What Are You Trying to Say? Format-Independent Semantic-Aware Streaming and Delivery

Joseph Thomas-Kerr[1], Ian Burnett[2] and Christian Ritz[3]
[1,3]*University of Wollongong*
[2]*Royal Melbourne Institute of Technology*
Australia

1. Introduction

> *"**[Elizabeth Bennett] looked at her father to entreat his interference**, lest Mary should be singing all night. He took the hint, and, when Mary had finished her second song, said aloud, 'That will do extremely well, child. **You have delighted us long enough.**' "* —
>
> Pride and Prejudice, 1813, Jane Austen (1813)

Users automatically associate many layers of meaning with the media content they consume, yet computers have barely begun to scrape the surface of this information. For example, consider the passage above. The subtle exchange of glances between Elizabeth and her father would be readily apparent to most human observers, but it is unlikely that a computer processing a video of the scene would be able to recognise their meaning. Furthermore, while the double-entendre in Mr Bennett's remark would be clear to most human listeners, algorithmic recoginition of this or other modes of speech are in their infancy (Paleari & Huet, 2008).

Other research communities are developing means to communicate such semantic information (whether computed or manually generated) in ways that are able to transcend the original context of the information.This work—originating from Knowledge Representation, but more popularly known as the Semantic Web—has provided languages such as the Resource Description Framework (RDF) (Beckett, 2004) and Web Ontology Language (OWL) (Dean & Schreiber, 2004) which can be used to express concepts in such a way that "this picture has many buildings" may also imply that "it is a cityscape", and "it contains man-made objects."

Recent multimedia coding formats developed by MPEG and ITU-T such as Scalable Video Coding (SVC) (ISO/IEC, 2007) and Scalable-to-Lossless Coding (SLS) (ISO/IEC, 2004a) offer the ability to dynamically adapt their bitrate to changing conditions. Current systems perform this adaptation on the basis of static channel parameters such as terminal and network capabilities (Timmerer et al., 2006) or dynamic estimation of channel capacity (Chou, 2006). There are, in fact, numerous examples of using *content semantics* to identify the best way to adapt content to dynamic conditions: Section 2 describes this in further detail. However, while others have proposed specific semantics to be used in the delivery process, there exists no generic system for connecting arbitrary semantics to the adaptation/delivery process.

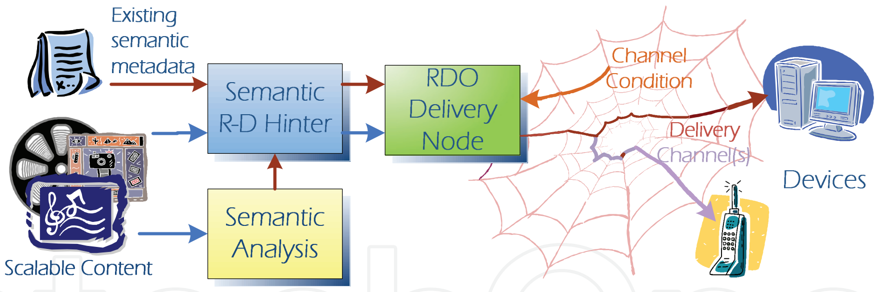


Fig. 1. A framework for semantic-aware multimedia delivery

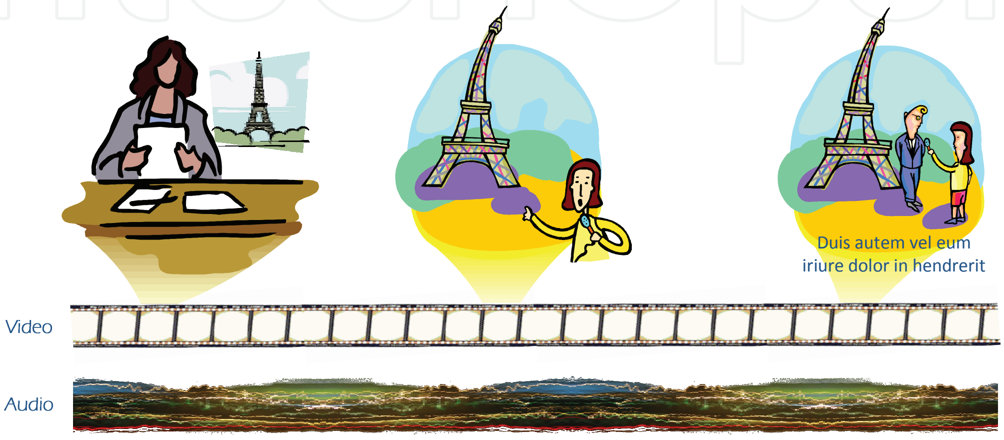


Fig. 2. News reports have a regular structure

This chapter proposes and demonstrates just such a system, as shown in Figure 1. An overview of the system was presented in Thomas-Kerr et al. (2009); the present work greatly expands upon the complexities of its concept and design. Combining semantics and multimedia delivery in a generic fashion is, unsurprisingly, a task that draws on numerous disparate fields. When possible, sufficient detail has been provided to appreciate the background concepts, however, the reader is referred to the relevant citations for further information.

