8] See also the forthcoming W3C Media Fragments Working Group: `http://www.w3.org/2008/01/media-fragments-wg.html`

[9] `http://www.annodex.net/TR/URI_fragments.html`

[10] `http://www.w3.org/TR/xpath20/`

the images with, e.g., his biography as it does not expect semantic labels of still regions as part of the `<KeywordAnnotation>` element.

Web interoperability. Nathalie would like to link the multimedia presentation to historical information about the key figures of the Yalta Conference or the various G8 summits that is already available. She has also found semantic metadata about the relationships between these figures that could improve the automatic generation of the multimedia presentation. However, she realizes that MPEG-7 cannot be combined with these concepts defined in domain-specific ontologies because of its closing to the web. As this example demonstrates, although MPEG-7 provides ways of associating semantics with (parts of) non-textual media assets, it is incompatible with (semantic) web technologies and has no formal description of the semantics encapsulated implicitly in the standard.

Embedding into compound documents. Nathalie needs to compile the semantic annotations of the images, videos, and textual stories into a semantically annotated compound document. However, the current state of the art does not provide a framework which allows the semantic annotation of compound documents. MPEG-7 solves only partially the problem as it is restricted to the description of audiovisual compound documents. Bearing the growing number of multimedia office documents in mind, this limitation is a serious drawback.

Querying. Eventually, Nathalie and other consumers of Nathalie's compound document may want to pick out specific events, related to specific persons or locations. Depending on such a condition and depending on what they want to pick out, e.g., a 2 minute video stream or a key frame out of a video, they need to formulate a query and receive the corresponding results. The query language and corresponding engine receiving such a request must be able to drill down into the compound document at an arbitrary level of granularity. For instance, if a person like Churchill appears in a keyframe that is part of a video scene that is part of a video shot, Churchill will also appear in the video shot as a whole. The engine must return results also at the desired level of granularity, e.g., the video scene.

3 Requirements for Designing a Multimedia Ontology

Requirements for designing a multimedia ontology have been gathered and reported in the literature, e.g., in [35]. Here, we compile these and use our scenario from the previous section to present a list of requirements for a web-compliant multimedia ontology.

MPEG-7 compliance. As an international standard, MPEG-7 is used both in the signal processing and the broadcasting communities. It contains a wealth of accumulated experience that needs to be included in a web-based multimedia ontology. In addition, existing annotations in MPEG-7 should be easily expressible in this multimedia ontology.

Semantic interoperability. Annotations are only re-usable when the captured semantics can be shared among multiple systems and applications. Obtaining similar results from reasoning processes about terms in different environments can only be guaranteed if the semantics is sufficiently explicitly described. A multimedia ontology has to ensure that the intended meaning of the captured semantics can be shared among different systems.

Syntactic interoperability. Systems are only able to share the semantics of annotations if there is a means of conveying this in some agreed-upon syntax. Given that the (semantic) web is an important repository of both media assets and annotations, a semantic description of the multimedia ontology should be expressible in a web language such as OWL, RDF/XML, or RDFa[11].

Separation of concerns. Clear separation of subject matter (i.e., knowledge about depicted entities, such as the person Winston Churchill) from knowledge that is related to the administrative management or the structure and the features of multimedia documents (e.g., Churchill's face is to the left of Roosevelt's face) is required. Reusability of multimedia annotations can only be achieved if the connection between both ontologies is clearly specified by the multimedia ontology.

Modularity. As demonstrated by MPEG-7, a complete multimedia ontology can be very large. The design of a multimedia ontology should thus be made modular, to minimize the execution overhead when used for multimedia annotation. Modularity is also a good engineering principle.

Extensibility. While we intend to construct a comprehensive multimedia ontology, as ontology development methodologies demonstrate, this can never be complete. New concepts will always need to be added to the ontology. This requires a design that can always be extended, without changing the underlying model and assumptions and without affecting legacy annotations.

4 Low Level Multimedia Processing and Classification

In this section, chosen low-level methods (in the sense of signal processing) for describing and classifying multimedia assets are reviewed. Section 4.1 presents briefly some multimedia description techniques with the focus on visual information, while in Section 4.2 few algorithms for automatic classification of multimedia assets are discussed.

4.1 Multimedia Content Description

Multimedia assets can be represented by features in order to reduce and simplify the amount of resources required to describe a large set of data accurately.

[11] RDFa allows for representing structured information in XHTML documents such as calendar items, business contact information, licenses of the document, or creator and camera settings of images. It is available from `http://www.w3.org/2006/07/SWD/RDFa/primer/`.

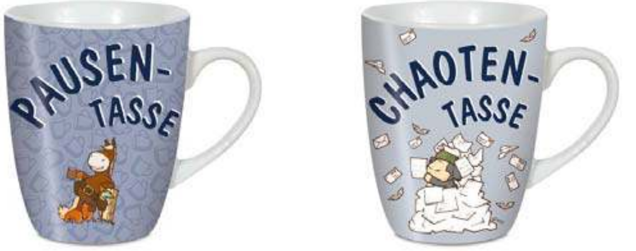


Fig. 2. Example of two objects with identical shape and different texture

According to current state of the art, for analysis with a large number of variables a large amount of memory and computation power is needed. For this reason, feature computation is a very important and unavoidable step in the multimedia processing chain.

Considering visual media assets, the feature computation techniques can be divided into two categories, namely the shape-based and the texture-based. Shape-based methods make use of geometric features such as lines or corners extracted by segmentation operations. These features and their relationships are then used for visual content description [7, 31, 39, 44]. However, the segmentation-based approach often suffers from errors due to loss of image details or other inaccuracies resulting from the segmentation process. Texture-based approaches avoid these disadvantages by directly using the visual data on the pixel level without a previous segmentation step [53, 70, 72]. Depending on the problem definition, both approaches have their advantages and disadvantages. For instance, objects depicted in Figure 2 can only be distinguished by texture features.

On the other hand, shape features of only one of the cups already describe fully the whole general class "cup". Concluding, shape-based description of multimedia contents seems to be more useful for classification into general categories, while texture-based features allow to distinguish visual contents belonging to the same general category from each other.

In the last decades many different algorithms for feature extraction from multimedia content have been proposed. Therefore, the MPEG-7 standard has been introduced to describe multimedia assets. Among many other things, the MPEG-7 standard defines visual descriptions for elementary features, such as color, texture, shape, and motion. Subsequently, we briefly present these descriptors.

MPEG-7 Color Descriptors. Color is the most basic attribute of visual media assets. MPEG-7 Visual defines five different description methods, each of which represents a different aspect of the color attribute. Color distribution includes a representative color description (Dominant Color), basic color distribution description (Scalable Color) and an advanced color distribution description (Color Structure). The remaining extraction techniques include Color Layout describing spatial distribution of colors, and Color Temperature describing perceptual feeling of illumination color.

Dominant Color. The Dominant Color descriptor characterizes an image or region by a small number of representative colors. These are selected by quantizing pixel colors into (up to seven) principal clusters. The description then consists of the fraction of the image represented by each color cluster and the variance of each one. A measure of overall spatial coherency of the clusters is also defined. This descriptor is a very compact description of the color distribution in the image.

Scalable Color. The Scalable Color descriptor is a color Histogram in the HSV Color Space [65], which is encoded by a Haar transform [65]. It has a binary representation that is scalable, in terms of bin numbers and bit representation accuracy, over a broad range of granularity. Retrieval accuracy can therefore be balanced against descriptor size. Inversion of the Haar transform [65] is not necessary for consumption of the description, since similarity matching is also effective in the transform domain.

Color Layout. The Color Layout descriptor represents the spatial layout of color images in a very compact form. It is based on generating a tiny (8 × 8) thumbnail of an image, which is encoded via Discrete Cosinus Transformation (DCT) and quantized. As well as efficient visual matching, this also offers a quick way to visualize the appearance of an image.

Color Structure. The Color Structure descriptor captures both color content and information about the spatial arrangement of the colors. Specifically, it is a histogram that counts the number of times a color is present in an 8 × 8 windowed neighborhood, as this window progresses over the image rows and columns. This enables it to distinguish, e.g., between an image in which pixels of each color are distributed uniformly and an image in which the same colors occur in the same proportions but are located in distinct blocks.

MPEG-7 Texture Descriptors

Edge Histogram. The Edge Histogram descriptor represents the spatial distribution of five types of edges (four directional edges and one non-directional). It consists of local histograms of these edge directions, which may optionally be aggregated into global or semi-global histograms.

Homogeneous Texture. The Homogeneous Texture descriptor is designed to characterize the properties of texture in an image (or region), based on the assumption that the texture is homogeneous, i.e., the visual properties of the texture are relatively constant over the region. It consists of the mean, the standard deviation value of an image, energy, and energy deviation values of Fourier transform [65] of the image.

Texture Browsing. The Texture Browsing descriptor is useful for representing homogeneous texture for browsing type applications, and requires only 12 bits (maximum). It provides a perceptual characterization of texture, similar to a human characterization, in terms of regularity, coarseness and directionality. The computation of this descriptor proceeds similarly as the Homogeneous Texture descriptor. First, the image is filtered with a bank of orientation and scale tuned filters (modeled using Gabor functions) [97]; from the filtered outputs, two dominant texture orientations are identified. Three bits are used to represent each of the dominant orientations. This is followed by analyzing the filtered image projections along the dominant orientations to determine the regularity (quantified to 2 bits) and coarseness (2 bits × 2). The second dominant orientation and second scale feature are optional. This descriptor, combined with the Homogeneous Texture descriptor, provide a scalable solution to representing homogeneous texture regions in images.

MPEG-7 Shape Descriptors

Region Shape. The shape of an object may consist of either a single region or a set of regions as well as some holes in the object. Since the Region Shape descriptor makes use of all pixels constituting the shape within a frame, it can describe any shapes, i.e. not only a simple shape with a single connected region but also a complex shape that consists of holes in the object or several disjoint regions. The Region Shape descriptor not only can describe such diverse shapes efficiently in a single descriptor, but is also robust to minor deformation along the boundary of the object.

Contour Shape. The Contour Shape descriptor captures characteristic shape features of an object or region based on its contour. It uses so-called Curvature Scale Space representation [50], which captures perceptually meaningful features of the shape. The Contour Shape descriptor has a number of important properties, namely: (i) it captures very well characteristic features of the shape, enabling similarity-based retrieval; (ii) it reflects properties of the perception of human visual system and offers good generalization; (iii) it is robust to non-rigid motion; (iv) it is robust to partial occlusion of the shape; (v) it is robust to perspective transformations which result from the changes of the camera parameters and are common in images and video; (vi) it is compact.

MPEG-7 Motion Descriptors

Camera Motion. This descriptor characterizes 3D camera motion parameters. It is based on 3D camera motion parameter information, which can be automatically extracted or generated by capture devices. The camera motion descriptor supports the following well-known basic camera operations: fixed, panning, tracking, tilting, booming, zooming, dollying, and rolling.

Motion Trajectory. The motion trajectory of an object is a simple feature defined as the localization in time and space of one representative point of this object. This descriptor is useful for content-based retrieval in object-oriented visual databases.

Parametric Motion. The parametric motion is associated with arbitrary (foreground or background) objects, defined as regions (group of pixels) in the image over a specified time interval. Such an approach leads to a very efficient description of several types of motions, including simple translation, rotation and zoom, or more complex motions such as combinations of the above-mentioned elementary motions.

Motion Activity. The Motion Activity descriptor captures the intuitive notion of "intensity of action" or "pace of action" in a video segment. This descriptor is useful for applications such as video re-purposing, surveillance, fast browsing, dynamic video summarization, content-based querying, and others.

4.2 Multimedia Content Classification

In the previous section, we introduced how media assets can be described by feature vectors, sometimes referred to as histograms. In this section, we present how these assets can be classified using automatic computer-aided approaches. In order to classify multimedia assets into concepts (classes), computers need to model sample data of these concepts. This process is called training. In the training phase, annotated and representative training data for all concepts (e.g., images for visual concepts, or music samples for audio concepts) is required. Once the concepts have been modeled in the training phase, unknown and not annotated multimedia assets can be assigned to the trained concepts by classification algorithms (classifiers). Considering visual media assets, the most known classification techniques are: Template Matching [5, 26, 71], Artificial Neural Networks [60, 64, 86, 97, 99], Support Vector Machines (SVM) [9, 96], and the Eigenspace Approach [27, 48, 49, 94]. Today, the SVM algorithm is widely applied to classify multimedia content. Thus, it is elaborated in more detail in the following using the example of object classification in images.

Support Vector Machines have been proposed as a very effective method for general purpose pattern recognition [9, 96]. Intuitively, given a set of points which belong to either of two classes, a SVM finds the hyperplane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyperplane. In the sense of object classification in digital images, a simple two-class problem has to be solved for all objects $\Omega_{\kappa=1,\ldots,N_\Omega}$ considered in the task. The first class is the object class Ω_κ itself. The second class represents everything which is not the class Ω_κ. It can be denoted by Ω'_κ. For the training of the class Ω_κ images of this object Ω_κ from different viewpoints are taken into account, while for the learning of the anti-class Ω'_κ images of all other objects $\Omega_{i\neq\kappa}$ are used. In the recognition phase, the SVM decides which of the objects $\Omega_{\kappa=1,\ldots,N_\Omega}$ occurs in a test scene. The two-class problem is regarded for each object class Ω_κ, i.e., N_Ω times. It is expected that $N_\Omega - 1$ times the anti-class $\Omega'_{\widehat{\kappa}}$ wins the two-class problem. The actual classification result $\Omega_{\widehat{\kappa}}$ is supposed to win the two-class problem only once[12].

[12] Assuming that exactly one of the trained objects $\Omega_{\kappa=1,\ldots,N_\Omega}$ occurs in the scene.

