Proceedings of the Federated Conference on Computer Science and Information Systems pp. 91–98
DOI: 10.15439/2017F66
ISSN 2300-5963 ACSIS, Vol. 11

# Utilizing Multimedia Ontologies in Video Scene Interpretation via Information Fusion and Automated Reasoning

Leslie F. Sikos
School of Computer Science,
Engineering and Mathematics
Flinders University
GPO Box 2100
Adelaide SA 5001
Australia
Email: leslie.sikos@ieee.org

***Abstract*—There is an overwhelming variety of multimedia ontologies used to narrow the semantic gap, many of which are overlapping, not richly axiomatized, do not provide a proper taxonomical structure, and do not define complex correlations between concepts and roles. Moreover, not all ontologies used for image annotation are suitable for video scene representation, due to the lack of rich high-level semantics and spatiotemporal formalisms. This paper presents an approach for combining multimedia ontologies for video scene representation, while taking into account the specificity of the scenes to describe, minimizing the number of ontologies, complying with standards, minimizing reasoning complexity, and whenever possible, maintaining decidability.**

## I. Introduction

IN THE last 15 years, narrowing the notorious *semantic gap* in video understanding has very much been neglected compared to image interpretation [1]. For this reason, most research efforts have been limited to frame-based concept mapping so that the corresponding techniques could be applied from the results of the research communities of image semantics. However, these approaches failed to exploit the temporal information and multiple modalities typical to videos.

Most domain ontologies developed for defining multimedia concepts with or without standards alignment went from one extreme to the other; they attempted to cover either a very narrow and specific knowledge domain that cannot be used for unconstrained videos, or an overly generic taxonomy for the most commonly depicted objects of video databases, which do not hold rich semantics.

Further structured data sources used for concept mapping include commonsense knowledge bases, upper ontologies, and Linked Open Data (LOD) datasets. Very few research have actually been done to standardize the corresponding resources, without which combining low-level image, audio, and video descriptors, and sophisticated high-level descriptors with rule-based video event definitions cannot be efficient. An early implementation in this field was a core audiovisual ontology based on MPEG-7, ProgramGuideML, and TV Anytime [2]. A more recent research outcome is the core reference ontology VidOnt,[1] which aims to act as a mediator between de facto standard and standard video and video-related ontologies [3].

## II. Problem Statement

Despite the large number of multimedia ontologies mentioned in the literature, there are very few ontologies that can be employed in video scene representation. Most problems and limitations of these ontologies indicate ontology engineering issues, such as lack of formal grounding, failure to determine the scope of the ontology, overgeneralization, and using a basic subset of the mathematical constructors available in the implementation language [4]. Capturing the associated semantics has quite often been exhausted by creating a taxonomical structure for a specific knowledge domain using the Protégé ontology editor,[2] and not only domain and range definitions are not used for properties, but even the property type is often incorrect.

As a result, implementing multimedia ontologies in video scene representation is not straightforward. For this reason, a novel approach has been introduced, which captures the highest possible semantics in video scenes.

## III. Towards a Methodology for Combining Multimedia Ontologies for Video Scene Representation

The representation of video scenes largely depends on the target application, such as content-based video scene retrieval and hypervideo playback. Hence, the different requirements have to be set on a case-by-case basis. Nevertheless, there are common steps for structured video annotation, such as determining the desired balance between expressivity and reasoning complexity, capturing the intended semantics for the knowledge domain featured in the video or required by the application, and standards compliance. The proposed approach guides through the key factors to be con-

[1] http://vidont.org
[2] http://protege.stanford.edu

sidered in order to achieve the optimal level of semantic enrichment for video scenes.

### A. Intended Semantics

In contrast to image annotation, in which the intended semantics can typically be captured using concepts from domain or upper ontologies, the spatiotemporal annotation of video scenes requires a wide range of highly specialized ontologies.

The numeric representation of audio waveforms, the edges, interest points, regions of interest, ridges, and other visual features of video frames and video clips employ low-level descriptors, usually from an OWL mapping of MPEG-7's XSD vocabulary. They correspond to local and global characteristics of video frames, and audio and video signals, such as intensity, frequency, distribution, pixel groups, and low-level feature aggregates, such as various histograms and moments based on low-level features. Some examples for audio descriptors include the zero crossing rate descriptor, which can be used to determine whether the audio channel contains speech or music, the descriptors of formants parameters, which are suitable for phoneme and vowel identification, and the attack duration descriptor, which is used for sound identification. Two feature aggregates frequently used for video representation are SIFT (Scale-Invariant Feature Transform), which is suitable for object recognition and tracking in videos [5], and HOF (Histogram of Optical Flow) [6], which can be used for, among others, detecting humans in videos. The most common motion descriptors include the camera motion descriptor, which can characterize a video scene in a particular time according to professional video camera movements, the motion activity descriptor, which can be used to indicate the spatial and temporal distribution of activities, and the motion trajectory descriptor, which represents the displacement of objects over time.