2. Semantics in the delivery process

A typical news report (Figure 2) provides a good example of how content semantics could be useful in multimedia delivery. News reports often have a fairly consistent structure, beginning with a studio introduction, then footage of the event (often with commentary overlaid on top of audio from the event), using subtitles if subjects are speaking in a foreign language, and sometimes concluding with further studio footage. As a report proceeds through these various stages, the relative importance of the audio and video varies. For example, in the studio introduction, virtually all of the semantic content of the presentation is carried in the audio. On a low-bandwidth (e.g. mobile) channel, reduction of the frame-rate in this region would have little impact on the transmission of the content semantics. When the report cuts to on-site footage, a much greater proportion of the semantic content is carried by the visuals, though the amount would vary from one report to another. If subtitles are overlaid on the video, virtually all of the semantic content is conveyed by the video, and bits spent on audio in this section will contribute much less to the successful delivery of the semantics.

As a second example, instead of comparing relative semantic importance along the modal dimension (as is the case above), similar comparisons could be made on the temporal axis. Here, segments of the news report would be annotated with an indication of the relative importance of the segment to the story as a whole. Users could receive a short “digest”, the full story, or something in between. This approach could also work for coverage of sporting events.

Numerous other types of semantic metadata have been identified which can assist in delivery optimisation: Bertini et al. (2006), Xu et al. (2006), and Baba et al. (2004) all argue that applying the same adaptation operation to different parts of a multimedia presentation will have differing effects in the perceptual quality of the presentation as a whole.

More specifically, both Bertini et al. (2006) and Xu et al. (2006) propose adaptation on the basis of semantic classification of sporting events into categories such as Shot on Goal, Corner (for soccer), or Shot, Foul, Penalty (for Basketball), among others. User preferences are used to prioritize the categories, and this priority information is then used to guide the adaptation. That is, given a bandwidth constraint such that the full content can’t be delivered, the adaptation engine reduces the bit rate of lower priority sections before those with higher priority.

The semantic metadata used in the preceding examples can be considered as very high-level, and coarse-grained. That is, it identifies relatively large segments of content, using concepts with a high level of abstraction from the digital representation of the content. In the first case, heuristic methods are proposed to automatically classify content segments, with a precision of 83% to 96% Bertini et al. (2006).

Baba et al. (2004) propose adaptation of speech signals on the basis of a much lower-level semantic concept: sound volume. They argue that regions of (relative) silence within a speech signal carry no *semantic information*, and as such may be truncated during playback. In fact, this feature of speech[1] may be used to guide adaptation, allowing regions of silence to be constrained to a zero bit-rate (or as close as the scalable codec or synchronization scheme will allow) with no perceptible loss of fidelity.

Cranley & Murphy (2006) suggest further low-level semantics that may be used to optimize delivery. They use measures of the temporal and spatial complexity to trade-off frame rate with resolution for scalable codecs, to achieve a so-called Optimum Adaptation Trajectory.

The semantic-aware content delivery framework proposed in this chapter provides a way to incorporate these and other semantics into the delivery of multimedia. This is achieved in a way that is flexible enough to support the increasingly diverse range of formats, semantics, and networks that are used (or useful) for content delivery (Brightman, 2005). Before a detailed discussion of this framework, Section 3 (below) identifies a number of key features that are necessary for the framework to successfully address the challenges posed by this diversity. The proposed framework itself is then detailed in Section 4, along with an analysis of existing work that is able to fulfil some constituent parts. Section 5 describes subjective testing validating the approach, and Section 6 offers some concluding remarks.

3. Features

Multimedia semantics is an extremely diverse field. Similarly, multimedia delivery is categorised by an exponentially growing array of devices that access and process multimedia,

[1] or audio, although silence is less prevalent there

and an increasing number of formats in which multimedia content is encoded. Given this complexity, this section argues that sucessfully combining semantics and delivery requires a flexible approach, where the semantics and formats used are not hard-coded, but instead described declaratively as content metadata.

3.1 Format-independence

The present, exponential rate of growth in both multimedia devices (hardware and software) and content formats is increasing the difficulty of maintaining interoperability. To be effective in this environment, a semantic-aware delivery framework must support content that is encoded in any current, or future, format. As has been shown (Thomas-Kerr et al., 2008), for many multimedia operations, it is possible to abstract the format-specific details of any given codec into a data file (hardware and software is then format-independent). This greatly simplifies interoperability, since a new content format can be integrated into existing devices merely by dissemninating a file that describes its format-specific details. Crucially (given the exponential growth in the range and diversity of multimedia devices) no modification of hardware or software is necessary.

This argument also holds for the syntax in which semantic metadata is encoded; as discussed (Section 4.2.2 on page 12) there are many syntaxes used to encode the metadata needed for semantic-aware delivery. Further, as is the case for content formats, the framework must also cope with new metadata formats, as they are developed. In response to these observations, methods for adapting metadata syntax without requiring changes to software or hardware have been proposed (Thomas-Kerr et al., 2006) and are important to allow a semantic-aware delivery framework to be as widely applicable as possible.

3.2 Semantic-independence

The range of semantics that people associate with media content is effectively infinite. The examples cited in Section 2 therefore represent just a small sample of the possibilities for using semantics to guide multimedia delivery. As such, it is important that a semantic-aware delivery framework not be limited to using a small, defined set of concepts.