In the following, we present a simple example where the object class Ω_κ and its anti-class Ω'_κ are linearly separable. Let the feature vectors $\boldsymbol{c}_{j=1,\dots,N_S}$ representing all object classes $\Omega_{\kappa=1,\dots,K}$ build a set S, where

$$S = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_j, \dots, \boldsymbol{c}_{N_S}\} \quad . \tag{1}$$

Each feature vector $\boldsymbol{c}_j$ from S belongs either to the class Ω_κ or to the anti-class Ω'_κ, which is given with the corresponding labels $y_j = \{-1, 1\}$. The goal is to establish the equation of a hyperplane that divides the set S leaving all the feature vectors describing Ω_κ on its one side and all the feature vectors belonging to Ω'_κ on the other side of the hyperplane. Moreover, both the distance of the class Ω_κ and the anti-class Ω'_κ to the hyperplane has to be maximized. For this purpose, some preliminary definitions are needed.

Definition 1. *The set S is linearly separable if there exist a vector $\boldsymbol{v} \in \mathbb{R}^{N_c}$ and scalar $b \in \mathbb{R}$ such that*

$$y_j(\boldsymbol{v} \cdot \boldsymbol{c}_j + b) \geq 1 \tag{2}$$

for all $j = 1, 2, \dots, N_S$. Note that $\boldsymbol{c}_j \in \mathbb{R}^{N_c}$.

The pair $(\boldsymbol{v}, b)$ defines a hyperplane of equation

$$\boldsymbol{v} \cdot \boldsymbol{c} + b = 0 \quad , \tag{3}$$

named *separating hyperplane.* If with $|\boldsymbol{v}|$ the norm of the vector $\boldsymbol{v}$ is denoted, the distance d_j of a point $\boldsymbol{c}_j$ (feature vector) to the separating hyperplane $(\boldsymbol{v}, b)$ is given by

$$d_j = \frac{\boldsymbol{v} \cdot \boldsymbol{c}_j + b}{|\boldsymbol{v}|} \quad . \tag{4}$$

Combining inequality (2) and equation (4) for all $\boldsymbol{c}_j \in S$ we have

$$y_j d_j \geq \frac{1}{|\boldsymbol{v}|} \quad . \tag{5}$$

Therefore, $|\boldsymbol{v}|^{-1}$ is the lower bound on the distance between the feature vectors $\boldsymbol{c}_j$ and the separating hyperplane $(\boldsymbol{v}, b)$. A *canonical representation* of the separating hyperplane is obtained by rescaling the pair $(\boldsymbol{v}, b)$ into the pair $(\boldsymbol{v}', b')$ in such a way that the distance of the closest feature vector equals $|\boldsymbol{v}'|^{-1}$. For the canonical representation $(\boldsymbol{v}', b')$ of the hyperplane it can be written considering the equation (2) that

$$\min_{\boldsymbol{c}_j \in S}\{y_j(\boldsymbol{v}' \cdot \boldsymbol{c}_j + b')\} = 1 \quad . \tag{6}$$

Consequently, for a separating hyperplane in the canonical representation, the bound in inequality (5) is tight. The discussion comes to the point where the *optimal separating hyperplane* has to be defined.

Definition 2. *Given a linearly separable set S, the optimal separating hyperplane is the separating hyperplane, for which the distance to the closest point of S is maximum.*

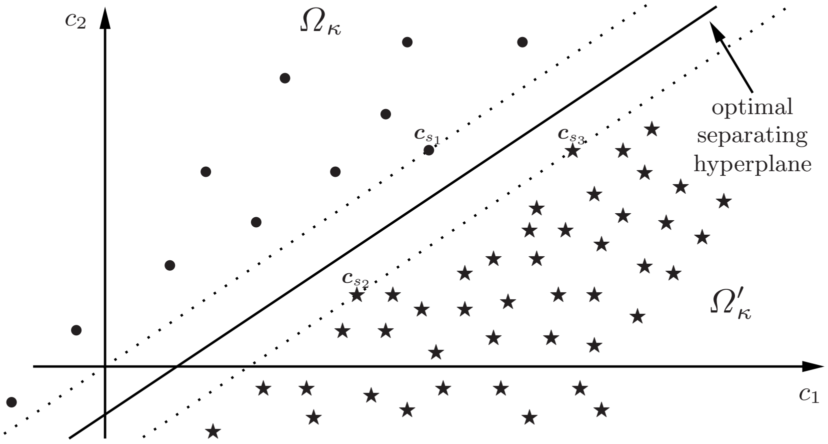


Fig. 3. Optimal separating hyperplane for two-dimensional feature space. With • feature vectors of the object class Ω_κ are denoted. By ★ the remaining feature vectors of all other object classes are represented. Three feature vectors $\boldsymbol{c}_{s_1}$, $\boldsymbol{c}_{s_2}$, and $\boldsymbol{c}_{s_3}$ lie in the minimum distance to the optimal separating hyperplane and are called support vectors.

Such an optimal separating hyperplane for a two-dimensional feature space, i. e., for $\boldsymbol{c} = (c_1, c_2)^{\mathrm{T}}$, is depicted in Figure 3. In this case, it is just a straight line. The feature vectors $\boldsymbol{c}_{s_1}$, $\boldsymbol{c}_{s_2}$, and $\boldsymbol{c}_{s_3}$, which are closest to the optimal separating hyperplane, are called *support vectors*. For object modeling (i.e., the training phase) it is sufficient to store the support vectors for each class Ω_κ, which significantly reduces the data amount. In the recognition phase, the classification algorithm starts with the extraction of feature vectors from a scene. Subsequently, it determines for each object class on which side of the optimal separating hyperplane the corresponding feature vectors lie. In this way, the objects which occur in the scene are found. A detailed discussion of object classification methods using the SVM approach can be found in [8].

So far, we considered the classification of single objects. Here, it is generally assumed that the probability of appearance of objects in the scene is equal (all objects have the same a priori probability). For example, if we consider ten objects for classification, we assume an a priori probability of 10 percent for all objects. If there is contextual information about the scene available, one can leverage this information to improve the classification results. For example, Grzegorzek and Izquierdo [28] are showing object classification at the example of three different contextual scenes: kitchen, office, and nursery. However, we can also imagine a scenario where we have to classify multiple objects in a scene. For example, in a tennis match we may detect a tennis player. In addition, we may detect another object being either a lemon or a tennis ball due to similar shape and texture. Taking contextual information into account and knowing that the probability of appearance for the two relations "player and ball" is higher than "player and lemon", we can rank the first classification higher and improve the overall classification quality. In another example, we may analyze a picture and identify a blue part at the top as sea or sky. Another part in the middle is also classified as sea or sky. Here, we can take contextual information about the spatial distribution into account saying that sky is typically above sea. This example of taking contextual information into account for object classification is presented in [74]. It pursues a knowledge-based approach for reasoning using the degree of classification confidence for the single objects as input to achieve overall annotation of the picture.

5 A Formal Ontological Foundation for Multimedia

As introduced in Section 1, MPEG-7 can be used to specify the connection between semantic annotations and parts of media assets. In Section 4, we presented concrete examples of different kinds of semantic annotations supported by MPEG-7. Here, we are aiming at defining a formal core ontology for multimedia called COMM (Core Ontology of MultiMedia). Based on early work [37, 87], COMM has been designed manually by re-engineering completely MPEG-7 according to the intended semantics of the written standard. We satisfy our semantic interoperability not by aligning our ontology to the XML Schema definition of MPEG-7 but by providing a formal semantics for MPEG-7. The foundational ontology DOLCE serves as the basis of COMM. More precisely, the Description and Situation (D&S) and Ontology of Information Objects (OIO) patterns are extended into various multimedia patterns that formalize the MPEG-7 concepts. For designing COMM, we employ a methodology by Sure et al. [85] that bases on a foundational, or top level, ontology. This provides a domain independent vocabulary that explicitly includes formal definitions of foundational categories, such as processes or physical objects, and eases the linkage of domain-specific ontologies because of the shared definitions of top level concepts.

COMM covers the most important part of MPEG-7 that is commonly used for describing the structure and the content of multimedia documents. Current investigations show that parts of MPEG-7 that have not yet been considered (e.g., navigation & access) can be formalized analogously to the other descriptors through the definition of other multimedia patterns.

COMM is an OWL DL ontology that can be viewed using Protégé. Its consistency has been validated using Fact++-v1.1.5. Other reasoners failed to classify it due to the enormous amount of DL axioms that are present in DOLCE. The presented OWL DL version of the core module is just an approximation of the intended semantics of COMM since the use of OWL 1.1 (e.g., qualified cardinality restrictions for number restrictions of MPEG-7 low-level descriptors) and even more expressive logic formalisms are required for capturing its complete semantics[13].

Firstly, we briefly introduce our chosen foundational ontology in Section 5.1. The multimedia ontology COMM is presented in Sections 5.2 and 5.3. Subsequently,

[13] The reification schema of DOLCE D&S is even not completely expressible in OWL 1.1.

we discuss why our ontology satisfies all the requirements stated in Section 5.4. Finally, we discuss related work in Section 5.5 and provide a comparison to COMM. Please note that the interested reader may also download the COMM ontology and its documentation from `http://multimedia.semanticweb.org/COMM/`.

5.1 DOLCE as Modeling Basis

Using the review in [61], we select the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (cf. [18]) as a modeling basis. Our choice is influenced by two of the main design patterns: *Descriptions & Situations* (D&S) and *Ontology of Information Objects* (OIO) [17]. The former can be used to formalize contextual knowledge, while the latter, based on D&S, implements a semiotics model of communication theory. We consider that the annotation process is a *situation* (i.e., a reified context) that needs to be described.

5.2 Multimedia Patterns

The patterns for D&S and OIO need to be extended for representing MPEG-7 concepts since they are not sufficiently specialized to the domain of multimedia annotation. This section introduces these extended multimedia design patterns, while Section 5.3 details two central concepts underlying these patterns: digital data and algorithms (cf. [61]). In order to define design patterns, one has to identify repetitive structures and describe them at an abstract level. The two most important functionalities provided by MPEG-7 are: the *decomposition* of media assets and the (semantic) *annotation* of their parts, which we include in our multimedia ontology.

Decomposition. MPEG-7 provides descriptors for spatial, temporal, spatio-temporal and media source decompositions of multimedia content into segments. A segment is the most general abstract concept in MPEG-7 and can refer to a region of an image, a piece of text, a temporal scene of a video or even to a moving object tracked during a period of time.

Annotation. MPEG-7 defines a very large collection of descriptors that can be used to annotate a segment. These descriptors can be low-level visual features, audio features or more abstract concepts. They allow the annotation of the content of multimedia documents or the media asset itself.

In the following, we first introduce the notion of multimedia data and then present the patterns that formalize the decomposition of multimedia content into segments, or allow the annotation of these segments. The decomposition pattern handles the structure of a multimedia document, while the media annotation pattern, the content annotation pattern, and the semantic annotation pattern are useful for annotating the media, the features, and the semantic content of the multimedia document respectively.

Multimedia Data. This encapsulates the MPEG-7 notion of multimedia content and is a subconcept of digital-data[14] (introduced in more detail in Section 5.3). multimedia-data is an abstract concept that has to be further specialized for concrete multimedia content types (e.g., image-data corresponds to the pixel matrix of an image). According to the OIO pattern, multimedia-data is realized by some physical media (e.g., an image). This concept is needed for annotating the physical realization of multimedia content.

Decomposition Pattern. Following the D&S pattern, we consider that a decomposition of a multimedia-data entity is a situation (a segment-decomposition) that satisfies a description such as a segmentation-algorithm or a method (e.g., a user drawing a bounding box around a depicted face), which has been applied to perform the decomposition, see Fig. 4-B. Of particular importance are the roles that are defined by a segmentation-algorithm or a method. The output-segment-roles express that some multimedia-data entities are segments of a multimedia-data entity that plays the role of an input segment (input-segment-role). These data entities have as setting a segment-decomposition situation that satisfies the roles of the applied segmentation-algorithm or method. The output-segment-roles as well as segment-decompositions are then specialized according to the segment and decomposition hierarchies of MPEG-7 ([52], part 5, section 11). In terms of MPEG-7, unsegmented (complete) multimedia content also corresponds to a segment. Consequently, annotations of complete multimedia content start with a root segment. In order to designate multimedia-data instances that correspond to these root segments the decomposition pattern provides the root-segment-role concept. Note that root-segment-roles are not defined by methods which describe segment-decompositions. They are rather defined by methods which cause the production of multimedia content. These methods as well as annotation modes which allow the description of the production process (e.g., [52], part 5, section 9) are currently not covered by our ontology. Nevertheless, the prerequisite for enhancing COMM into this direction is already given.

The decomposition pattern also reflects the need for localizing segments within the input segment of a decomposition as each output-segment-role requires a mask-role. Such a role has to be played by one or more digital-data entities which express one localization-descriptor. An example of such a descriptor is an ontological representation of the MPEG-7 `RegionLocatorType`[15] for localizing regions in an image (see Fig. 4-C). Hence, the mask-role concept corresponds to the notion of a mask in MPEG-7.

The specialization of the pattern for describing image decompositions is shown in Fig. 5-F. According to MPEG-7, an image or an image segment (image-data) can be composed into still regions. Following this modeling, the concepts output-segment-role and root-segment-role are specialized by the concepts still-region-role and root-still-region-role respectively. Note, that root-still-region-role is a subconcept of still-region-role *and* root-segment-role. The MPEG-7 decomposition mode which can be applied to

[14] Sans serif font indicates ontology concepts.

[15] Italic type writer font indicates MPEG-7 language descriptors.