The MPEG-7 descriptors can be used for tasks such as generating video summaries [7] and matching video clips [8], however, they do not convey information about the meaning of audiovisual contents, i.e., they cannot provide high-level semantics [9]. Nevertheless, MPEG-7 terms can be used for low-level descriptors. However, using partial mappings of MPEG-7 limits semantic enrichment, because video representation requires a wide range of multimedia descriptors. Therefore, an ontology supporting only the visual descriptors of MPEG-7, such as the Visual Descriptor Ontology (VDO) [10], for example, omits audio descriptors that can be used for describing the audio channel of videos. In fact, even a complete mapping of MPEG-7 does not guarantee semantic enrichment, such as the ones created via a transparent XSD-OWL translation (e.g., Rhizomik),[3] particularly when the mathematical constructors are not exploited to their full potential [11].

Common high-level video concepts can be utilized from Schema.org.[4] For example, generic video metadata can be provided for video objects using `schema:video` and `schema:VideoObject`. Movies, series, seasons, and episodes of series can be described using `schema:Movie`, `schema:MovieSeries`, `schema:CreativeWorkSeason`, and `schema:Episode`. Analogously, video metadata can be described using `schema:duration`, `schema:genre`, `schema:inLanguage`, and similar properties. Rich video semantics can be described using specialized ontologies, such as the STIMONT ontology, which can capture the emotional responses associated with videos [12]. The use of more specific high-level concepts depends on the knowledge domain to represent, and often includes Linked Data [13].

*Criteria*

- A1. The ontology or dataset captures the intended semantics or the semantics closest to the intended semantics in terms of concept and property definitions.
- A2. The terms to be used for annotation are defined in a standardized ontology or dataset. If this is not available, or there are similar or identical definitions available in multiple ontologies or datasets, the choice is determined by the following precedence order: 1) standard, 2) standard-aligned, 3) de facto standard, 4) proprietary.

### B. Quality of Conceptualization

Another important consideration beyond capturing the intended semantics is the quality of conceptualization. For example, the MPEG-7 mappings known for the literature transformed semistructured definitions to structured data, but this did not make them suitable for reasoning over visual contents. Since MPEG-7 provides low-level descriptors, their OWL mapping does not provide real-world semantics, which can be achieved through high-level descriptors only. The MPEG-7 descriptors provide metadata and technical characteristics to be processed by computers, so their structured definition does not contribute to the semantic enrichment of the corresponding multimedia resources. To demonstrate this, take a closer look at a code fragment of the Core Ontology for Multimedia (COMM):[5]

```
<owl:Class rdf:about="#cbac-coefficient-14">
  <rdfs:comment rdf:datatype="&xsd;string"
>Corrresponds to the &quot;CbACCoeff14&#8221;
element of the &quot;ColorLayoutType&quot;
(part 3, page 45)</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#cbac-crac-
coefficient-14-descriptor-parameter"/>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
```

[3] http://rhizomik.net/ontologies/2017/05/Mpeg7-2001.owl

[4] https://schema.org

[5] http://multimedia.semanticweb.org/COMM/visual.owl

```
    <owl:Class rdf:about="&p1;unsigned-5-
vector-dim-14"/>
  </rdfs:subClassOf>
</owl:Class>
```

This part of the ontology is related to the color layout descriptor (CLD) of MPEG-7, which is used for capturing the spatial distribution of colors in images. To compute the CLD, RGB images are typically converted to the YCbCr color space, partitioned into 8×8 subimages, after which the dominant color of each subimage is calculated. Applying the discrete cosine transform (DCT) of the 8×8 dominant color matrix to the luminance (Y), the blue chrominance (Cb), and the red chrominance (Cr) results in three sets of 64 signal amplitudes, i.e., DCT coefficients DCTY, DCTCb, and DCTCr. The DCT coefficients can be grouped into two categories: those with a waveform mean value of 0 and those that have non-zero frequencies (DC and AC coefficients). Finally, the DCT coefficients are quantized and zig-zag scanned. This means that the `cbac-coefficient-14` listed above is suitable for the representation of blue chrominance AC coefficients, which can be used, among others, to filter video keyframes [14], however, they do not convey high-level semantics about the visual content. The `cbac-coefficient-14 class` is defined in COMM as a subclass of `cbac-crac-coefficient-14-descriptor-parameter` and `unsigned-5-vector-dim-14`, neither of which correspond to any real-world object class. Apparently, these coefficients would have been better defined as roles rather than concepts to enable them to hold the corresponding values. In this case, the OWL definitions do not advance the corresponding XSD vocabulary definitions with richer semantics, due to the previous modeling issues and the limited use of mathematical constructors in the implementation language.