3.3 Multiple optimisation algorithms

As will be seen in Section 4.1, a considerable number of algorithms have been developed for optimising the Rate-Distortion (R-D) performance of multimedia delivery (Chakareski et al., 2004a; Chou, 2006; Cranley & Murphy, 2006; Eichhorn, 2006). These algorithms vary in their guarantees of tractability, complexity, and the range of metadata required as inputs to the process. As a result, different algorithms may be preferable in certain scenarios, and so flexibility in this regard is an important characteristic of a semantic-aware delivery framework.

3.4 Segmentation and association

The examples cited in Section 2 on page 2 differentiate the semantic importance of segments of content that have been segmented along numerous axes. The most straightforward is with the sporting analysis and speech sound-level concepts, where some *temporal* segments are more important than others. This is also the case in the news example, but here a distinction is also made along the *mode* axis: in some temporal segments the video has more semantic importance, in other segments it is the audio. Cranley et al. (2003) distinguish

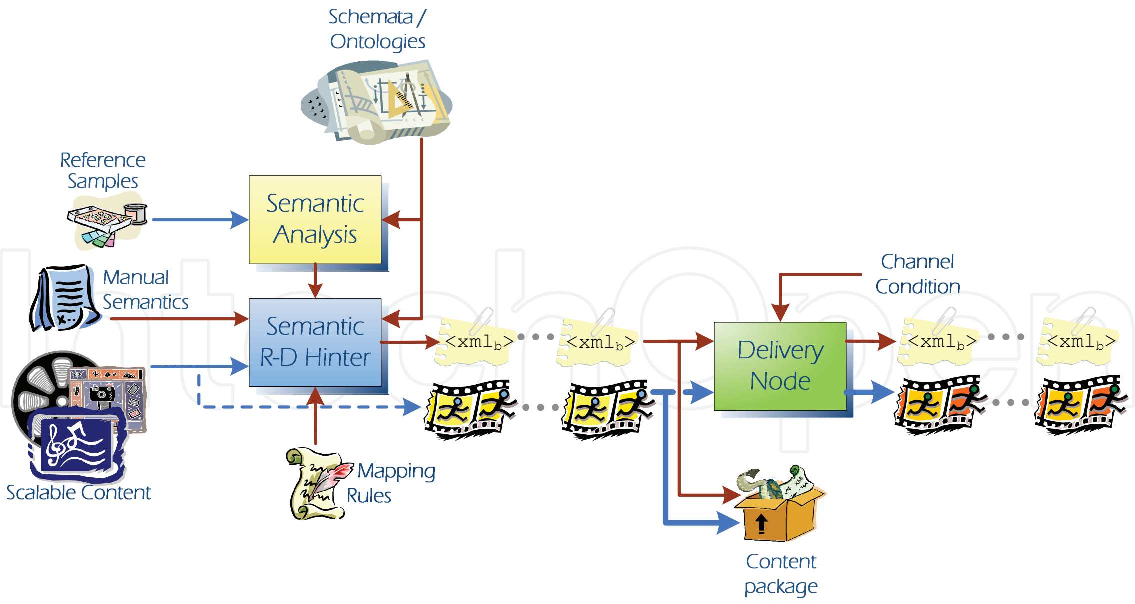


Fig. 3. An architecture for Mulimedia Delivery that incorporates content semantics

between semantic importance along the *temporal* and *spatial* axes. Although it has not been widely utilised, MPEG-4 (ISO/IEC, 2004b) generalises this concept still further by introducing other modes (text, graphics, and hybrid coding), and additionally provides the ability to arrange multiple audiovisual "objects" within a scene. In such a scenario, it may be highly advantageous to attach (time-varying) semantic importance to each of these modes and objects.

Clearly, the utility of a semantic delivery framework would depend considerably on it having the flexibility to segment content along all of these (and potentially other) dimensions. After segmentation, such a framework would need to be able to associate semantic and other metadata with these segments, in such a way that they can be input to an algorithm that makes the trade-offs described.

4. A framework for format-independent semantic-aware multimedia delivery

Figure 3 depicts the proposed architecture of a semantic-aware delivery framework. As proposed by Chakareski et al. (2004b), the Rate-Distortion Optimisation (RDO) process is split into two parts: generation of R-D metadata is performed offline by a *hinter*, minimising the amount of computation that must be done by the real-time *delivery node*. The present work extends this concept by proposing an architecture for the hinter that is format-independent, for the reasons outlined earlier in Section 3. Additionally, the hinter in Figure 3 provides for *Semantic Distortion* (see below, Section 4.2.2 on page 12) to be combined with the "classical" approach to distortion where decoded samples are compared to the samples that were originally encoded, using a measure such as (peak-)SNR, referred to as *Sample Distortion*.