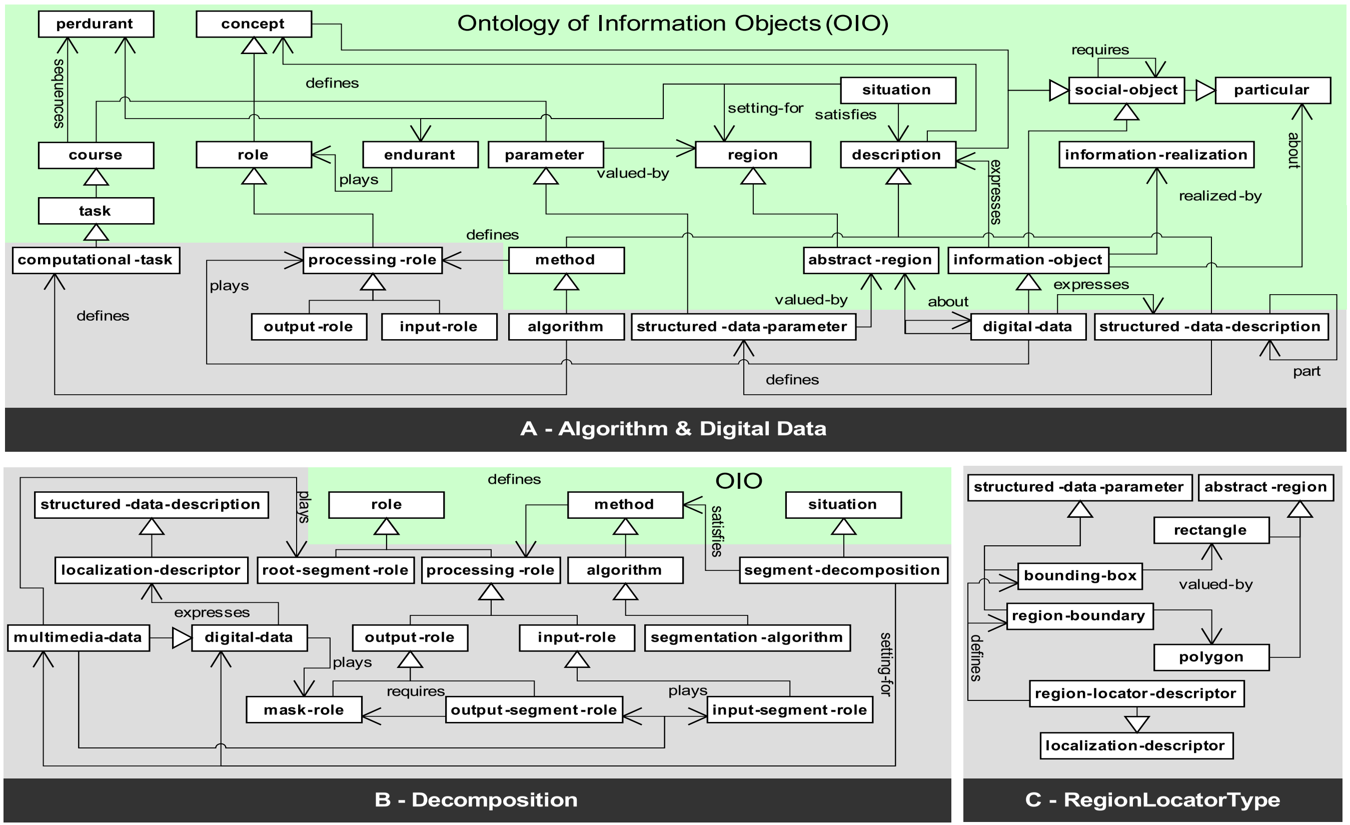


Fig. 4. COMM: Design patterns in UML notation: Basic design patterns (A), multimedia patterns Decomposition (B) and modeling example (C)

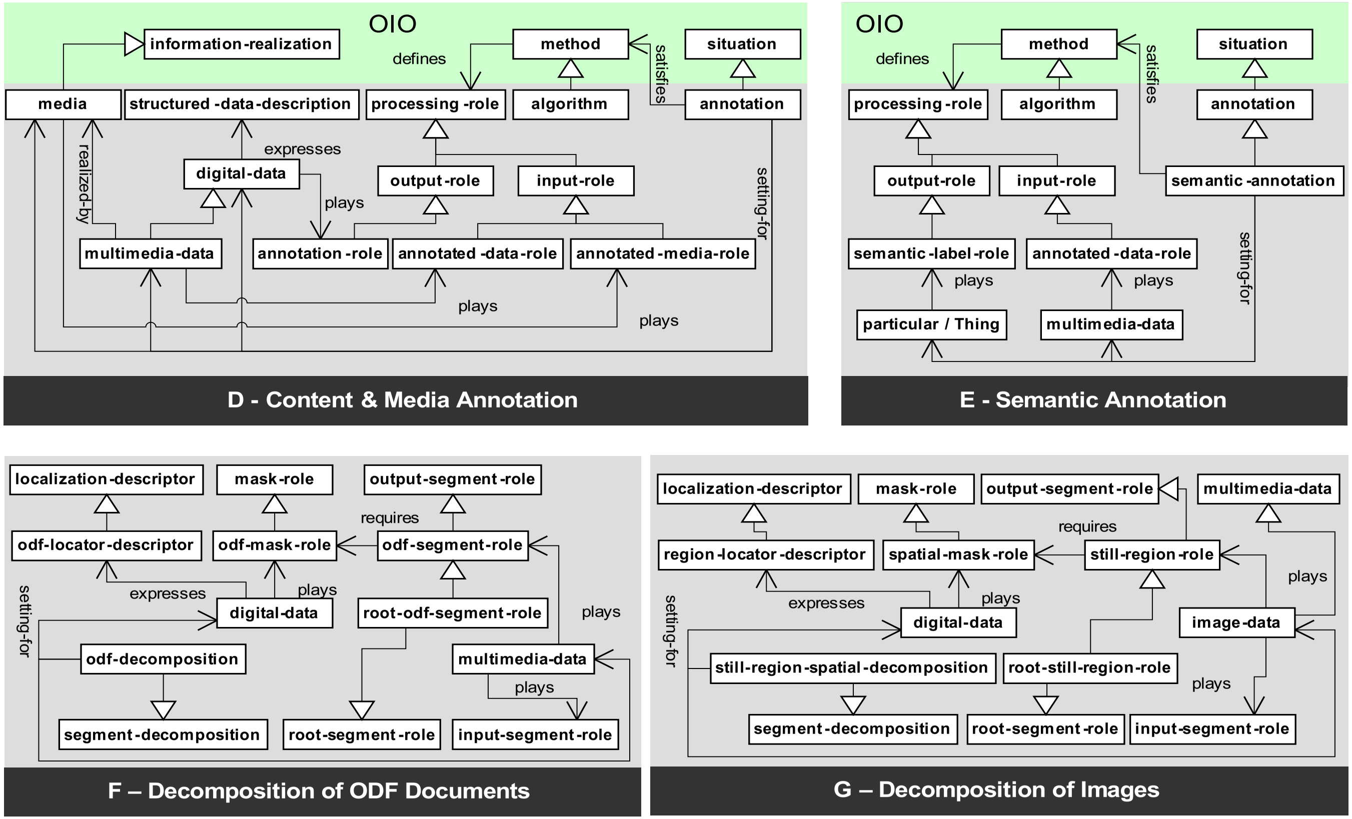


Fig. 5. COMM: Design patterns in UML notation continued: Multimedia patterns Content & Media Annotation and Semantic Annotation (D and E) and modeling example (F)

still regions is called StillRegionSpatialDecompositionType. Consequently, the concept still-region-spatial-decomposition is added as a subconcept of segment-decomposition. Finally, the mask-role concept is specialized by the concept spatial-mask-role. Analogously, the pattern can be used to describe the decomposition of a video asset or of an ODF document (see Fig. 7).

Content Annotation Pattern. This pattern formalizes the attachment of metadata (i.e., annotations) to multimedia-data (Fig. 5-D). Using the D&S pattern, annotations also become situations that represent the state of affairs of all related digital-data (metadata and annotated multimedia-data). digital-data entities represent the attached metadata by playing an annotation-role. These roles are defined by methods or algorithms. The former are used to express manual (or semi-automatic) annotation while the latter serve as an explanation for the attachment of automatically computed features such as the dominant colors of a still region. It is mandatory that the multimedia-data entity being annotated plays an annotated-data-role.

The actual metadata that is carried by a digital-data entity depends on the structured-data-description that is expressed by it. These descriptions are formalized using the digital data pattern (see Section 5.3). Applying the content annotation pattern for formalizing a specific annotation, e.g., a dominant-color-annotation which corresponds to the connection of a MPEG-7 DominantColorType with a segment, requires only the specialization of the concept annotation, e.g., dominant-color-annotation. This concept is defined by being a setting for a digital-data entity that expresses one dominant-color-descriptor (a subconcept of structured-data-description which corresponds to the DominantColorType).

Media Annotation Pattern. This pattern forms the basis for describing the physical instances of multimedia content (Fig. 5-D). It differs from the content annotation pattern in only one respect: it is the media that is being annotated and therefore plays an annotated-media-role.

One can thus represent that some visual content (e.g., the picture of a digital camera) is realized by a JPEG image with a size of 462848 bytes, using the MPEG-7 MediaFormatType. Using the media annotation pattern, the metadata is attached by connecting a digital-data entity with the image. The digital-data plays an annotation-role while the image plays an annotated-media-role. An ontological representation of the MediaFormatType, namely an instance of the structured-data-description subconcept media-format-descriptor, is expressed by the digital-data entity. The tuple formed with the scalar "462848" and the string "JPEG" is the value of the two instances of the concepts file-size and file-format respectively. Both concepts are subconcepts of structured-data-parameter.

Semantic Annotation Pattern. MPEG-7 provides some general concepts (see [52], part 5, section 12) that can be used to describe the perceivable content of a multimedia segment. It is germane to the approach pursued with MPEG-7 that the real world objects appearing in a multimedia document are modeled within the realm of MPEG-7, too. We argue that it is indeed useful to create an ontology specific to multimedia. However, we decline that it was useful to try to model the real world within the very same approach. An ontology-based multimedia annotation framework should rely on domain-specific ontologies for the representation of the real world entities that might be depicted in multimedia content. Consequently, this pattern specializes the content annotation pattern to allow the connection of multimedia descriptions with domain descriptions provided by independent world ontologies (Fig. 5-E).

An OWL Thing or a DOLCE particular (belonging to a domain-specific ontology) that is depicted by some multimedia content is not directly connected to it but rather through the way the annotation is obtained. Actually, a manual annotation method or its subconcept algorithm, such as a classification algorithm, has to be applied to determine this connection. It is embodied through a semantic-annotation that satisfies the applied method. This description specifies that the annotated multimedia-data has to play an annotated-data-role and the depicted Thing/particular has to play a semantic-label-role. The pattern also allows the integration of features which might be evaluated in the context of a classification algorithm. In that case, digital-data entities that represent these features would play an input-role.

5.3 Basic Patterns

Specializing the D&S and OIO patterns for defining multimedia design patterns is enabled through the definition of basic design patterns, which formalize the notion of digital data and algorithm.

Digital Data Pattern. Within the domain of multimedia annotation, the notion of digital data is central—both the multimedia content being annotated and the annotations themselves are expressed as digital data. We consider digital-data entities of arbitrary size to be information-objects, which are used for communication between machines. The OIO design pattern states that descriptions are expressed by information-objects, which have to be about facts (represented by particulars). These facts are settings for situations that have to satisfy the descriptions that are expressed by information-objects. This chain of constraints allows the modeling of complex data structures to store digital information. Our approach is as follows (see Fig. 4-A): digital-data entities express descriptions, namely structured-data-descriptions, which define meaningful labels for the information contained by digital-data. This information is represented by numerical entities such as scalars, matrices, strings, rectangles, or polygons. In DOLCE terms, these entities are abstract-regions. In the context of a description, these regions are described by parameters. structured-data-descriptions thus define structured-data-parameters for which abstract-regions carried by digital-data entities assign values.

The digital data pattern can be used to formalize complex MPEG-7 low-level descriptors. Fig. 4-C shows the application of this pattern by formalizing the MPEG-7 `RegionLocatorType`, which mainly consists of two elements: a `Box` and a `Polygon`. The concept region-locator-descriptor corresponds to the `RegionLocatorType`. The element `Box` is represented by the structured-data-parameter subconcept BoundingBox while the element `Polygon` is represented by the region-boundary concept.

The MPEG-7 code example given in Fig. 1 highlights that the formalization of data structures is not sufficient so far. Complex MPEG-7 types can include nested types that again have to be represented by structured-data-descriptions. In our example, the MPEG-7 `SemanticType` contains the element `Definition` which is of complex type `TextAnnotationType`. The digital data pattern covers such cases by allowing a digital-data instance dd1 to be about a digital-data instance dd2 that expresses a structured-data-description corresponding to a nested type (see Fig. 4-A). In this case, the structured-data-description of instance dd2 would be a part of the one expressed by dd1.

Algorithm Pattern. The production of multimedia annotation can involve the execution of algorithms or the application of computer assisted methods which are used to produce or manipulate digital-data. The recognition of a face in an image region is an example of the former, while manual annotation of the characters is an example of the latter.

We consider algorithms to be methods that are applied to solve a computational problem (see Fig. 4-A). The associated (DOLCE) situations represent the work that is being done by algorithms. Such a situation encompasses digital-data[16] involved in the computation, regions that represent the values of parameters of an algorithm, and perdurants[17] that act as computational-tasks (i.e., the processing steps of an algorithm). An algorithm defines roles that are played by digital-data. These roles encode the meaning of data. In order to solve a problem, an algorithm has to process input data and return some output data. Thus, every algorithm defines at least one input-role and one output-role that both have to be played by digital-data.