Beyond the aforementioned OWL mappings of MPEG-7 that suffer from design issues, there is a more advanced MPEG-7 ontology, which does not inherit conceptual ambiguity issues from the standard and has been implemented in OWL 2.[6] This ontology has been grounded using a description logic formalism, covers the entire range of concepts and properties of MPEG-7 with property domains and ranges, and complex role inclusion axioms. Also, it captures correlations between properties.

*Criteria*

B1. The ontology to be used correctly conceptualizes the terms related to the scene and has a correct taxonomical structure.
B2. The ontology is axiomatized in a way that it can be used for reasoning.
B3. The ontology provides rich semantics for the concepts and/or events.

[6] http://mpeg7.org

### C. Specificity

Video scene representation employs not only domain ontologies, but also upper ontologies, application ontologies, commonsense ontologies, and core reference ontologies. For example, the Large Scale Concept Ontology for Multimedia (LSCOM) collects high-level concepts commonly depicted in videos (based on the comprehensive TRECVID dataset), however, many of the concepts are too general for precise high-level video scene descriptions. Also, video contents are not limited to concepts, and there are no events defined in LSCOM. The Linked Movie Database[7] is too specific, and can be used only for categorizing Hollywood movies, and even for this intended application it is not comprehensive enough.

The four fundamental ontologies that can be employed in video representation, and are imported by several higher-level video ontologies, are the SWRL Temporal Ontology,[8] the Event Ontology,[9] the Timeline Ontology,[10] and the Multitrack Ontology.[11]

There are many common terms that are defined by multiple ontologies (which is discouraged according to Semantic Web best practices [15]), sometimes with a slightly different name. These have to be assessed, and it has to be determined whether the represented concept or role corresponds to the same real-world entity or property. This should not be confused with those terms that are similar, but have been defined for different application scenarios, such as `dc:creator` and `foaf:maker`.[12]

*Criteria*

C1. The ontology clearly falls into one of the standard ontology categories.
C2. The ontology terms are not overly generic.
C3. Specific ontology terms are used from a highly specialized domain ontology or application ontology.
C4. The ontology terms used for annotation are defined by only one ontology or dataset. If there are similar or identical definitions available in multiple ontologies or datasets, the choice is determined by the following precedence order: 1) standard, 2) standard-aligned, 3) de facto standard, 4) proprietary.

### D. DL Expressivity

A common issue with multimedia ontologies is the lack of formal grounding, which is crucial not only for capturing the intended semantics, but also to reach high levels of, or maximize, reasoning potential. For example, the Visual De-

[7] http://www.linkedmdb.org
[8] http://swrl.stanford.edu/ontologies/built-ins/3.3/temporal.owl
[9] http://purl.org/NET/c4dm/event.owl#
[10] http://purl.org/NET/c4dm/timeline.owl#
[11] http://purl.org/ontology/studio/multitrack
[12] For creators described using a string literal, and without domain and range, `dc:creator` should be used, while `foaf:maker` is ideal for those creators who are identified by a URI.

scriptor Ontology (VDO),[13] which was published as an "ontology for multimedia reasoning" [16] has a very low DL expressivity (corresponds to $\mathcal{AL}$). This prevents capturing the correlation of classes and properties. In fact, VDO has fundamental problems with its concept definitions. For example, `colorSpace` is defined as an object property using the class `ColorSpaceDescriptor` as the range:

```
<owl:ObjectProperty
rdf:about="&VDO;colorSpace">
  <a:comment></a:comment>
  <a:range
rdf:resource="&VDO;ColorSpaceDescriptor"/>
  <a:subPropertyOf
rdf:resource="&VDO;DEFAULT_ROOT_RELATION"/>
  <a:domain
rdf:resource="&VDO;DominantColorDescriptor"/>
</owl:ObjectProperty>
```

Depending on the granularity of the ontology, color space could be defined as a concept instantiated with individuals, or a datatype property with all the permissible string values enumerated.[14] In VDO, neither of these is the case, and `colorSpace` is an object property, despite that it does not define a relation between classes or individuals. Moreover, there is no formal definition provided in VDO about the color spaces defined in the MPEG-7 standard the ontology is based on. Without rich semantics, no simple statements can be inferred, let alone complex statements, therefore VDO has a very limited potential in multimedia reasoning.