4.1 Delivery node

With the static content analysis performed offline by a hinter, a delivery node (Figure 4 on the next page) is left only to decide whether and when to forward, drop or truncate (where applicable) each packet. That decision is made on the basis of some type of rate-distortion

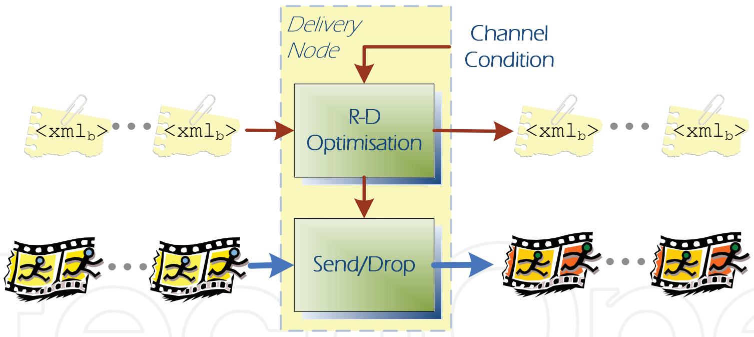


Fig. 4. A delivery node uses content hints to perform R-D optimisation

optimisation algorithm, which takes as its inputs feedback about the channel condition, and metadata from the semantic hinter. These elements are described in the following sections.

4.1.1 R-D optimisation

There are a number of rate-distortion optimisation algorithms. Different algorithms perform better in particular scenarios, and so this semantic-aware framework avoids prescribing one over another. Instead, the framework allows the most suitable algorithm(s) to be implemented on any given delivery node.

Chou (2006) proposes the use of classical optimisation of R-D performance $D(R)$ by minimising the Lagrangian $D + \lambda R$ for some λ. The formulation of distortion must consider the error probability-cost functions for each unit of data, as well as the interdependencies between Data Units, since descendent packets (e.g. any motion-compensated frame, or enhancement layers in SVC) generally cannot be decoded if their ancestors are not received.

Chakareski et al. (2004a) note that although the algorithm proposed by Chou is theoretically optimal and suitable for certain applications, it comes at the cost of significant computational complexity. Consequently, Chakareski proposes a low-complexity approximation of the lagrangian optimisation problem, by ignoring interdependencies between Data Units and instead assuming that distortion from packet loss on subsequent packets is additive.

Eichhorn (2006) suggests the opposite approximation: Chakareski ignores actual dependencies; Eichhorn ignores actual distortions, and asserts that dependency alone may be sufficient. Finally, Cranley & Murphy (2006) trade temporal resolution against spatial resolution and use subjective testing to arrive at a so-called Optimum Adaptation Trajectory.

4.1.2 Serialisation of hinter metadata

On the one hand, a binary syntax could be specified for hinter metadata, in order to maximise space efficiency over-the-wire[2]. However, this makes extension of the data set (as is likely inevitable as new optimisation techniques are developed) difficult to achieve without breaking existing implementations. For this and other reasons, most recent metadata uses XML rather than binary syntax, because of the ease with which it is processed and parsed, despite its inherent verbosity. As it turns out, it is possible to achieve most of the benefits of both, using the so-called Binary format for Metadata (BiM) (Niedermeier et al., 2002). BiM uses XML Schema to provide efficient binary encodings of XML data. This means that the R-D metadata can be created and processed as XML, but if it must be transmitted, BiM can achieve

[2] That is, when this metadata must be transmitted on-the-wire, which is only the case if the Delivery Node is remote from the content, for example if it is a gateway node.

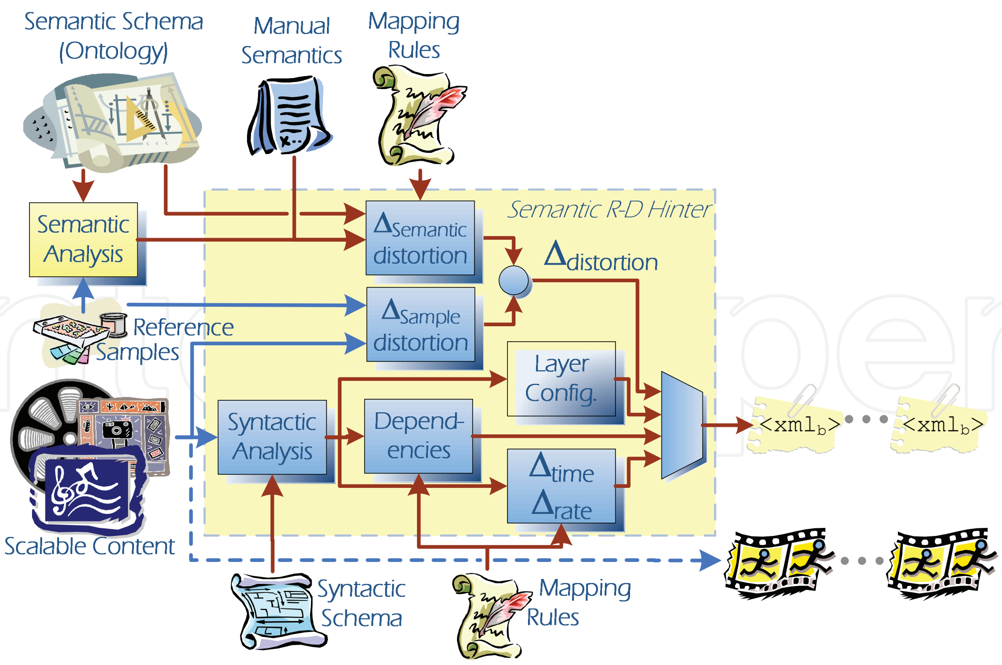


Fig. 5. The semantic hinter computes R-D metadata based on content syntax and semantics

transmission efficiencies close to those of a dedicated binary syntax. Furthermore, at the downstream node, the binary representation may be parsed directly, without decompression, avoiding any additional time complexity.