5.4 Comparison with Requirements

In the previous sections, we have introduced COMM as a formal ontological foundation for multimedia. We now discuss whether the requirements stated in Section 3 are satisfied with our proposed modeling of the multimedia ontology.

MPEG-7 compliance. The ontology is MPEG-7 compliant since the patterns have been designed with the aim of translating the standard into DOLCE. It covers the most important part of MPEG-7 that is commonly used for describing the structure and the content of multimedia documents. Our current investigation shows that parts of MPEG-7 that have not yet been considered (e.g., navigation & access) can be formalized analogously to the other descriptors through the definition of further patterns. The technical realization of the basic MPEG-7 data types (e.g., matrices and vectors) is not within the scope of the multimedia ontology. They are represented as ontological concepts, because

[16] digital-data entities are DOLCE endurants, i.e., entities that exist in time and space.

[17] Events, processes, or phenomena are examples of perdurants. endurants participate in perdurants.

the about relationship that connects digital-data with numerical entities is only defined between concepts. Thus, the definition of OWL data type properties is required to connect instances of data type concepts (subconcepts of the DOLCE abstract-region) with the actual numeric information (e.g., xsd:string). Currently, simple string representation formats are used for serializing data type concepts (e.g., rectangle) that are currently not covered by W3C standards. Future work includes the integration of the extended data types of OWL 1.1.

Semantic and syntactic interoperability. The syntactic and semantic interoperability of our multimedia ontology is achieved by an OWL DL formalization[18]. Similar to DOLCE, we provide a rich axiomatization of each pattern using first order logic. Our ontology can be linked to any web-based, domain-specific ontology through the semantic annotation pattern.

Separation of concerns. A clear separation of concerns is ensured through the use of the multimedia patterns: the decomposition pattern for handling the structure and the annotation pattern for dealing with the metadata.

Modularity. The decomposition and annotation patterns form the core of the modular architecture of the multimedia ontology. We follow the various MPEG-7 parts and organize the multimedia ontology into modules which cover *i)* the descriptors related to a specific media type (e.g., visual, audio or text) and *ii)* the descriptors that are generic to a particular media (e.g., media descriptors). We also design a separate module for data types in order to abstract from their technical realization.

Extensibility. Through the use of multimedia design patterns, our ontology is also extensible. It allows inclusion of further media types and descriptors (e.g., new low-level features) using the same patterns. As our patterns are grounded in the D&S pattern, it is straightforward to include further contextual knowledge (e.g., about provenance) by adding roles or parameters. Such extensions will not change the patterns, so that legacy annotations will remain valid.

5.5 Related Work

In the field of semantic image understanding, using a multimedia ontology infrastructure is regarded to be the first step for closing the, so-called, semantic gap between low-level signal processing results and explicit semantic descriptions of the concepts depicted in images. Furthermore, multimedia ontologies have the potential to increase the interoperability of applications producing and consuming multimedia annotations. The application of multimedia reasoning techniques on top of semantic multimedia annotations is also a research topic which is currently investigated [59]. A number of drawbacks of MPEG-7 have been reported [58, 63]. As a solution, multimedia ontologies based on MPEG-7 have been proposed.

[18] Examples of the axiomatization are available on the COMM website.

Table 1. Summary of the different MPEG-7 based Multimedia Ontologies

	Hunter	**DS-MIRF**	**Rhizomik**	**COMM**
Foundations	ABC	none	none	DOLCE
Complexity	OWL-Full[a]	OWL-DL[b]	OWL-DL[c]	OWL-DL[d]
Coverage	MDS+Visual	MDS+CS	All	MDS+Visual
Reference	[32]	[92]	[19]	[3]
Applications	Digital Libraries, e-Research	Digital Libraries, e-Learning	Digital Rights Management, e-Business	Multimedia Analysis and Annotations

[a] `http://metadata.net/mpeg7/`
[b] `http://www.music.tuc.gr/ontologies/MPEG703.zip`
[c] `http://rhizomik.net/ontologies/mpeg7ontos`
[d] `http://multimedia.semanticweb.org/COMM/`

From 2001 until the present time, there are four main ontologies that formalize the MPEG-7 standard using Semantic Web languages. Besides COMM, these are the ontology by Hunter, DS-MIRF and the ontology by Rhizomik. In the following, we describe these four ontologies, and the main characteristics as well as the context in which they have been developed are summarized in the Table 1.

Hunter's MPEG-7 ontology. In 2001, Hunter proposed an initial manual translation of MPEG-7 into RDFS (and then into DAML+OIL) and provided a rationale for its use within the Semantic Web [32]. This multimedia ontology was translated into OWL, extended and harmonized using the ABC upper ontology [43] for applications in the digital libraries [33, 34] and eResearch fields [36].

The current version is an OWL Full ontology containing classes defining the media types (Audio, AudioVisual, Image, Multimedia, Video) and the decompositions from the MPEG-7 Multimedia Description Schemes (MDS) part [52]. The descriptors for recording information about the production and creation, usage, structure, and the media features are also defined. The ontology can be viewed in Protégé[19] and has been validated using the WonderWeb OWL Validator[20].

This ontology has usually been applied to describe the decomposition of images and their visual descriptors for use in larger semantic frameworks. Harmonizing through an upper ontology, such as ABC, enables queries for abstract concepts such as subclasses of *events* or *agents* to return media objects or segments of media objects. While the ontology has most often been applied in conjunction with the ABC upper model, it is independent of that ontology and can also be harmonized with other upper ontologies such as SUMO [66] or DOLCE [18].

DS-MIRF ontology. In 2004, Tsinaraki et al. have proposed the DS-MIRF ontology that fully captures in OWL DL the semantics of the MPEG-7 MDS and the Classification Schemes. The ontology can be visualized with GraphOnto or Protégé and has been validated and classified with the WonderWeb OWL

[19] `http://protege.stanford.edu/`
[20] `http://www.mygrid.org.uk/OWL/Validator`

Validator. The ontology has been integrated with OWL domain ontologies for soccer and Formula 1 [93] in order to demonstrate how domain knowledge can be systematically integrated in the general-purpose constructs of MPEG-7. This ontological infrastructure has been utilized in several applications, including audiovisual digital libraries and e-learning.

The DS-MIRF Ontology has been conceptualized manually, according to the methodology outlined in [92]. The XML Schema simple datatypes defined in MPEG-7 are stored in a separate XML Schema to be imported in the DS-MIRF ontology. The naming of the XML elements are generally kept in the `rdf:ID`s of the corresponding OWL entities, except when two different XML Schema constructs have the same names. The mapping between the original names of the MPEG-7 descriptors and the `rdf:ID`s of the corresponding OWL entities is represented in an OWL DL mapping ontology. Therefore, this ontology will represent, e.g., that the `Name` element of the MPEG-7 type `TermUseType` is represented by the `TermName` object property, while the `Name` element of the MPEG-7 type `PlaceType` is represented by the `Name` object property in the DS-MIRF ontology. The mapping ontology also captures the semantics of the XML Schemas that cannot be mapped to OWL constructs such as the sequence element order or the default values of the attributes. Hence, it is possible to return to an original MPEG-7 description from the RDF metadata using this mapping ontology. This process has been partially implemented in GraphOnto [68], for the OWL entities that represent the `SemanticBaseType` and its descendants.

The generalization of this approach has led to the development of a transformation model for capturing the semantics of any XML Schema in an OWL DL ontology [91]. The original XML Schema is converted into a main OWL DL ontology while a OWL DL mapping ontology keeps trace of the constructs mapped in order to allow circular conversions.

Rhizomik Ontology. In 2005, Garcia and Celma have presented the Rhizomik approach that consists in mapping XML Schema constructs to OWL constructs following a generic XML Schema to OWL together with an XML to RDF conversion [19]. Applied to the MPEG-7 schemas, the resulting ontology covers the whole standard as well as the Classification Schemes and TV Anytime[21]. It can be visualized with Protégé or Swoop[22] and has been validated and classified using the Wonderweb OWL Validator and Pellet.

The Rhizomik ontology was originally expressed in OWL Full, since 23 properties must be modeled using an `rdf:Property` as they have both a data type and object type range, i.e., the corresponding elements are both defined as containers of complex types and simple types. An OWL DL version of the ontology has been produced, solving this problem by creating two different properties (`owl:DatatypeProperty` and `owl:ObjectProperty`) for each of them. This change is also incorporated into the XML2RDF step in order to map the affected input XML elements to the appropriate OWL property (object or datatype) depending on the kind of content of the input XML element.

[21] `http://www.tv-anytime.org`

[22] `http://code.google.com/p/swoop`

The main contribution of this approach is that it benefits from the great amount of metadata that has been already been produced by the XML community. Moreover, it is implemented in the ReDeFer project[23], which allows to automatically map input XML Schemas to OWL ontologies and XML data based on them to RDF metadata following the resulting ontologies. This approach has been used with other large XML Schemas in the Digital Rights Management domain such as MPEG-21 and ODRL [21] or in the E-Business domain [20].

Comparison and Summary. These ontologies have been recently compared with COMM according to three criteria:[24] *i)* the way the multimedia ontology is linked with domain semantics, *ii)* the MPEG-7 coverage of the multimedia ontology, and *iii)* the scalability and modeling rationale of the conceptualization [89]. Unlike COMM, all the other ontologies perform a one to one translation of MPEG-7 types into OWL concepts and properties. However, this translation does not guarantee that the intended semantics of MPEG-7 is fully captured and formalized. On the contrary, the syntactic interoperability and conceptual ambiguity problems illustrated in Section 2 remain. Although COMM is based on a foundational ontology, the annotations proved to be no more verbose than those in MPEG-7.

Finally, general models for annotations of non-multimedia content have been proposed by librarians. The Functional Requirements for Bibliographic Records (FRBR)[25] model specifies the conventions for bibliographic description of traditional books. The CIDOC Conceptual Reference Model (CRM)[26] defines the formal structure for describing the concepts and relationships used in cultural heritage documentation. Hunter has described how an MPEG-7 ontology could specialize CIDOC-CRM for describing multimedia objects in museums [33]. Interoperability with such models is an issue, but interestingly, the design rationale used in these models are often comparable and complementary to foundational ontologies approach.

6 KAT—The K-Space Annotation Tool

The K-Space Annotation Tool (KAT) is a platform for an efficient, semi-automatic semantic annotation of multimedia content. It provides a plugin infrastructure to integrate different annotation support. KAT further consists of a core that allows for instantiation, communication, visualization, and threaded execution of plugins. Plugins communicate using a message mechanism and exchange metadata based on COMM (cf. Section 5). The development of KAT is based on the tool M-Ontomat Annotizer [4], which was developed as a tool for extracting features from multimedia content and linking those features to domain ontologies. However, M-Ontomat did not provide the same flexible

[23] `http://rhizomik.net/redefer`

[24] Available from: `http://mklab.iti.gr/mareso/files/proceedings.pdf`

[25] `http://www.ifla.org/VII/s13/frbr/index.htm`

[26] `http://cidoc.ics.forth.gr/`

infrastructure, was not geared towards annotation and retrieval, and was further not based on such a generic multimedia ontology as COMM.

Within the KAT, a plugin is required to understand COMM annotations in order to determine whether and how it has to process a certain content item and to produce its output according to COMM. There are two major types of plugins, the analysis plugins and visual plugins.

Analysis plugins provide automatic or semi-automatic analysis functionalities of media assets. Examples of analysis plugins could be an image segmentation algorithm that decomposes an image into regions or a key-frame extraction algorithm that extracts the most important frames from a video. The location, size, or boundaries of both segments or key-frames are described as COMM annotations. Since a key-frame is a kind of image data, a key-frame can be processed by the image segmentation algorithm (given that the key-frame data, i.e., the pixels are stored in an appropriate format). The resulting segments are added as annotations to the key-frame in the same way as it was done for an image. In other words, an algorithm does not have to distinguish between key-frames or images. It only has to check the COMM annotations whether the data provides all information that it requires. Besides this, all image-data is treated equally. The fact that a key-frame is part of a video is only important in the context of video processing.

Visual plugins provide the means for visualization of COMM annotations and the associated content. They are responsible for any kind of user interaction. A plugin might register a view, which is responsible for displaying a certain type of content and certain types of annotations. One of the standard plugins delivered with the KAT is the Image Annotation Tool, which is capable of displaying images and their decompositions. A user might add additional regions using different drawing tools and regions might be annotated with ontology concepts and instances. The concepts and instances are displayed by another default plugin, the ontology browser. Using a simple drag&drop mechanism, a region is dropped on a concept or an instance of the ontology, which creates an according annotation. Another visual plugin is the annotation browser, which provides a more structured and media-type independent view on the resulting COMM graph. It does not display the content itself but only the COMM annotation in a tree view.

One might also consider other types of plugins, e.g., plugins to browse and import content from Web 2.0 sites such as Flickr or plugins that provide retrieval functionalities. The plugin architecture of KAT is kept very simple and generic in order to provide for implementing also unforeseen semantic multimedia applications. The foundation on the formally defined and extensible COMM offers easy extension to other types of annotations and content.