While one might argue that many ontologies have a low expressivity by design (in order to be lightweight and computationally cheap to reason over), in most cases low expressivity is the result of limiting the ontology to a taxonomical structure, which prevents advanced reasoning altogether.

*Criteria*

D1. The ontology is formally grounded.
D2. The ontology exploits all the mathematical constructors needed to formally describe constraints, complex roles, and correlations, rather than providing a class hierarchy and roles only.
D3. The ontology is as lightweight as possible.
D4. The ontology is underpinned by a decidable formalism.

### *E. Standards Alignment*

While international standards should be preferred over proprietary implementations, even ISO-standard-based ontologies are most often exposed through a nonstandard namespace URI, and standards alignment is often partial only.

General video metadata, such as title and language, can be represented using Dublin Core (ISO 15836-2009).[15] Low-level image, audio, and video descriptors can be annotated using the aforementioned MPEG-7 (ISO/IEC 15938).[16]

The most common de facto standards used in structured video annotations are W3C's Ontology for Media Resources,[17] DBpedia,[18] and the aforementioned Schema.org.

*Criteria*

E1. The ontology defines terms according to the corresponding standard specification and schema, and does not redefine them if an official ontology file is available.
E2. Standardized terms are used via the standard or, if this is not available, the de facto standard namespace URL.
E3. The ontology from which standardized terms are used covers the entire vocabulary of the standard with all datatypes and constraints adequately defined.

### *F. Namespace and Documentation Stability*

Many of the multimedia ontologies mentioned in the literature do not have a reliable namespace, making video annotations obsolete if the namespace becomes unavailable. A best practice to prevent this is to use a permanent URL, such as PURL,[19] which corresponds to a pointer that can be changed if the ontology file is moved. Another issue regarding ontology namespaces is that many of the namespace URLs are symbolic URLs only.

*Criteria*

F1. The namespace URL of the ontology to be used is preferably an actual web address (not a symbolic URL) and by using content negotiation, it
    a. serves the machine-readable ontology file (RDFS or OWL) to semantic agents, and
    b. serves a human-readable description of the ontology to web browsers (HTML5).
F2. The ontology namespace URL is a permanent URL.
F3. The human-readable content behind the URL is a comprehensive and up-to-date documentation of the ontology that reveals the intended implementation for each ontology term.

### *G. Spatiotemporal Annotation Support*

Although the mathematical constructors available in OWL 2 are not exploited in most multimedia ontologies, and they can express not only 2D, but also 3D information [17], vid-

[13] https://github.com/gatemezing/MMOntologies/blob/master/ACEMEDIA/acemedia-visual-descriptor-ontology-v09.rdfs.owl
[14] In MPEG-7, the following color spaces are supported: RGB, YCbCr, HSV, HMMD, and Monochrome. Linear transformation matrix with reference to RGB is also allowed.
[15] https://www.iso.org/standard/52142.html
[16] https://www.iso.org/standard/34230.html
[17] https://www.w3.org/TR/mediaont-10/
[18] http://wiki.dbpedia.org/
[19] https://archive.org/services/purl/

eo events require an even higher expressivity than what is supported by $\mathcal{SROIQ}^{(\mathcal{D})}$, the description logic underpinning OWL 2. Rule-based mechanisms, such as SWRL rules, are proven efficient in expressing video events [18], however, they often break decidability. Another option to push the expressivity boundaries is to employ formal grounding with spatial and temporal description logics, although many of these are not decidable either [19].

Spatial description logics vary greatly in terms of expressivity, and not all support qualitative spatial representations, which can address different aspects of space, including topology, orientation, shape, size, and distance. Some spatial description logics implement a Region Connection Calculus, such as RCC8 (see $\mathcal{ALC}(\mathcal{D}_{RCC8})$, for example [20]), while others, such as $\mathcal{ALC}(\mathcal{CDC})$, employ the Cardinal Direction Calculus (CDC) [21].

Temporal description logics also vary greatly, because some feature datatypes for time points, others for time intervals or sets of time intervals. Temporal description logics, such as $\mathcal{TL}$-$\mathcal{F}$ and $\mathcal{T}$-$\mathcal{ALC}$, are suitable for the formal representation of video actions and video event recognition via reasoning [22, 23].