4.1.3 Summary

In deciding whether to forward or drop each packet as it is received, the Delivery Node utilises some sort of optimisation algorithm. Several of these were discussed above (Section 4.1.1). Depending on the algorithm chosen, different metadata is required from the R-D Hinter, although a common subset of data including Δ_{time}, Δ_{rate} and segmentation is required for the forward/drop routine. Otherwise, this metadata may contain the set of distortion increments ΔD_l, the Data Unit dependence graph, or Spatial Information (SI) and Temporal Information (TI) values[3]. This also points to the need for a negotiation process between the Delivery Node and the node holding the content to identify the desired optimisation algorithm based on the available metadata, although in some cases a node may be able to generate missing metadata on-the-fly (with the concordant time penalty).

If the Delivery Node is remote from the content and metadata, for example if it is a gateway node that spans two heterogeneous networks, then it may be desirable to minimise the bandwidth used by the R-D metadata, by utilising BiM (Niedermeier et al., 2002) to binarise the data. In this case, the Delivery Node would use a BiM parser that directly interprets the binarised metadata and passes the output data points directly to the RDO algorithm.

4.2 Semantic R-D hinter

The role of the hinter (Figure 5) is to prepare the metadata needed by the R-D Optimisation algorithm. This metadata can then be stored in a file (such as an ISO (ISO/IEC, 2005a) or Quicktime (Apple, 2001) container) for later use, or transmitted with the content to a local or remote delivery node. The hinter itself is composed by elements that analyse the semantics

[3] refer to Cranley & Murphy (2006) for a detailed discussion of these parameters.

and the syntax of the content. The former (semantic analysis) obtains the higher-level characteristics of the content which are typically not evident in the compressed domain, but must be identified from the original (reference) samples or entered manually. Section 4.2.2 on page 12 considers semantic analysis in greater detail. On the other hand, syntactic analysis (discussed in Section 4.2.1) extracts the interdependency, temporal and scalability metadata that are direct parameters of the compressed bitstream.

4.2.1 Syntactic analysis

Syntactic analysis is the process by which the hinter exposes the syntactical elements of multimedia that are needed by a given RDO algorithm. This occurs in two stages. First, the underlying syntactic structure of the content must be exposed so as to provide access to the internal data fields. In this work, *binary schemata* (Thomas-Kerr et al., 2007) are used to achieve this functionality. Secondly, a *mapping* must be made from the arbitrary raw data fields exposed by a schema, to the specific concepts needed for RDO.

Binary Schemata

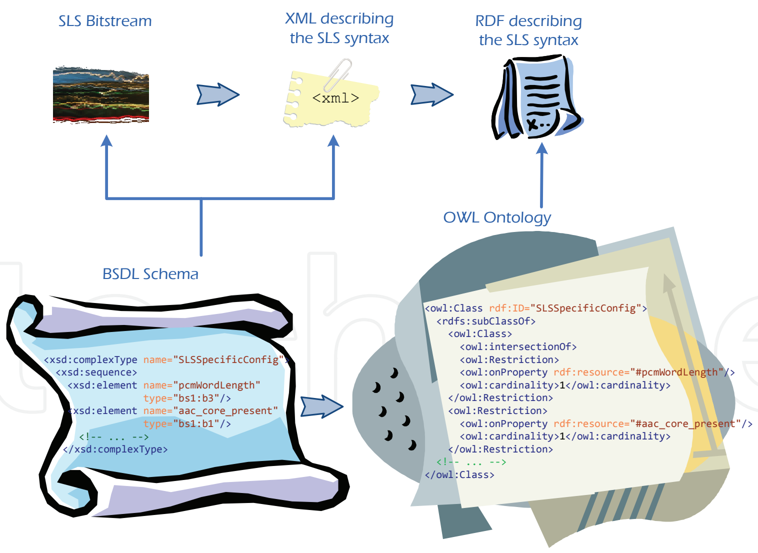
Recent coding formats utilise increasingly complex multi-layer structures to encode media in ever-fewer bits. As a result, identifying the timestamp, interdependencies or even byte-boundaries of an encoded Data Unit requires significant parsing. In most systems, this parsing is performed by format-specific software or hardware, that is, the format of the codec is “hard coded” into the parser. However, because the number of coding formats is large and growing, such a hard-coded approach makes it increasingly difficult to maintain interoperability with the available coded content.

An alternative is to use a reconfigurable or generic parser for syntactic analysis, where the specific syntax of individual codecs is stored in a schema *data file*. Support for additional formats may then be added via a new file, rather than new hardware or software. While there are numerous syntax description languages (such as the common EBNF (Klint et al., 2005)), only a few of which provide sufficient expressivity to function as a schema language for a generic parser (see Thomas-Kerr, Janneck, Mattavelli, Burnett & Ritz (2007)): BSDL and XFlavor (Hong & Eleftheriadis, 2002) (and a hybrid of the two—BFlavor (Neve et al., 2006)).