A screenshot of the current version of KAT is depicted in Figure 6 showing the image of the "Big Three". Each of them is marked with a bounding box (a type of a segment) and annotated with an instance of the concept Man identifying the specific person. The ontology is displayed on the left-hand side, while the annotation browser is displayed in the lower right corner. In the screenshot, the left most bounding box (referring to Churchill) is selected and all concepts and instances associated with the selected region are displayed in the annotation

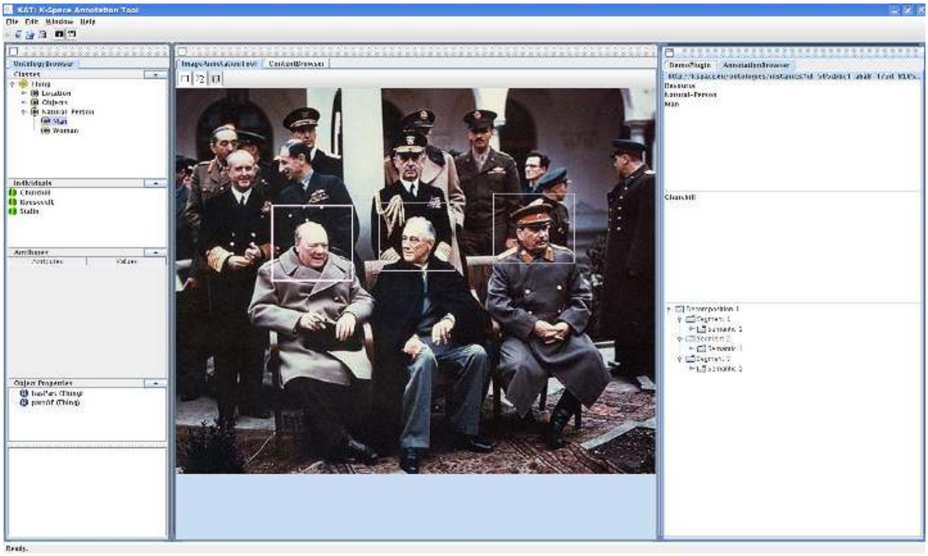


Fig. 6. The KAT showing the annotation of the "Big Three"

browser. The latest information about the KAT as well as binary and source releases are available from `http://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/kat`.

7 Expressing the “Big Three” Scenario in COMM

The interoperability problem with which Nathalie is faced in Section 2 can be solved by using a tool like KAT employing the COMM ontology for representing the metadata of all relevant multimedia objects and the presentation itself throughout the whole creation workflow. The student is shielded from details of the multimedia ontology by embedding it in authoring tools like KAT and feature analysis web services.

The application of the Winston Churchill face recognizer results in an annotation RDF graph that is depicted in the upper part of Fig. 7 (visualized by an UML object diagram[27]). The decomposition of Fig. 1-A, whose content is represented by id0, into one still region (the bounding box of Churchill's face) is represented by the lighter middle part of the UML diagram. The segment is represented by the image-data instance id1 that plays the still-region-role srr1. It is located by the digital-data instance dd1 which expresses the region-locator-descriptor rld1 (lower part of the diagram). Using the semantic annotation pattern, the face recognizer can annotate the still region by connecting it with the URI `http://en.wikipedia.org/wiki/Winston_Churchill`. An instance of an arbitrary domain ontology concept could also have been used for identifying the resource.

[27] The scheme used in Fig. 7 is instance:Concept, the usual UML notation.

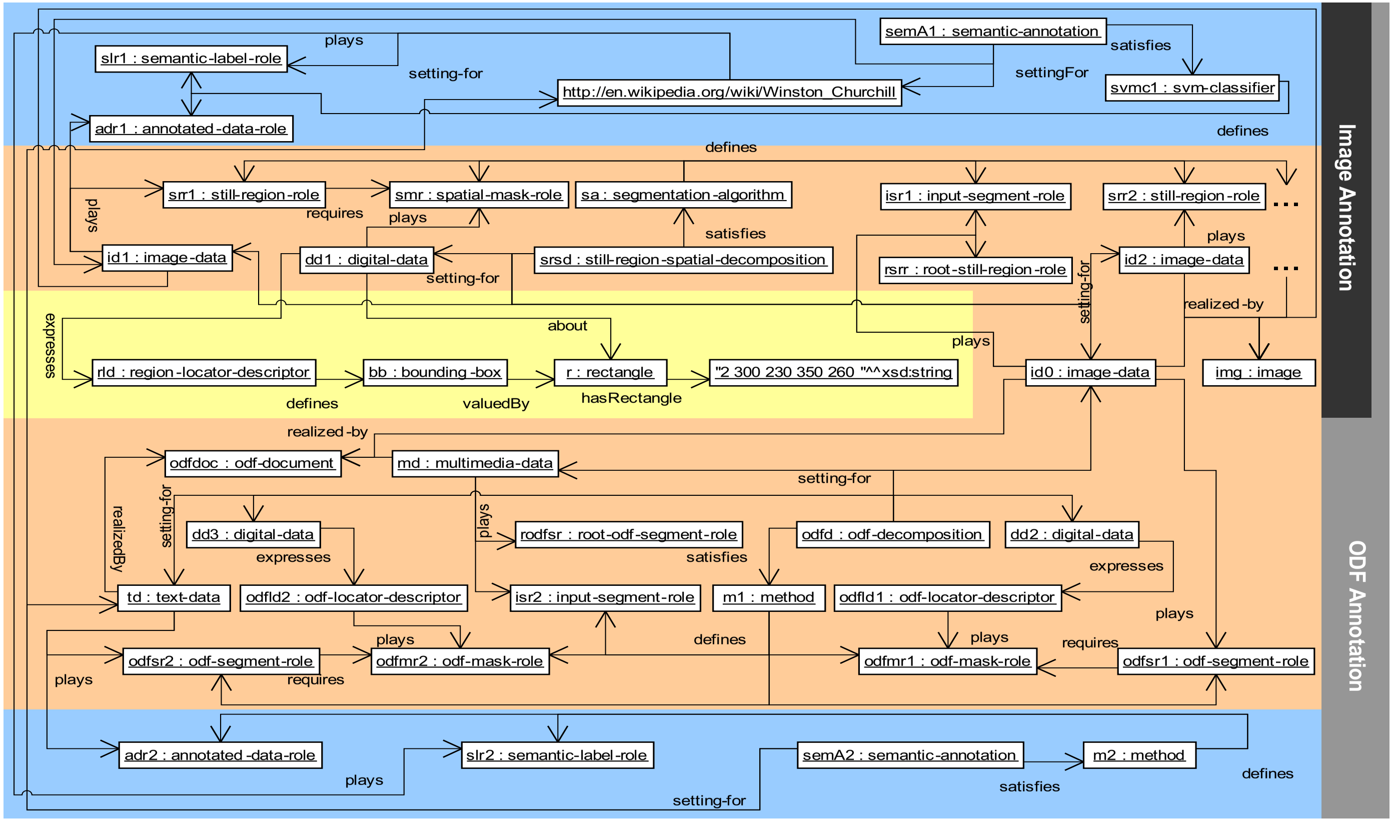


Fig. 7. Annotation of one segment of the Yalta picture and its embedding into an ODF document which contains a text segment that is also about Winston Churchill

Running the two remaining face recognizers for Roosevelt and Stalin will extend the decomposition further by two still regions, i.e., the image-data instances id2 and id3 as well as the corresponding still-region-roles, spatial-mask-roles, and digital-data instances expressing two more region-locator-descriptors (indicated at the right border of Fig. 7). The domain ontologies that provide the instances Roosevelt and Stalin for annotating id2 and id3 with the semantic annotation pattern do not have to be identical to the one that contains Churchill. If several domain ontologies are used, Nathalie can use the OWL sameAs and equivalentClass constructs to align the three face recognition results to the domain ontology that is best suited for enhancing the automatic generation of the multimedia presentation.

Decomposition of ODF documents is formalized analogously to image segmentation (see Fig. 5-F). Therefore, embedding the image annotation into an ODF document annotation is straightforward. The lower part of Fig. 7 shows the decomposition of a compound ODF document into textual and image content. This decomposition description could result from copying an image from the desktop and pasting it into an ODF editor such as OpenOffice. A plugin of this program could produce COMM metadata of the document in the background while it is produced by the user. The media independent design patterns of COMM allow the implementation of a generic mechanism for inserting metadata of arbitrary media assets into already existing metadata of an ODF document. In the case of Fig. 7, the instance id0 (which represents the whole content of the Yalta image) needs to be connected with three instances of the ODF annotation: *i)* the odf-decomposition instance odfd which is a setting-for all top level segments of the odf-document, *ii)* the odf-segment-role instance odfsr1 which identifies id0 as a part of the whole ODF content md (a multimedia-data instance), and *iii)* the instance odfdoc as the image now is also realized-by the odf-document.

Fig. 7 also demonstrates how a domain ontology[28] can be used to define semantically meaningful relations between arbitrary segments. The textual content td as well as the image segment id1 are about Winston Churchill. Consequently, the URI `http://en.wikipedia.org/wiki/Winston_Churchill` is used for annotating both instances using the media independent semantic annotation pattern.

The two segments td and id1 are located within md by two digital-data instances (dd2 and dd3) that express two corresponding odf-locator-descriptor instances. The complete instantiations of the two odf-locator-descriptors are not shown in Fig. 7. The modeling of the region-locator-descriptor, which is completely instantiated in Fig. 7, is shown in Fig. 4-C. The technical details of the odf-locator-descriptor are not presented. However, it is possible to locate segments in ODF documents by storing an XPath which points to the beginning and the end of an ODF segment. Thus, the modeling of the odf-locator-descriptor can be carried out analogously to the region-locator-descriptor.

In order to ease the creation of multimedia annotations with our ontology, we have developed a Java API[29] providing a MPEG-7 class interface for the construction of meta-data at runtime. Annotations that are generated in memory

[28] In this example, the domain ontology corresponds to a collection of Wikipedia URI's.

[29] The Java API is available at `http://multimedia.semanticweb.org/COMM/api/`.

can be exported to Java-based RDF triple stores such as Sesame. To this end, the API translates the objects of the MPEG-7 classes into instances of the COMM concepts. The API also facilitates the implementation of multimedia retrieval tools as it is capable of loading RDF annotation graphs (e.g., the complete annotation of an image including the annotation of arbitrary regions) from a store and converting them back to the MPEG-7 class interface. Using this API, the face recognition web service will automatically create the annotation which is depicted in the upper part of Fig. 7 by executing the code described below.

First of all an image has to be created. In the COMM, an image is formalized as some image-data, that plays a root-segment-role. This is abstracted in the API by creating an image object and assigning a still region (which refers to the image-data) to it (lines 1–3). The bounding box that refers to the recognized face is added as a decomposition to the root still region representing the image. The resulting regions are added as output segments to decomposition object (lines 4–14). Finally the semantic annotation is performed by creating a `Semantic` object. This is assigned a label, which has to be an individual of the domain ontology (in this case the individual representing Winston Churchill). This semantic annotation is then added to the segment (lines 15–18).

```
1  Image img0 = new Image();
2  StillRegion isr0 = new StillRegion();
3  img0.setImage(isr0);
4  StillRegionSpatialDecomposition srsd1 =
5    new StillRegionSpatialDecomposition();
6  isr0.addSpatialDecomposition(srsd1);
7  srsd1.setDescription(new SegmentationAlgorithm());
8  StillRegion srr1 = new StillRegion();
9  srsd1.addStillRegion(srr1);
10 SpatialMask smr1 = new SpatialMask();
11 srr1.setSpatialMask(smr1);
12 RegionLocatorDescriptor rld1 = new RegionLocatorDescriptor();
13 smr1.addSubRegion(rld1);
14 rld1.setBox(new Rectangle(300, 230, 50, 30));
15 Semantic s1 = new Semantic();
16 s1.addLabel("http://en.wikipedia.org/wiki/Winston_Churchill");
17 s1.setDescription(new SVMClassifier());
18 srr1.addSemantic(s1);
```

8 Querying for Semantic Multimedia

So far, we have presented sophisticated support for processing, classifying, and semantically annotating media assets. To be of actual use, these annotations and metadata shall be leveraged to query for media assets. In the K-Space project[30] a database based on Sesame[31] has been developed that allows for storing and querying over RDF triples describing the semantics of media assets. Queries on

[30] http://www.k-space.eu/

[31] Available from http://www.openrdf.org/

semantically-enriched media assets vary from navigating the decomposition of a video into shots and keyframes to retrieving all documents annotated with a certain pattern. Sophisticated queries may even take background knowledge into account. In the scenario presented in Section 2, we might be interested, e.g., in all images showing the heads of the United States and the Soviet Union together. To answer this query, we need to take decompositions of images, semantic annotations, and domain-specific knowledge into account in order to determine whether the persons depicted are heads of the USA or USSR.

In order to process and answering such queries, we are faced with various challenges with respect to the potential size of the dataset, complexity of queries, recursiveness of queries, and interactive access to media asset annotations. These challenges are elaborated below.

Large Datasets. The queried datasets may become extremely large. We estimate annotations of one million triples for one hour of video, which is decomposed into keyframes and annotated region based. If basic inferencing is done to compute subclass and instance relations, this may easily result in an increase by a large constant factor. On the other hand, most state of the art RDF repositories scale to tens or hundreds of million of statements[32]. Only at the time of writing this report, the billion triples border is being attacked [30, 62]. However, such repositories usually require powerful hardware or even clusters of repositories. Compared to this scale, typical datasets of background world knowledge, like DBPedia[33], can almost be considered small.

Complex queries. Queries can become extremely complex. A typical instantiation of a COMM pattern results in up to 20 statements. This complexity is not COMM specific, but typical for multimedia annotation, in order to capture the necessary expressivity [89]. In turn, this results in a query with 20 statement patterns and 19 joins. Given the size of the datasets, this is a challenge that also most existing relational databases fail to meet. Special care needs to be taken to find a very good query execution plan for this kind of queries. In order to avoid errors, it is desirable to hide these complex queries from application developers. In the case of COMM, COMM-API provides an abstraction layer for developers, which allows to access COMM items as Java objects without writing SPARL queries.