*Criteria*

G1. Spatial annotations employ a formalism that supports qualitative spatial representation and reasoning.
G2. Temporal annotations use a formalism that allows both point-based and interval-based annotations.
G3. Spatiotemporal annotations employ a formalism that supports not only still regions, but also moving regions.
G4. Not only visual, but also audio descriptors are available to support video understanding via information fusion.
G5. If the description of a video scene requires spatiotemporal annotation, the formalism underlying the implemented ontology or ontologies is decidable, unless this would limit the semantics of the annotation.

### *H. Annotation Support for Uncertainty*

Video contents are inherently ambiguous. Fuzzy description logics can be used to express the certainty of the depiction of concepts [24], events, and video scenes [25]. This can be achieved by enabling normalized certainty degree values assigned to objects of fuzzy concepts.

*Criteria*

H1. The ontology is grounded in a formalism that supports fuzzy concept and fuzzy role axioms, and defines the associated semantics and interpretation.
H2. The formalism behind the fuzzy ontology is decidable.
H3. The core TBox axioms that represent background knowledge are formally grounded in a standard description logic.

## IV. Experimental Case Study

To evaluate the efficiency of the proposed approach, ontologies have been assessed, selected, and implemented for the spatiotemporal annotation of 10 video scenes, one of which is briefly presented here.

The iconic scene of the movie “Life of Pi” has been annotated with the regions of interest depicting Pi Patel and the tiger, Richard Parker (see Fig. 1).

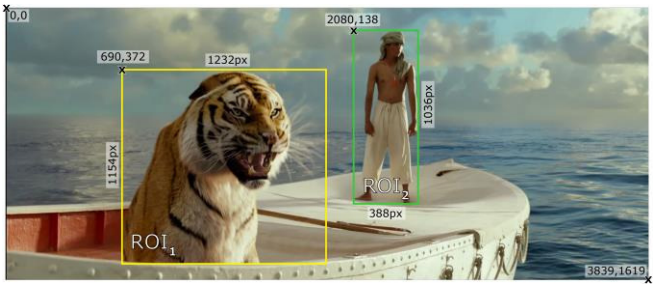


Fig. 1 Regions of interest coordinates and dimensions in a 4K Blu-Ray video scene. Movie scene by 20th Century Fox [26]

How the most suitable vocabularies and ontologies have been selected is demonstrated here via concepts related to this scene. Searching for vocabularies and ontologies that contain the corresponding terms is not adequate, because the ad-hoc selection of vocabularies and ontologies will not give satisfactory results, even if the selection is limited to high-quality structured data resources that have been checked for consistency. The Linked Open Vocabularies (LOV)[20] catalogue is maintained to help determine which vocabularies and ontologies to use for formal descriptions. Even though the list of rigorous criteria to meet before a vocabulary or ontology will be listed on LOV assures design quality [27], it does no guarantee that the best vocabulary will be selected for a particular scenario. For example, when searching for the term “video,” the LOV website suggests OpenGraph in the first, the Library extension of Schema.org in the second, and the NEPOMUK File Ontology in the third place. Among these, OpenGraph supports a URL to a video file without any semantics whatsoever, while the other two ontologies have not even been available at the time of writing (404 Not found).

Therefore, the proposed approach complements automated assessment with human judgment. Table I shows a comparison of three ontologies from the literature for representing the low-level video features of video scenes, namely the aforementioned VDO, COMM, and the only formally grounded MPEG-7 ontology, using the proposed approach, upon which the MPEG-7 Ontology has been selected.

[20] http://lov.okfn.org/dataset/lov/

TABLE I.
COMPARING ONTOLOGIES FOR REPRESENTING VIDEO PROPERTIES

| Criterion | VDO | COMM | MPEG-7 |
|---|---|---|---|
| A1 | – | – | Partially |
| A2 | Priority 2 | Priority 2 | Priority 1 |
| B1 | – | – | Partially |
| B2 | – | – | Partially |
| B3 | – | – | Partially |
| C1 | – | – | + |
| C2 | + | + | + |
| C3 | – | – | – |
| C4 | Priority 2 | Priority 2 | Priority 1 |
| D1 | – | – | + |
| D2 | – | – | + |
| D3 | + | + | + |
| D4 | + | + | + |
| E1 | – | – | + |
| E2 | – | – | + |
| E3 | – | – | + |
| F1 | – | – | + |
| F2 | – | – | + |
| F3 | – | – | + |
| G1 | – | – | + |
| G2 | – | – | – |
| G3 | + | + | + |
| G4 | + | + | + |
| G5 | + | + | + |
| H1 | – | – | – |
| H2 | N/A | N/A | N/A |
| H3 | N/A | N/A | N/A |