Any of these languages are suitable for syntax schemata in the model. Each provides an XML “view" of binary data which can be used to construct the rules required for further analysis. In XFlavor and BFlavor there is a level of indirection between the binary schema and the XML schema, whereas in BSDL they are one and the same: a BSDL Schema is an augmented XML Schema (Thompson et al., 2004).

Lehti & Fankhauser (2004) show that the object-based structure of XML Schemata (and the XML data they describe) means that it is possible to map from XML Schema complex and simple types (which directly or indirectly represent binary structures in the case of a BSDL Schema) to OWL classes and properties (respectively).

This approach is far from elegant, because XML Schemata describe syntax whereas OWL (Dean & Schreiber, 2004) and RDF (Beckett, 2004) describe semantics, and mixing the two in this way can lead to significant ambiguity. Nonetheless, it is useful, since combining it with one of the binary schemata languages described above allows binary data to be directly integrated with OWL/RDF-based data). This means that binary content may be processed and queried as if it were RDF triples. Figure 6 on the next page depicts an example of the



```
<xsd:complexType name="SLSSpecificConfig">
  <xsd:sequence>
    <xsd:element name="pcmWordLength"
                 type="bs1:b3"/>
    <xsd:element name="aac_core_present"
                 type="bs1:b1"/>
      <!-- ... -->
  </xsd:complexType>
```

```
<owl:Class rdf:ID="SLSSpecificConfig">
  <rdfs:subClassOf>
    <owl:Class>
      <owl:intersectionOf>
      <owl:Restriction>
      <owl:onProperty rdf:resource="#pcmWordLength"/>
      <owl:cardinality>1</owl:cardinality>
    </owl:Restriction>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#aac_core_present"/>
      <owl:cardinality>1</owl:cardinality>
    </owl:Restriction>
  <!-- ... -->
</owl:Class>
```

Fig. 6. A binary schema can be used to expose the structure and data of a bitstream as OWL/RDF

approach. A BSDL schema describes the binary structure of an SLS bitstream (the example shows an SLSSpecificConfig structure (ISO/IEC, 2005b)). At the same time, this BSDL schema describes the structure of an XML representation of the syntax of the binary bitstream. Lehti's technique is then used to map the BSDL/XML schema into OWL classes, and to map the XML metadata into RDF triples. This allows the binary structure of the SLS bitstream, along with the data it contains, to be reasoned on and combined with other OWL/RDF semantic metadata. [4]

Mapping Rules

The metadata exposed by using a binary schema will be specific to the format that the schema describes (e.g. SLS, Flash, SVC). In order to use this metadata in a semantic-aware delivery framework, it is necessary to be able to map from the format-specific structures exposed by the binary schema, to the set of format-independent metadata needed by the RDO algorithm being used. The list of metadata required will vary depending on the particular RDO algorithm being used, but will generally include items such as

- segmentation of the content into Data Units;
- decoding interdependencies between Data Units; and
- temporal relations between Data Units;

One such set of mapping rules may be used to describe the extraction of RDO metadata from SLS bitstreams, while another set of mappings describe the process for Flash, and a third for H.264/SVC (as shown – Figure 7).

[4] It should also be noted that BSDL allows the bitstream to be described at whatever level is required, in order to avoid unnecessary verbosity. That is, if the reasoning to be performed requires only that the binary data be split into frames, then the BSDL Schema may be written in such a way that it emits a single XML element per frame. On the other hand, if certain fields within a frame are necessary for reasoning (such as a timestamp, sample rate, etc.) then the schema is able to expose these fields without showing the entire detail of the inside of a frame. See (Thomas-Kerr et al., 2007) for more information.

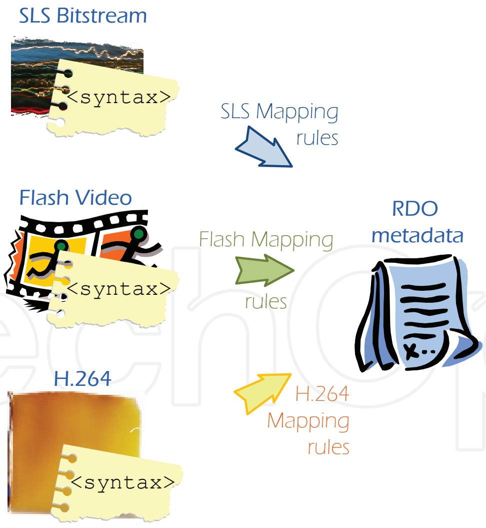


Fig. 7. Mapping rules can be used to translate from format-specific structures into the format-independent metadata needed for a semantic-aware RDO delivery framework

```
<xsd:element name="pic_parameter_set_rbsp">
  <xsd:annotation><xsd:appinfo>
    <bs2x:variable name="pps" bs2x:position="pic_parameter_set_id + 1"/>
  </xsd:appinfo></xsd:annotation>
  <xsd:complexType>
    <xsd:sequence>
      <xsd:element name="pic_parameter_set_id" type="bs1:unsignedExpGolomb"/>
      <xsd:element name="seq_parameter_set_id" type="bs1:unsignedExpGolomb"/>
      <!−− ... −−>
    </xsd:sequence>
    <xsd:attribute ref="rdo:ancestors" bs0:value="for $spsID in svc:seq_parameter_set_id return $svc:sps
    [$spsID+1]/../@rdo:id"/>
  </xsd:complexType>
</xsd:element>
```

Listing 1. Augmented BSDL test schema exposing the ancestor Data Units of an SVC PPS

A number of options exist for expressing such rules:

➢ **In the binary domain:** A BSDL Schema may be extended so that it appends attributes to the output which correspond to the needed RDO content features (size, dependencies, etc). The advantage of this approach is that the description of how these features are extracted from the binary data is very concise. The disadvantage is that this description is embedded in the BSDL Schema and is therefore tightly coupled, limiting reusability.