Complex recursive queries. Annotations to multimedia items can be done on a variety of levels of decomposition. For example, a whole image can be annotated with a concept but also only a segment showing the concept or a compound ODF document containing the image. Hence, retrieval queries need to recursively follow the decompositions. Standard query languages like the Semantic Web Query Language[34] (SPARQL) do not allow for formulating such recursion. There are extensions to SPARQL that add support for regular path expressions [42, 67].

[32] See http://esw.w3.org/topic/RdfStoreBenchmarking for a good overview of RDF benchmarks.

[33] http://wiki.dbpedia.org/Datasets

[34] http://www.w3.org/2001/sw/DataAccess/

However, such regular expressions are not expressive enough to capture the patterns used in COMM to annotate media assets. For this reason, a metadata repository must additionally support a specialized set of rules that allows to (recursively) follow decompositions during retrieval.

Interactive access to annotations. Multimedia data is often browsed in interactive manner. Hence, drill down and query refinement must be supported for querying semantic multimedia. Given the potential complexity of the dataset and the queries, it must be possible to start a new query from any given point in the annotation graph. For example, if we want to drill down into the annotation of a keyframe in a shot of a video, we should start from the already known shot instead of searching the whole database again. This is facilitated by using RDF for media assets annotations, as in RDF everything is assigned to an URI, e.g., a label, a segment, or a situation.

To illustrate these challenges, we consider two example queries. The first query selects all media assets that show both an US and an USSR leader (lines 10–18 and 28, lines 19–27 and 29). In addition, the direct types, e.g., the image or ODF document and the URLs of realizations of the assets are selected (lines 3 and 5). Please note that we do not specify what an US or USSR leader is. Hence, the query makes use of additional inferencing done over a domain ontology. However, the semantics of all concepts is still clear; in contrast to annotation done in MPEG-7, where the link to a domain ontology can be missing.

```
 1 SELECT ?ITEM ?URI ?TYPE
 2 WHERE {
 3 ?ITEM            custom:directInstanceOf ?TYPE;
 4                  a                core:multimedia-data;
 5                  core:plays       core:root-segment-role;
 6                  core:realized-by ?URI;
 7                  core:plays       ?annotated-data-role1.
 8                  core:plays       ?annotated-data-role2.
 9 ?annotated-data-role  a           core:annotated-data-role.
10 ?annotation1     a                core:semantic-annotation;
11                  core:setting-for ?ITEM;
12                  core:setting-for ?label1;
13                  core:satifies    [
14                          a core:method;
15                          core:defines ?annotated-data-role1;
16                          core:defines ?semantic-label-role1].
17 ?label1          core:plays       ?semantic-label-role1.
18 ?semantic-label-role1 a           core:semantic-label-role.
19 ?annotation2     a                core:semantic-annotation;
20                  core:setting-for ?ITEM;
21                  core:setting-for ?label2;
22                  core:satifies    [
23                          a core:method;
24                          core:defines ?annotated-data-role2;
25                          core:defines ?semantic-label-role2].
26 ?label2          core:plays       ?semantic-label-role2.
?semantic-label-role2 a           core:semantic-label-role.
?label1          a                ex:USLeader.
?label2          a                ex:USSRLeader.
}
```

The second query selects all subsegments of any input segment (lines 13–25) and propagates the semantic annotation of the subsegment (selected in lines 26–34) to the input segment. Here, new COMM annotations are generated using SPARQL construct queries (lines 1–11). If this rule is evaluated recursively, e.g., using Networked Graphs [75], the first query can ignore decompositions and can be formulated in a much shorter way.

```
1  CONSTRUCT {
2  ?ITEM            core:plays       _:annotated-data-role.
3  _:annotation     a                core:semantic-annotation;
4                   core:setting-for ?ITEM;
5                   core:setting-for ?LABEL;
6                   core:satifies    [
7                           a core:method;
8                           core:defines _:annotated-data-role;
9                           core:defines _:semantic-label-role].
10 ?LABEL           core:plays       _:semantic-label-role.
11 _:semantic-label-role a           core:semantic-label-role }
12 WHERE {
13 ?ITEM            a                core:multimedia-data;
14                  core:plays       ?input-segment-role.
15 ?input-segment-role  a            core:input-segment-role.
16 ?decomposition   a                core:decomposition;
17                  core:setting-for ?ITEM;
18                  core:settingFor  ?segment;
19                  core:satisfies   [
20                           a core:method;
21                           core:defines ?segment-role;
22                           core:defines ?input-segment-role].
23 ?segment         core:plays       ?segment-role;
24                  core:plays       ?annotated-data-role.
25 ?segment-role    a                core:segment-role.
26 ?annotation      a                core:semantic-annotation;
27                  core:setting-for ?segment
28                  core:setting-for ?label;
29                  core:satifies    [
30                          a core:method;
31                          core:defines ?annotated-data-role;
32                          core:defines ?semantic-label-role].
33 ?LABEL          core:plays        ?semantic-label-role.
34 ?semantic-label-role a            core:semantic-label-role.
35 }
```

Having presented the challenges of querying multimedia semantics and demonstrated these challenges at the example of two representative queries, we now propose a selection of approaches to deal with the enormous amounts of data we are faced with here.

Partitioning Datasets. In contrast to many sources of world knowledge, multimedia metadata can easily be split horizontally. This means that annotations of two media assets are to a very large degree independent of each other. The links between them are usually indirect, specified through world knowledge. For example, two images could show the same scenery from different angles. However, the scenery is not part of the actual multimedia annotation but world knowledge. As a result, one possible approach to scaling querying of multimedia metadata is to distinguish between multimedia annotation and world knowledge and to accordingly split the datasets and queries. This allows us to come up with easier problems due to shorter queries and a smaller dataset. On the other hand, new challenges arise when splitting queries and datasets such as determining relevant fragments for answering (a part of) a query or joining query results like efficiently handling distributed joins. Even though many of these challenges are well known from distributed and federated relational databases, they are more problematic for RDF as schema information is not reflected in the structure of data and an extremely high number of joins has to be handled compared to relational databases. For illustration, please remember that in relational databases the table structure implicitly reflects the schema of the data. In contrast, in RDF we have triples as the only structure and schema information is expressed explicitly using special predicates.

Appropriate Expressiveness of Languages. State of the art reasoners are not able to deal with the very large datasets we are facing here. To alleviate this issue, again intelligent splitting of data can be applied, using different expressiveness when reasoning with different parts of the dataset. For example, the COMM ontology can still be classified by some OWL-DL reasoners. While this takes a long time, it can be precomputed. Using a pre-classified COMM and some comparable simple query rewriting, we are able to completely avoid reasoning at runtime for many queries. Similar strategies can be used when including knowledge from domain ontologies. We also use a small extension of SPARQL to query for meta-knowledge such as fuzzy values or provenance [82] in order to determine what is more recent, reliable, and so on. Another approach uses fuzzy logic and probabilities to express and manage uncertainty and vagueness, respectively [84].

On Demand Access to Annotation. Due to the enormous size of the metadata, applications cannot hold whole annotation graphs even for moderately complex problems in the main memory. For this reason, we are pursuing a RDF persistence framework to rebuild the COMM API upon. Similar to approaches like Hibernate[35] for relational databases, this allows to read and write only fractions of multimedia annotations on demand. Consequently, we avoid to deal with the whole, very large dataset in memory.

Related Work on Querying. Besides the approach described above for querying media assets by the use of a semantic database, there are also other

[35] `http://www.hibernate.org/`

approaches and solutions to query semantic multimedia. For example, the commercial database Oracle with its Oracle Multimedia[36] feature provides for retrieving images, audio, and video. The Multimedia package is an extension of the relational Oracle database. It supports the extraction of metadata from media assets and allows querying for media assets by specific indices. The Digital Memory Engineering group at the Research Studios in Austria developed with the multimedia database METIS a sophisticated storage and management solution for structured multimedia content and its semantics [40, 73]. The METIS database provides a flexible concept for the definition and management of arbitrary media elements and their semantics. It is adaptable and extensible to the requirements of a concrete application domain by integrating application-specific plugins and defining domain-specific (complex) media types. For it, the semantic relationship of specific media elements and their semantics can be described to form new, independent multimedia data types. Those domain-specific media types can be bundled up and distributed in form of so-called *semantic packs*. The research approach QBIC [16] from IBM is known to be one of the first databases that supports content-based features for querying the content.[37] QBIC supports queries with respect to content-based attributes of images such as color distribution, color layout, and specific textures in the images. Both approaches, Oracle's Intermedia and IBM's QBIC use a relational database and do not provide support for a fully-fledged semantic description of the content such as supported by the K-Space database.

The multimedia presentation algebra (MPA) by Adali et al. [1] extends the relational model of data and allows for dynamically creating new presentations from (parts of) existing presentations. With the MPA, a page-oriented view on multimedia content is given. A multimedia presentation is considered as an interactive presentation that consists of a tree, which is stored in a database. Each node of this tree represents a non-interactive presentation, e.g., a sequence of slides, a video element, or a HTML page. The branches of the tree reflect different possible playback variants of a set of presentations. A transition from a parent node to a child node in this tree corresponds to an interaction. The proposed MPA allows for specifying a query on the database based on the contents of individual nodes as well as querying based on the presentation's tree structure. For it, the MPA provides extensions and generalizations of the `select` and `project` operations in the relational algebra. However, it also allows to author new presentations based on the nodes and tree structure stored in the database. For it, the MPA defines operations such as `merge`, `join`, `path-union`, `path-intersection`, and `path-difference`. These extend the algebraic join operation to tree structures and allow to author new presentations by combining existing presentations and parts of presentations. Another approach comprises a multimedia calculus and algebra allowing for querying on tree-based multimedia content stored in multimedia databases [46, 47]. Here, the new multimedia presentations are created on basis of a given query and a set of inclusion and

[36] `http://www.oracle.com/technology/products/intermedia/index.html`

[37] `http://wwwqbic.almaden.ibm.com/`

exclusion constraints stored in the database. The main advantage of these approaches based on algebras is that the requested multimedia content is specified as a query in a formal language. However, typically high effort is necessary to learn the algebra and their operators and it is very difficult to apply such a formal approach. Consequently, the presented algebras remain purely academic so far.

9 Further and Future Issues of Semantic Multimedia

In this section, we reconsider selected aspects of multimedia semantics. We briefly motivate and summarize them in order to give an outlook to future work.

Semantics for Multimedia. Multimedia semantics exhibits multiple semantics influenced by many different factors like time and contextual use. As motivated in Section 1 and described in Section 4, researchers are looking into the bits and bytes of multimedia content in order to determine its semantics. They also take contextual information about the media assets into account such as EXIF[38] information provided by digital still cameras. In recent time, there is also much research that aims at combining both content-based analysis and context-based analysis in order to improve the results. However, today's approaches and systems typically only look at particular factors that influence multimedia semantics and do not consider the problem in its entirety. Thus, they only look at particular aspects that determine the semantics of multimedia. In order to better understand, describe, and communicate multimedia semantics, a holistic approach is needed that describes and embraces this complex and challenging problem.

A multimedia ontology like COMM presented in Section 5 is an annotation model that can be used to organize and structure multimedia semantics. However, it does not provide support in terms of a method or "high-level" model that helps one in understanding the different factors that make the semantics of multimedia content. Thus, it does not provide for a holistic view we are looking for to better understand multimedia semantics.

The WeKnowIt project[39] aims at understanding the semantics of social media for personal, organizational, and social use through a so-called collective intelligence. The goal of the project is to develop novel techniques for exploiting multiple layers of intelligence from user-generated content. These multiple layers of intelligence together form the collective intelligence that emerges from the collaboration and competition among many individuals. To this end, various sources of information from digital content items and contextual information (media intelligence), massive user feedback (mass intelligence), and users' social interaction (social intelligence) so as to benefit end-users (personal intelligence) as well as organizations (organizational intelligence) will be we analyzed and combined. Thus, it aims at understanding different factors that influence multimedia semantics.

[38] http://www.exif.org/

[39] http://www.weknowit.eu/

With semiotics, we find a general philosophical theory for understanding signs and symbols.[40] It especially deals with the function of signs and symbols in languages and can be broken up into three branches: semantics, syntactics, and pragmatics. Semantics describes the relation between signs and the things they refer to. Syntactics deals with the relation of signs in formal structures. Finally, pragmatics describes the relation of signs to their users and the environment in which they occur. Prominent work in the field of semantics is, e.g., the classification of ten fundamental visual codes by Eco [12]. These codes are an instrument to shape images: codes of perception, codes of transmission, codes of recognition, tonal codes, iconic codes, iconographic codes, codes of taste and sensibility, rhetorical codes, stylistic codes, and codes of the unconscious.[41] Based on this work, concrete systems like the semiotic-aware architecture for hypermedia [56] and the automated video editing tool AUTEUR [55] have been developed, providing valuable achievements in order to understand multimedia semantics.