By using the criteria of the proposed approach for other video scene aspects, further ontologies and datasets have been selected for the video scene representation, including DBpedia, Schema.org, VidOnt, and the SWRL Temporal Ontology. For datatype definitions, the XML Schema vocabulary has been used to maximize interoperability. By declaring the corresponding namespaces, the background knowledge has been formalized as follows:

```
@prefix dbpedia:
<http://dbpedia.org/resource/> .
@prefix mpeg-7: <http://mpeg7.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-
rdf-syntax-ns#> .
@prefix schema: <http://schema.org/> .
@prefix vidont: <http://vidont.org/> .

dbpedia:Life_of_Pi_(film) a schema:Movie ;
vidont:filmAdaptationOf dbpedia:Life_of_Pi ;
mpeg-7:Video .
dbpedia:Suraj_Sharma a schema:Actor .
vidont:PiPatel a vidont:MovieCharacter ;
vidont:portrayedBy dbpedia:Suraj_Sharma ;
vidont:characterFrom
dbpedia:Life_of_Pi_(film) .
vidont:RichardParker a vidont:MovieCharacter
; vidont:portrayedBy dbpedia:Bengal_tiger ;
vidont:characterFrom
dbpedia:Life_of_Pi_(film) .
```

In this case study, the scene description utilized the previous individuals and highly specific concepts via spatiotemporal annotation and moving regions as follows:

```
@prefix mpeg-7: <http://mpeg7.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-
rdf-syntax-ns#> .
@prefix temporal:
<http://swrl.stanford.edu/ontologies/built-
ins/3.3/temporal.owl> .
@prefix vidont: <http://vidont.org/> .
@prefix xsd:
<http://www.w3.org/2001/XMLSchema#> .

<http://example.com/lifeofpi.mp4#t=1:14:38,1:
14:41> a vidont:Scene ;
vidont:sceneFrom dbpedia:Life_of_Pi_(film) ;
temporal:hasStartTime "01:14:38"^^xsd:time ;
temporal:duration "PT00M03S"^^xsd:duration ;
temporal:hasFinishTime "01:14:41"^^xsd:time ;
vidont:depicts vidont:PiPatel ,
vidont:RichardParker .
<http://example.com/lifeofpi.mp4#t=1:14:40&xy
wh=690,372,1232,1154> a mpeg-7:MovingRegion ;
vidont:depicts vidont:RichardParker ;
vidont:inFrontOf vidont:PiPatel ; vidont:isIn
dbpedia:Lifeboat_(shipboard).
<http://example.com/lifeofpi.mp4#t=smpte:01:
14:40:03> mpeg-7:width
"3840"^^xsd:positiveInteger ;
mpeg-7:height "1620"^^xsd:positiveInteger .
<http://example.com/lifeofpi.mp4#t=1:14:40&xy
wh=2080,138,1036,388> a mpeg-7:MovingRegion ;
vidont:depicts dbpedia:PiPatel ; vidont:isIn
dbpedia:Lifeboat_(shipboard) .
```

Note that the spatiotemporal segmentation employs not only the SWRL Temporal Ontology, but also Media Fragment URI 1.0 identifiers,[21] where the URL identifies the minimum bounding boxes of the regions of interests using the top left corner coordinates and the dimensions, so that the media segments are globally unique and dereferencable.

Based on the previous video scene description, reasoners can infer new, useful information by utilizing axioms of the vocabularies and ontologies selected using the proposed approach. For example, based on the statement that `<http://example.com/lifeofpi.mp4#t=1:14:40&xywh=690,372,1232,1154>` is a moving region, and the axiom of the MPEG-7 Ontology that defines moving regions as subclasses of spatiotemporal video segments, it can be inferred using concept subsumption reasoning, according to which concept $D$ subsumes concept $C$ with reference to knowledge base $\mathcal{K}$ if and only if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ for all interpretations $\mathcal{I}$ (that are models of knowledge base $\mathcal{K}$), that `<http://example.com/lifeofpi.mp4#t=1:14:40&xy`

[21] https://www.w3.org/TR/media-frags/

wh=690,372,1232,1154> is a spatiotemporal decomposition, which was not explicitly stated. This could not have been deducted using terms from VDO or COMM, because they do not define moving regions at all, let alone doing so in a taxonomical structure.