Listing 1 is an example of this approach. It shows part of a BSDL schema that outputs an rdo:ancestors attribute expressing the interdependency of Data Units[5]. Thus, attribute declarations like this are one way to provide mapping rules from the format-specific binary structure of the schema, to the format-independent concepts needed for RDO (which are represented by members of the rdo namespace).

[5] The for structure used by the attribute value specification is not a loop, but rather a workaround for the fact that XPath does not have a current() function (cf. the sps variable in Listing 3).

```
<xsl:template match="pps">
  <xsl:variable name="sps"
    select="preceding::sps[seq_parameter_set_id =
                                   current()/seq_parameter_set_id][1]"/>
  <xsl:copy>
    <xsl:attribute name="rdo:ancestors" select="$sps/@rdo:groupID"/>
  </xsl:copy>
</xsl:template>
```

Listing 3. XSLT fragment annotating pps elements with ancestor metadata

- **In the XML domain:** A second option is to describe the identification of the necessary features using XQuery (Boag et al., 2007) or an XSLT (Clark, 1999) stylesheet. This removes the tight coupling with the BSDL Schema, but is less succint, and adds an additional layer of complexity to the process. Listing 3 shows a fragment of an XSLT stylesheet that adds the same rdo:ancestors attribute to a BSDL description.

- **In the semantic domain:** Alternatively, the BSDL Schema may be directly converted to OWL classes, allowing the feature identification process to be specified using an ontological reasoning tool such as the Semantic Web Rule Language (SWRL) (Horrocks et al., 2004). One disadvantage of this approach is that RDF is inherently unordered, and so Data Unit order must be explicitly imposed using sequence numbers, timestamps or the like. Furthermore, some assertions about the order of such sequence numbers are non-monotonic (see for example Listing 4).

  Examples of mapping rules using SWRL are (the prefix svc is used for the BSDL Schema, and rdo for the RDO ontology):

  $$\text{svc:nalUnit(?x)} \rightarrow \text{rdo:dataUnit(?x)} \qquad \ldots(2)$$

  which implies that a Network Abstraction Layer (NAL) Unit is an atomic unit of data for the purposes of RDO. This deceptively simple rule is in fact making use of the inheritance properties afforded by SWRL and the semantic web, since there are no direct instances of svc:nalUnit within an svc instance, but rather it is the abstract superclass of all other top-level objects in an SVC stream. This inheritance is unavailable to an XSLT-based rule (e.g. Listing 3), where separate rules must be specified for each instance type); and Listing 4 which (almost) states that a Picture Parameter Set (PPS) has a dependency to the *most recent SPS with an ID that matches the one given in the PPS*. If multiple SPS' with the given ID are present in the bitstream prior to the PPS, then rule 4 on the next page will incorrectly match all of them. The missing constraint—"most recent"—is nonmonotonic and hence not supported by SWRL or OWL-DL[6]. Consequently, an XML-based approach has been applied to the mapping rules for syntactic parameters in the example system implemented in this work (see Section 5 on page 19). Future work on SWRL and/or other Semantic Web rule languages may provide the expressivity needed for this and other rules required for RDO parameters.

Examples of mapping rules for Δ_{rate} and Δ_{time} are given in Section 5 on page 19.

[6] The missing "most recent" constraint is specified in the XSLT example (Listing 3) by the preceeding:: axis and [1] predicate

```
svc:pps(?pps) ∧ svc:spsID(?pps,?spsID) ∧ svc:sps(?sps) ∧ svc:spsID(?sps,?spsID) ∧
svc:seqNo(?pps,?ppsSeqNo) ∧ svc:seqNo(?sps,?spsSeqNo) ∧ swrlb:lessThan(?spsSeqNo,?ppsSeqNo)
                        → rdo:dependsOn(?pps,?sps)
```

Listing 4. A not-quite-complete rule for specifying the interdependency between a PPS and the SPS it references

4.2.2 Semantic analysis

The aim of semantic analysis is to generate metadata that can subsequently be reasoned on to compute Semantic Distortion. There are many options for generating and obtaining semantic metadata, as discussed below. Crucially, there are also several disparate widely-used methods for serialising this metadata (RDF (Beckett, 2004), XML (Bray et al., 2008), as well as numerous binary forms (ISO/IEC, 2002b; Matroska, n.d.; Nilsson, 2000)). To be able to reason on such metadata (in order to compute Semantic Distortion) it must generally be translated into a single form. There are several options for this, but for the sake of brevity, and because it is the most powerful option for semantic reasoning, this section will focus on translation of binary and XML metadata into RDF/OWL.