Finally, we find with the semantics ecosystem a theoretical approach for understanding and modeling semantics [79]. The ecosystem bases on work from the philosopher Popper [69] and defines five different types of semantics (natural, analytical, user, expressive, and emergent semantics) and their relationships. It aims at integrating existing work in the field rather than reinventing it. With natural semantics, we understand the semantics of the non-living physical objects, living things, and events of our physical world. It is the result of the long-term natural language communication between humans. Natural semantics associates basic objects and actions with symbols. Analytical semantics bases on natural semantics. It aims at understanding more complex objects, concepts, and situations. Analytical semantics is applied to dismantle these more complex objects, identify the individual parts, and interpreted them by applying natural semantics. User semantics is the human's perception of the physical world based on his or her personal background. It is the perception of the items, biological objects, and events of the physical world based on a multitude of very different aspects. Among them are the individual's knowledge, preferences, interests, needs, and cultural background [6, 15, 41] and the location, time, used end device, and social situation [10, 80, 81]. With expressive semantics, we consider how the products of the physical world are created. A product can be a gesture, a spoken sentence, or any kind of a non-living object like a book, CD, or multimedia presentation. Expressive semantics describes the intention of the creator when creating such a product (why is the product created in that certain way and what is the intention of the creator in creating it like this). The expressive semantics heavily depends on the individual's background and contextual situation as introduced above. Thus, it depends on the user semantics. Finally, emergent semantics considers the change of semantics over time and use. This means that the individual's semantics and observation of a physical world item, biological object, or event can and will change over time and will change through the different contexts in which it is used. Emergent semantics can be

[40] http://www.merriam-webster.com/dictionary/semiotics

[41] http://www.aber.ac.uk/media/Documents/S4B/sem08.html

short-termed (a couple of seconds up to some minutes) or very long-termed (like a couple of years). However, the key to emergent semantics is the interaction of expressive semantics and analytical semantics. This interaction leads to a modification of user semantics, i.e., the personal ontologies and understanding of the physical world of the individual. Early results of applying parts of the ecosystem in the area of authoring semantically-rich multimedia albums are very promising. However, the ecosystem is still in an early stage and requires maturation. In a future work, it will be very interesting to elaborate how the work on the five types of semantics defined in the ecosystem, the layers of intelligence considered in the WeKnowIt project, and the work in the field of semiotics can be integrated. Thus, what we need is bringing the different ideas and approaches together to provide a better understanding of multimedia semantics.

Organizing, Sharing, and Communicating semantically-rich Multimedia Content. Looking at the field of multimedia semantics and understanding the different contextual factors that determine the multimedia content's semantics raises the question of an appropriate support to organize, share, and communicate such semantically-rich content. Here, we find different systems and applications like Flickr, Picasa[42], and YouTube. The goal of these applications is to provide the users a means to organize and share their experiences. However, these systems and applications focus on the media assets that accompany these experiences, thus they are media-centric. In recent years, it has been reinforced that events are a much better abstraction of human experience [98]. Thus, events are much better for managing media assets captured during events. As a consequence, we find today approaches and applications like SenseCam [23], MyLifeBits [22], PhotoCompas [54], World Explorer [2], FotoFiti [45], PhotoFinder [38, 83], and many more that integrate the concept of events into their media management solution. These are very important and valuable steps towards an event-centric media management. However, the existing approaches and applications typically consider events only as second-class entities, i.e., as some semantics that can be extracted from the media assets and attached to them as additional metadata. Thus, in media-centric approaches, events are considered only one concept among many such as the actual media management, a social network support, and others to describe the multimedia semantics. However, an event-centric management of media assets promises strong advantages over a media-centric approach [76]. Thus, it would be a much better approach for managing the multimedia content's semantics. Early work in this area has been done such as the EMMA system [76]. However, extensive user studies have to be conducted to further underpin this claim.

Annotating Multimedia Assets. Annotating multimedia assets has been introduced in Section 2 at the example of the "Big Three" picture. Looking at today's support for annotating multimedia assets, we typically find support for adding tags (Flickr and YouTube), attaching geo-positions to photos (Zonetag[43]

[42] `http://picasa.google.com/`

[43] `http://zonetag.research.yahoo.com/`

and Locr[44]), defining and annotating regions of interest (ROI) on Flickr, detecting faces with Riya[45], or manually writing and adding comments. Most of these systems are mono-media and allow only for annotations that refer to entire media assets like images and videos. Only a few approaches and systems actually look into fragments of the media assets like ROIs in Flickr and face detection in Riya. For modeling the annotations, typically proprietary formats are used rather then employing standards. This is very unfortunate as many of these standards exist for the different media types (examples are listed in [3]). So far there has not been a broad uptake of these standards for annotating multimedia assets.

In addition, there is a huge lack in providing appropriate annotation support for structured multimedia content such as Flash[46], SMIL[47], SVG[48], or LASeR[49] presentations. Although today's systems and approaches like the Cuypers Multimedia Transformation Engine [24, 95] and the Semi-automatic Multimedia Presentation Generation Environment [13, 14] generate rich multimedia content, exploit semantically-rich annotations and metadata, and even derive further information while authoring the content, this valuable source of information is thrown away once the content creation task is finished. Thus, the created multimedia presentations carry none or only very few annotations.

An approach to (semi-)automatically annotate structured multimedia content during the multimedia authoring process is provided with the SemanticMM4U framework [77, 78]. The framework itself does not define a model for semantic annotation but provides the ability to integrate and use arbitrary ones. These can be simple models like Dublin Core[50] but also complex ones like MPEG-7 or the COMM model introduced in Section 5.

Low Level Multimedia Processing and Classification. We find research in the field of low level multimedia processing and classification already for a couple of decades. A good introduction to this field gives Section 4. The related work shows that classification can be done to a certain degree of accuracy using the different technologies described. However, despite the long-term research in the field there is until today no approach that overcomes the semantic gap. Shape-based approaches can be used to classify arbitrary media assets into a set of classes. However, they remain on the concept level like people/faces, landscapes, nature, and so on. This approach has not reached high-level semantics and annotations of the media assets by proper nouns like determining the peoples' names in the "Big Three" example and identifying that the picture has been taken at the Yalta Conference. Classification using textures allows for identifying objects on the proper noun level (if the objects have been assigned one in the training phase). However, this approach is only applicable for a limited set of objects

[44] `http://www.locr.com/`

[45] `http://www.riya.com/`

[46] `http://www.adobe.com/devnet/swf/`

[47] `http://www.w3.org/AudioVideo/`

[48] `http://www.w3.org/Graphics/SVG/`

[49] `http://www.mpeg-laser.org/html/techSection_laserSpec.htm`

[50] `http://dublincore.org/`

to classify. To alleviate the problem, researchers recently combine traditional content-based classification with additional contextual information such as location, compass, calendar, weather station, and so on. Another promising step to enhance the current state of the art is to combine the so-far uncombined research areas of shape-based processing and texture-based processing.

Most of the work we find today on low level multimedia processing and classification focuses on single media assets like images, video, and audio. However, only little work has been done on analyzing and classifying structured multimedia content such as Flash presentations. An example of low level processing and classification of Flash presentations is by Ding et al. [11].

A Formal Ontological Foundation for Multimedia. MPEG-7 is one of the most renowned metadata standards for annotating media assets. However, as elaborated in Section 1 it became semantically ambiguous due to its complexity. Thus, it lacks from a formal semantics that provides guidelines to the users of the standard how to apply it. With COMM presented in Section 5, we find an approach to describe parts of MPEG-7 using formal semantics based on DOLCE. With the example of expressing complex semantics in the "Big Three" scenario by using the annotation tool KAT in Section 6 and manually applying COMM in Section 7, applicability of COMM for rich semantic annotations is shown. A major challenge for COMM is the high burden and effort needed to start using it. In future, it is to become more practicable and applicable. Thus, what is missing are methods and guidelines how to apply an ontology like COMM to annotate multimedia content and providing tools working with COMM. A fundamental issue here is introducing the concept of modules into ontologies. By this, the complex problem is broken down into smaller bricks and at the same time allows for providing very domain-specific and thus easier to use ontologies. On top of such modularized ontologies we can then define appropriate methods and tools. Enhancing the state of the art here is a key research issue of the NEON project[51].

Querying for Semantic Multimedia. For querying semantic multimedia, we presented in Section 8 a database based on Sesame providing for storing and querying over RDF triples. We also considered related approaches and systems in the field of querying for semantic multimedia. Looking at the current state of the art, a future research issue is providing efficient support for a recursive querying in structured multimedia content over a large dataset. For it, we need effective query optimization algorithms taking pattern similarities of the queries into account. The further, current research allows for querying using fuzzy logic and provenance [82, 84]. Feasibility of this approach is shown by first systems that actually integrate fuzzy logic. However, it remains future challenge to proof real benefit of using fuzzy logic. With respect to provenance, a future challenge is to leverage this information to make decisions about the trustworthiness of specific statements made about the multimedia content. Thus, to establish trust to the user. Finally, we can state that querying for semantic multimedia is a vehicle to bring a vitally needed, sophisticated expressiveness to multimedia metadata.

[51] http://www.neon-project.org/

However, as we could only sketch in Section 8, this sophisticated expressiveness also puts very high demands on the semantic infrastructure used. Consequently, we expect that the demands of semantic multimedia applications significantly drive the development of a semantic web infrastructure in the next years, both in terms of scaleability but also with respect to the expressivity of query languages.

10 Summary and Outlook

In this paper, we presented current research in multimedia semantics. We looked into the field of annotating media assets and elaborated the drawbacks of todays support for annotation such as fragment identification, semantic annotation, web interoperability, and embedding semantic annotations into compound documents. Research in the area low level multimedia processing and classification has been been presented. We identified requirements for designing a multimedia ontology and introduced a formal ontological foundation for multimedia with the multimedia ontology COMM. The multimedia ontology COMM has been used for implementing the multimedia annotation tool KAT and has been applied to annotate the "Big Three" scenario. We investigated the retrieval of multimedia semantics based on SPARQL and considered further and future aspects of multimedia semantics.

As a quintessence of the discussion in Section 9, we conclude with identifying the major challenges for future research in semantic multimedia. These are combining existing research approaches and streams and providing semantics support for structured multimedia content.

Combining research approaches and streams. Recent approaches of combing, e.g., content-based analysis with context-based analysis of media assets have shown that the results achieved here are much better compared to applying the techniques solitary. To further enhance the state of the art in annotating, processing, and classifying media assets, a big challenge for the future is to bring different fields and streams of research and thus different approaches together. Current efforts towards integration of content-based and context-based media understanding reflects this trend. Another example is the so far uncombined research in shape-based classification and texture-based classification. It seems very promising to combine both approaches to bring the field one step further.

Support for structured multimedia content. Most approaches for annotating, processing, and classifying is focused on single media assets such as images, video, and audio. The challenge for the research results in these areas is to extend and to apply it to rich, structured multimedia content.

Acknowledgment. The research that lead to this paper was partially supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content (K-Space), and under contract FP6-026978, X-Media Integrated Project.

References

1. Adali, S., Sapino, M.L., Subrahmanian, V.S.: An algebra for creating and querying multimedia presentations. Multimedia Syst. 8(3), 212–230 (2000)
2. Ahern, S., Naaman, M., Nair, R., Yang, J.H.-I.: World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: Proceedings of the 7th ACM/IEEE joint conference on Digital libraries, pp. 1–10. ACM Press, New York (2007)
3. Arndt, R., Troncy, R., Staab, S., Hardman, L., Vacura, M.: COMM: Designing a Well-Founded Multimedia Ontology for the Web. In: 6th Int. Semantic Web Conference (2007)
4. Blöhdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., Strintzis, M.: Semantic Annotation of Images and Videos for Multimedia Analysis. In: 2nd European Semantic Web Conference (2005)
5. Brunelli, R., Poggio, T.: Template matching: Matched spatial filters and beyond. Pattern Recognition 30(5), 751–768 (1997)
6. Brusilovsky, P., Maybury, M.T.: From adaptive hypermedia to the adaptive Web. Communications of the ACM 45(5), 30–33 (2002)
7. Chen, H., Shimshoni, I., Meer, P.: Model based object recognition by robust information fusion. In: 17th International Conference on Pattern Recognition, Cambrige, UK (August 2004)
8. Christianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)
9. Cortes, C., Vapnik, V.N.: Support vector networks. Machine Learning 20, 273–297 (1995)
10. Dey, A.K., Abowd, G.D.: Towards a Better Understanding of Context and Context-Awareness. Technical Report GIT-GVU-99-22, Graphics, Visualization and Usability Center and College of Computing, Georgia Institute of Technology, Atlanta, GA, USA (June 1999)
11. Ding, D., Yang, J., Li, Q., Liu, W., Wang, L.: What can expressive semantics tell: Retrieval model for a flash-movie search engine. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 123–133. Springer, Heidelberg (2005)
12. Eco, U.: Einfuehrung in die Semiotik. Wilhelm Fink Verlag, Munich (1985)
13. Falkovych, K., Nack, F.: Context Aware Guidance for Multimedia Authoring: Harmonizing domain and discourse knowledge. Multimedia Systems Journal 11(3) (2006)
14. Falkovych, K., Nack, F., van Ossenbruggen, J., Rutledge, L.: Sample: Towards a framework for system-supported multimedia authoring. In: Multimedia Modelling, p. 362. IEEE Computer Society, Los Alamitos (2004)
15. Fink, J., Kobsa, A., Schreck, J.: Personalized hypermedia information through adaptive and adaptable system features: User modeling, privacy and security issues. In: Mullery, A., Besson, M., Campolargo, M., Gobbi, R., Reed, R. (eds.) Intelligence in Services and Networks: Technology for Cooperative Competition, pp. 459–467. Springer, Heidelberg (1997)
16. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: the QBIC system. In: Readings in multimedia computing and networking, pp. 255–264. Morgan Kaufmann, San Francisco (2001)