More complex information can be automatically inferred using the RDFS entailment rules,[22] the Ter Horst reasoning rules [28], and the OWL reasoning rules.[23] For example, based on the axiom of the MPEG-7 Ontology that defines Frame as the domain of the height property and the height declaration for the screenshot of the Life of Pi video file, i.e.,

```
mpeg-7:height rdfs:domain mpeg-7:Frame .
<http://example.com/lifeofpi.mp4#t=
smpte:01:14:40:03> mpeg-7:height
"1620"^^xsd:positiveInteger .
```

and using the OWL 2 reasoning rules for axioms about properties, it can be automatically inferred that this temporal video segment corresponds to a video frame, formally,

```
<http://example.com/lifeofpi.mp4#t=smpte:01:
14:40:03> a mpeg-7:Frame .
```

which was not explicitly stated. Considering that these concepts and roles are not defined in the other MPEG-7-aligned ontologies, and therefore their reasoning potential would be inadequate for this scenario, it can be confirmed that the MPEG-7 Ontology suggested by the presented approach is the best choice.

## V. Conclusion

Based on the comprehensive review of the state of the art, an approach has been proposed to determine the list of DL-based multimedia ontologies to be used for the annotation of video scenes while taking into account all major aspects of ontology implementation. Some of these correspond to core requirements all selected ontologies have to meet, such as high-quality conceptualization and having a stable namespace. For others, such as spatiotemporal annotation support, it may be adequate if at least one of the ontologies qualifies. Some video scenes do not require fuzzy concepts. The integration of multimedia ontologies using the proposed approach can not only guide through selecting the most appropriate ontologies to obtain the formalism needed to describe a particular video scene, but also ensures standards alignment, avoids overgeneralization, eliminates overlapping definitions, and optimizes reasoning complexity.

## References


[1] L. F. Sikos, "Ontology-based structured video annotation for content-based video retrieval via spatiotemporal reasoning," In *Bridging the Semantic Gap in Image and Video Analysis*. *Intelligent Systems Reference Library*. H. Kwaśnicka and L. C. Jain, Eds., Cham: Springer, 2017

[2] A. Isaac and R. Troncy, "Designing and using an audio-visual description core ontology," presented at the Workshop on Core Ontologies in Ontology Engineering, Northamptonshire, October 8, 2004.

[3] L. F. Sikos, "VidOnt: a core reference ontology for reasoning over video," *J. Inf. Telecommun*, 2017.

[4] L. F. Sikos, "A novel approach to multimedia ontology engineering for automated reasoning over audiovisual LOD datasets," in *Intelligent information and database systems*, N. T. Nguyễn, B. Trawiński, H. Fujita, and T.-P. Hong, Eds. Heidelberg: Springer, 2016, pp. 3–12. doi: 10.1007/978-3-662-49381-6_1

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Conf. Proc. 1999 IEEE Int. Conf. Comput. Vis.*, pp. 1150–1157. doi: 10.1109/ICCV.1999.790410

[6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Conf. Proc. 2006 Eur. Conf. Comput. Vis.*, pp. 428–441. doi: 10.1007/11744047_33

[7] J.-H. Lee, G.-G. Lee, and W.-Y. Kim, "Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder," *IEEE Trans. Consumer Electronics*, vol. 49, pp. 742–749, 2003. doi: 10.1109/TCE.2003.1233813

[8] M. Bertini, A. Del Bimbo, and W. Nunziati, "Video clip matching using MPEG-7 descriptors and edit distance." In *Image and Video Retrieval*, H. Sundaram, M. Naphade, J. R. Smith, and Y. Rui, Eds., Heidelberg: Springer, 2006, pp. 133–142.

[9] L. F. Sikos, *Description Logics in Multimedia Reasoning*. Cham: Springer, 2017. doi: 10.1007/978-3-319-54066-5

[10] S. Blöhdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, and M. Strintzis, "Semantic annotation of images and videos for multimedia analysis," in *The Semantic Web: research and applications*, A. Gómez-Pérez and J. Euzenat, Eds. Heidelberg: Springer, 2005, pp. 592–607. doi: 10.1007/11431053_40

[11] L. F. Sikos and D. M. W. Powers, "Knowledge-driven video information retrieval with LOD: from semi-structured to structured video metadata," in *Proc. 8th Workshop on Exploiting Semantic Annotations in Information Retrieval*, New York, 2015, pp. 35–37. doi: 10.1145/2810133.2810141

[12] M. Horvat, N. Bogunović, and K. Ćosić, "STIMONT: a core ontology for multimedia stimuli description," *Multimed. Tools Appl.*, vol. 73, pp. 1103–1127, 2014. doi: 10.1007/s11042-013-1624-4

[13] L. F. Sikos, "RDF-powered semantic video annotation tools with concept mapping to Linked Data for next-generation video indexing: a comprehensive review. *Multim. Tools Appl.*, vol. 76, pp. 14437–14460, 2016. doi: 10.1007/s11042-016-3705-7

[14] M. Abdel-Mottaleb, N. Dimitrova, L. Agnihotri, S. Dagtas, S. Jeannin, S. Krishnamachari, T. McGee, and G. Vaithilingam, "MPEG 7: a content description standard beyond compression," in *Proc. 42nd IEEE Midwest Symp. Circuits Syst.*, New York, 1999, pp. 770–777. doi: 10.1109/MWSCAS.1999.867750

[15] E. Simperl, "Reusing ontologies on the Semantic Web: a feasibility study," *Data Knowl. Eng.*, vol. 68, pp. 905–925. doi: 10.1016/j.datak.2009.02.002

[16] N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias, "A visual descriptor ontology for multimedia reasoning," presented at the 6th International Workshop on Image Analysis for Multimedia Interactive Services, Montreux, April 13–15, 2005.

[17] L. F. Sikos, "A novel ontology for 3D semantics: from ontology-based 3D object indexing to content-based video retrieval. *Int. J. Metadata, Semant. Ontol.*, 2017

[18] M. Y. K. Tani, A. Lablack, A. Ghomari, and I. M. Bilasco, "Events detection using a video surveillance ontology and a rule-based approach," in *Computer Vision – ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer, 2014, pp. 299–308. doi: 10.1007/978-3-319-16181-5_21

[19] Sikos, L. F, "Spatiotemporal Reasoning for Complex Video Event Recognition in Content-Based Video Retrieval." In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*. *Advances in Intelligent Systems and Computing*,


[22] https://www.w3.org/TR/2004/REC-rdf-mt-20040210/#RDFSRules

[23] https://www.w3.org/TR/owl2-profiles/#Reasoning_in_OWL_2_RL_and_RDF_Graphs_using_Rules

vol. 639. A. Hassanien, K. Shaalan, T. Gaber, and M. F. Tolba, Eds., Cham: Springer, 2017, pp. 704–713. doi: 10.1007/978-3-319-64861-3_66

[20] K.-S. Na, H. Kong, and M. Cho, "Multimedia information retrieval based on spatiotemporal relationships using description logics for the Semantic Web," Int. J. Intell. Syst., vol. 21, pp. 679–692. doi: 10.1002/int.20153

[21] M. Cristani and N. Gabrielli, "Practical issues of description logics for spatial reasoning," in Proc. 2009 AAAI Spring Symp., Menlo Park, CA, 2009, pp. 5–10.

[22] L. Bai, S. Lao, W. Zhang, G. J. F. Jones, and A. F. Smeaton, "Video semantic content analysis framework based on ontology combined MPEG-7," in "Adaptive multimedia retrieval: retrieval, user, and semantics," N. Boujemaa, M. Detyniecki, and A. Nürnberger, Eds. Heidelberg: Springer, 2008, pp. 237–250. doi: 10.1007/978-3-540-79860-6_19

[23] W. Liu, W. Xu, D. Wang, Z. Liu, X. Zhang, "A temporal description logic for reasoning about action in event," Inf. Technol. J., vol. 11, pp. 1211–1218. doi: 10.3923/itj.2012.1211.1218

[24] N. Elleuch, M. Zarka, A. B. Ammar, and A. M. Alimi, "A fuzzy ontology-based framework for reasoning in visual video content analysis and indexing," in Proc. 11th Int. Workshop Multim. Data Min., New York, 2011, Article No. 1. doi: 10.1145/2237827.2237828

[25] E. Elbaşi, "Fuzzy logic-based scenario recognition from video sequences," J. Appl. Res. Technol., vol. 11, pp. 702–707. doi: 10.1016/S1665-6423(13)71578-5

[26] Netter, G., Lee, A., Womark, D. (Producers) and Lee, A. (Director), Life of Pi, 20th Century Fox, USA, 2012 [Motion picture, 2016 Ultra HD Blu-ray release].

[27] Vandenbussche, P.-Y., Atemezing, G. A., Poveda, M., and Vatant, B., "Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web," Semantic Web, vol. 8, pp. 437–452. doi: 10.3233/SW-160213

[28] Ter Horst, H. J., "Completeness, Decidability and Complexity of Entailment for RDF Schema and a Semantic Extension Involving the OWL Vocabulary," J. Web Semant. Sci. Serv. Agents World Wide Web, vol. 3, pp. 79–115. doi: 10.1016/j.websem.2005.06.001