17. Gangemi, A., Borgo, S., Catenacci, C., Lehmann, J.: Task Taxonomies for Knowledge Content. Technical report, Metokis Deliverable 7, (2004)
18. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with dolce. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)
19. Garcia, R., Celma, O.: Semantic Integration and Retrieval of Multimedia Metadata. In: 5th International Workshop on Knowledge Markup and Semantic Annotation (2005)
20. García, R., Gil, R.: Facilitating Business Interoperability from the Semantic Web. In: Abramowicz, W. (ed.) BIS 2007. LNCS, vol. 4439, pp. 220–232. Springer, Heidelberg (2007)
21. Garcia, R., Gil, R., Delgado, J.: A Web Ontologies Framework for Digital Rights Management. Journal of Artificial Intelligence and Law 15, 137–154 (2007)
22. Gemmell, J., Bell, G., Lueder, R.: Mylifebits: a personal database for everything. Commun. ACM 49(1), 88–95 (2006)
23. Gemmell, J., Williams, L., Wood, K., Lueder, R., Bell, G.: Passive capture and ensuing issues for a personal lifetime store. In: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences, pp. 48–55. ACM Press, New York (2004)
24. Geurts, J., van Ossenbruggen, J., Hardman, L.: Application-specific constraints for multimedia presentation generation. In: Multimedia Modeling. IEEE, Los Alamitos (2001)
25. Geurts, J., van Ossenbruggen, J., Hardman, L.: Requirements for practical multimedia annotation. In: Workshop on Multimedia and the Semantic Web (2005)
26. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, Englewood Cliffs (2001)
27. Gräßl, C., Deinzer, F., Nieman, H.: Continuous parametrization of normal distribution for improving the discrete statistical eigenspace approach for object recognition. In: Krasnoproshin, V., Ablameyko, S., Soldek, J. (eds.) Pattern Recognition and Information Processing 2003, Minsk, Belarus, May 2003, pp. 73–77 (2003)
28. Grzegorzek, M., Izquierdo, E.: Statistical 3d object classification and localization with context modeling. In: Domanski, M., Stasinski, R., Bartkowiak, M. (eds.) 15th European Signal Processing Conference, pp. 1585–1589. PTETiS, Poznan (2007)
29. Halasz, F., Schwartz, M.: The Dexter Hypertext Reference Model. Communications of the ACM 37(2), 30–39 (1994)
30. Harth, A., Umbrich, J., Hogan, A., Decker, S.: Yars2: A federated repository for searching and querying graph structured data. Technical report, Digital Enterprise Research Institute, Galway, 4 (2007)
31. Hornegger, J.: Statistische Modellierung, Klassifikation und Lokalisation von Objekten. Shaker Verlag, Aachen (1996)
32. Hunter, J.: Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In: 1st International Semantic Web Working Symposium, pp. 261–281 (2001)
33. Hunter, J.: Combining the CIDOC/CRM and MPEG-7 to Describe Multimedia in Museums. In: 6th Museums and the Web Conference (2002), http://www.archimuse.com/mw2002/papers/hunter/hunter.html
34. Hunter, J.: Enhancing the semantic interoperability of multimedia through a core ontology. IEEE Transactions on Circuits and Systems for Video Technology 13(1), 49–58 (2003)
35. Hunter, J., Armstrong, L.: A Comparison of Schemas for Video Metadata Representation. In: 8th International World Wide Web Conference, pp. 1431–1451 (1999)

36. Hunter, J., Little, S.: A Framework to Enable the Semantic Inferencing and Querying of Multimedia Content. International Journal of Web Engineering and Technology – Special Issue on the Semantic Web 2(2/3), 264–286 (2005)
37. Isaac, A., Troncy, R.: Designing and Using an Audio-Visual Description Core Ontology. In: Workshop on Core Ontologies in Ontology Engineering (2004)
38. Kang, H., Shneiderman, B.: Visualization methods for personal photo collections: Browsing and searching in the photofinder. In: IEEE International Conference on Multimedia and Expo (III), pp. 1539–1542 (August 2000)
39. Kerr, J., Compton, P.: Toward generic model-based object recognition by knowledge acquisition and machine learning. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, pp. 9–15 (2003)
40. King, R., Popitsch, N., Westermann, U.: METIS: a flexible database foundation for unified media management. In: Proc.of the 12th annual ACM Int. Conf. on Multimedia, pp. 744–745. ACM Press, New York (2004)
41. Kobsa, A., Koenemann, J., Pohl, W.: Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. In: The Knowledge Engineering Review, vol. 16, pp. 111–155. Cambridge University Press, Cambridge (2001)
42. Kochut, K., Janik, M.: Sparqler: Extended sparql for semantic association discovery. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 145–159. Springer, Heidelberg (2007)
43. Lagoze, C., Hunter, J.: The ABC Ontology and Model (v3.0). Journal of Digital Information 2(2) (2001)
44. Latecki, L.J., Lakaemper, R., Wolter, D.: Optimal partial shape similarity. Image and Vision Computing Journal 23, 227–236 (2005)
45. Lee, B.N., Chen, W., Chang, E.Y.: Fotofiti: web service for photo management. In: Proceedings of the 14th annual ACM international conference on Multimedia, pp. 485–486. ACM Press, New York (2006)
46. Lee, T., Sheng, L., Balkir, N.H., Al-Hamdani, A., Özsoyoglu, G., Özsoyoglu, Z.M.: Query Processing Techniques for Multimedia Presentations. Multimedia Tools Appl. 11(1), 63–99 (2000)
47. Lee, T., Sheng, L., Bozkaya, T., Balkir, N.H., Özsoyoglu, Z.M., Özsoyoglu, G.: Querying Multimedia Presentations Based on Content. IEEE Trans. on Knowledge and Data Engineering 11(3), 361–385 (1999)
48. Leonardis, A., Bischof, H.: Dealing with occlusions in the eigenspace approach. In: Pelillo, M., Hancock, E.R. (eds.) EMMCVPR 1997. LNCS, vol. 1223, pp. 453–458. Springer, Heidelberg (1997)
49. Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. PAMI 19(7), 696–710 (1997)
50. Mokhtarian, F., Bober, M.: Curvature Scale Space Representation: Theory, Applications, and MPEG7-Standardization. Springer, Heidelberg (2003)
51. MPEG-21. Part 17: Fragment Identification of MPEG Resources. Standard No. ISO/IEC 21000-17 (2006)
52. MPEG-7. Multimedia Content Description Interface. Standard No. ISO/IEC 15938 (2001)
53. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-d objects from appearance. International Journal of Computer Vision 14(1), 5–24 (1995)
54. Naaman, M., Yeh, R.B., Garcia-Molina, H., Paepcke, A.: Leveraging context to resolve identity in photo albums. In: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, pp. 178–187. ACM Press, New York (2005)

55. Nack, F.: AUTEUR: The Application of Video Semantics and Theme Representation for Automated Film Editing. PhD thesis, Lancaster University, UK (September 1996)
56. Nack, F., Hardman, L.: Denotative and connotative semantics in hypermedia: proposal for a semiotic-aware architecture. New Rev. Hypermedia Multimedia 7(1), 7–37 (2002)
57. Nack, F., Lindsay, A.T.: Everything you wanted to know about MPEG-7 (Parts I & II). IEEE Multimedia 6(3-4) (1999)
58. Nack, F., van Ossenbruggen, J., Hardman, L.: That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). IEEE Multimedia 12(1) (2005)
59. Neumann, B., Möller, R.: On Scene Interpretation with Description Logics. In: Cognitive Vision Systems, pp. 247–275. Springer, Heidelberg (2006)
60. Niemann, H.: Klassifikation von Mustern. Springer, Heidelberg (1983)
61. Oberle, D., Lamparter, S., Grimm, S., Vrandecic, D., Staab, S., Gangemi, A.: Towards Ontologies for Formalizing Modularization and Communication in Large Software Systems. Journal of Applied Ontology 1(2), 163–202 (2006)
62. Sirma Group Corp Ontotext Lab. Bigowlim: System documentation (2006) [15-05-2008], `http://www.ontotext.com/owlim/big/BigOWLIMSysDoc.pdf`
63. van Ossenbruggen, J., Nack, F., Hardman, L.: That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). IEEE Multimedia 11(4) (2004)
64. Park, S., Lee, J., Kim, S.: Content-based image classification using a neural network. Pattern Recognition Letters 25(3), 287–300 (2004)
65. Paulus, D., Hornegger, J.: Applied Pattern Recognition. Friedr. Vieweg & Sohn Verlagsgesellschaft GmbH, Braunschweig (2003)
66. Pease, A., Niles, I., Li, J.: The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In: Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web (2002)
67. Polleres, A., Scharffe, F., Schindlauer, R.: Sparql++ for mapping between rdf vocabularies. In: OTM Conferences (1), pp. 878–896 (2007)
68. Polydoros, P., Tsinaraki, C., Christodoulakis, S.: GraphOnto: OWL-based ontology management and multimedia annotation in the DS-MIRF framework. Journal of Digital Information Management (JDIM) 4(4), 214–219 (2006)
69. Popper, K.: Three worlds [the tanner lecture on human values: Delivered at the university of michigan], April (1978), `http://www.tannerlectures.utah.edu/lectures/documents/popper80.pdf`
70. Pösl, J.: Erscheinungsbasierte, statistische Objekterkennung. Shaker Verlag, Aachen (1999)
71. Pratt, W.K.: Digital Image Processing. John Wiley & Sons Ltd., New York (2001)
72. Reinhold, M.: Robuste, probabilistische, erscheinungsbasierte Objekterkennung. Logos Verlag, Berlin (2004)
73. Ross, K., Westermann, G.U., Popitsch, N.: METIS - A Flexible Database Solution for the Management of Multimedia Assets. In: Proc. of the 10th Int. Workshop on Multimedia Information Systems, College Park, MD, USA (August 2004)
74. Saathoff, C., Staab, S.: Exploiting Spatial Context in Images Using Fuzzy Constraint Reasoning. In: 9th Int. Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, Austria. IEEE, Los Alamitos (2008)
75. Schenk, S., Staab, S.: Networked graphs: A declarative mechanism for sparql rules, sparql views and rdf data integration on the web. In: Proceedings of the 17th International World Wide Web Conference, WWW2008, Bejing, China (2008)

76. Scherp, A., Agaram, S., Jain, R.: Event-centric media management. In: Gevers, T., Jain, R.C., Santini, S. (eds.) Multimedia Content Access: Algorithms and Systems II. Proceedings of the SPIE Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, vol. 6820, pp. 68200C-68200C-15 (January 2008)
77. Scherp, A.: Semantics support for personalized multimedia content. In: Int. Conf. Internet and Multimedia Systems and Applications, Innsbruck, Austria, March 2008, pp. 57–65. IASTED (2008)
78. Scherp, A., Boll, S., Cremer, H.: Emergent semantics in personalized multimedia content. J. of Digital Information Management 5(2) (April 2007)
79. Scherp, A., Jain, R.: Towards an ecosystem for semantics. In: MS 2007: Workshop on multimedia information retrieval on The many faces of multimedia semantics, pp. 3–12. ACM Press, New York (2007)
80. Schilit, B., Adams, N., Want, R.: Context-Aware Computing Applications. In: Workshop on Mobil Computing Systems and Applications, Santa Cruz, CA, USA, pp. 85–90. IEEE, Los Alamitos (1994)
81. Schmidt, A., Beigl, M., Gellersen, H.-W.: There is more to context than location. Computers & Graphics 23(6), 893–901 (1999)
82. Schueler, B., Sizov, S., Staab, S., Tran, D.T.: Querying for meta knowledge. In: WWW 2008: Proceeding of the 17th international conference on World Wide Web, pp. 625–634. ACM, New York (2008)
83. Shneiderman, B., Kang, H.: Direct annotation: A drag-and-drop strategy for labeling photos. In: Proceedings of the International Conference on Information Visualisation, p. 88. IEEE Computer Society, Washington (2000)
84. Straccia, U.: Managing Uncertainty and Vagueness in Description Logics, Logic Programs and Description Logic Programs. Springer, Heidelberg (2008)
85. Sure, Y., Staab, S., Studer, R.: Methodology for development and employment of ontology based knowledge management applications. SIGMOD Rec. 31(4), 18–23 (2002)
86. Tadeusiewicz, R.: Introduction to Practice of Application of Neural Networks (in Neuron Networks) StatSoft, Warsaw, Poland (1999)
87. Troncy, R.: Integrating Structure and Semantics into Audio-visual Documents. In: 2nd International Semantic Web Conference, pp. 566–581 (2003)
88. Troncy, R., Bailer, W., Hausenblas, M., Hofmair, P., Schlatte, R.: Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles. In: 1st International Conference on Semantics And digital Media Technology, pp. 41–55 (2006)
89. Troncy, R., Celma, Ó., Little, S., García, R., Tsinaraki, C.: MPEG-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In: 1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies, pp. 2–15 (2007)
90. Troncy, R., Hardman, L., van Ossenbruggen, J., Hausenblas, M.: Identifying Spatial and Temporal Media Fragments on the Web. In: W3C Video on the Web Workshop (2007), http://www.w3.org/2007/08/video/positions/Troncy.pdf
91. Tsinaraki, C., Christodoulakis, S.: Interoperability of XML Schema Applications with OWL Domain Knowledge and Semantic Web Tools. In: 6th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE) (2007)
92. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support for Ontology-based Video Retrieval Applications. In: 3rd International Conference on Image and Video Retrieval (CIVR), pp. 582–591 (2004)

93. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support between MPEG-7/21 and OWL in DS-MIRF. Transactions on Knowledge and Data Engineering (TKDE) 19(2), 219–232 (2007) (Special Issue on the Semantic Web Era)
94. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Conference on Computer Vision and Pattern Recognition, Maui, USA, pp. 586–591 (June 1991)
95. van Ossenbruggen, J., Hardman, L., Geurts, J., Rutledge, L.: Towards a multimedia formatting vocabulary. In: World Wide Web. ACM, New York (2003)
96. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
97. Walter, J., Arnrich, B.: Gabor filters for object localization and robot grasping. In: Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, September 2000. ICSP, pp. 124–127 (2000)
98. Westermann, U., Jain, R.: Toward a common event model for multimedia applications. IEEE MultiMedia 14(1), 19–29 (2007)
99. Yuan, C., Niemann, H.: Neural networks for the recognition and pose estimation of 3-d objects from a single 2-d perspective view. International Journal of Image and Vision Computing 19, 585–592 (2001)