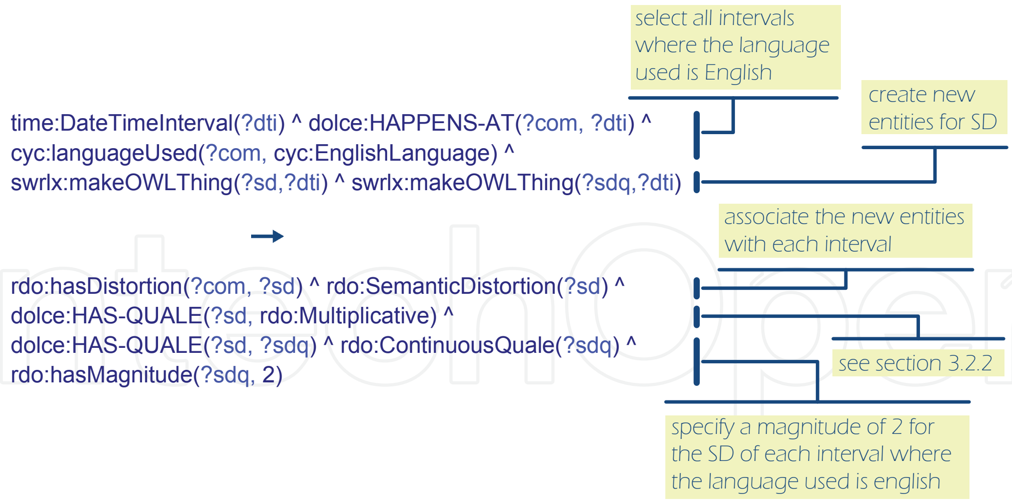
Generating the desired content semantics

The first stage of semantic analysis involves extracting the desired semantics from the content (e.g. "this scene depicts the studio anchor discussing the news story"—see Figure 9 on page 17). As described in the introduction, this remains a challenging problem, with many efforts directed toward algorithms able to expose various specific semantics for media content. This process typically uses the uncompressed samples of the original content and (partly because of the volume of this data) can be very computationally expensive (Figure 5 on page 7). While such semantic metadata may not be specifically designed for the delivery process, it can often, nonetheless, contribute to it. For either of these reasons the semantic analysis may often be performed asynchronononously to the operation of the RDO-hinter. Further, much semantic metadata is presently annotated by hand, consider Flickr/Youtube tags, or iTunes song ratings, for example.

Whether semantic metadata is the result of an (a)synchronous analysis step or manual annotation, the result is a set of metadata about the content that is expressed in some machine-processable form. Such metadata is increasingly specified using ontologies, which simplify the integration of heterogeneous data sources as well as the reuse of information for applications other than those for which it was first developed (Naphade et al., 2006). However, this is far from universal, and a great deal of existing semantic metadata is stored as XML or binary data (see, for example, (ISO/IEC, 2002b; Matroska, n.d.; Nilsson, 2000)).

Translating XML metadata

As discussed earlier (Section 4.2.1), it is possible to map XML-Schema-based metadata directly into the semantic domain (Lehti & Fankhauser, 2004). This may be imprecise—it is generally possible to express the same semantics using several different XML structures (e.g. element



```
time:DateTimeInterval(?dti) ^ dolce:HAPPENS-AT(?com, ?dti) ^
cyc:languageUsed(?com, cyc:EnglishLanguage) ^
swrlx:makeOWLThing(?sd,?dti) ^ swrlx:makeOWLThing(?sdq,?dti)

        →

rdo:hasDistortion(?com, ?sd) ^ rdo:SemanticDistortion(?sd) ^
dolce:HAS-QUALE(?sd, rdo:Multiplicative) ^
dolce:HAS-QUALE(?sd, ?sdq) ^ rdo:ContinuousQuale(?sdq) ^
rdo:hasMagnitude(?sdq, 2)
```

select all intervals where the language used is English

create new entities for SD

associate the new entities with each interval

see section 3.2.2

specify a magnitude of 2 for the SD of each interval where the language used is english

Listing 5. A SWRL rule that specifies the Semantic Distortion of English Communication

content vs attributes)[7]. An alternative proposed by Hunter is to use an upper-level ontology (Hunter, 2001), which is more robust, specifically because it involves a time-consuming manual mapping from the (implicit) semantics underlying the XML representation to a set of explicit ontological relations. Both of these approaches are feasible for a semantic-aware delivery framework.

Translating binary metadata

There are a number of very widely used binary formats for semantic metadata: ID3 (Nilsson, 2000), EXIF (JEITA, 2002), and MPEG-4/Quicktime (Apple, 2001; ISO/IEC, 2004b) for instance. In this case, a syntactic analysis of this metadata must precede further processing of the semantics themselves. This syntactic analysis can be performed in the same manner that interdependencies and other constructs are exposed (Section 4.2.1). This will yield an XML description of the structure of the metadata, including the name and value of all of the desired metadata fields. This XML may then be mapped into the semantic domain, as above.

Computation of Semantic Distortion

This is the second stage of semantic analysis, and is one of the central contributions of this work. *Semantic Distortion* (SD) is defined as a measure of the "SNR" between the intended semantic (meaning) of the content before it is encoded, as compared to the semantics conveyed by the content that is rendered for its recipient(s). The contribution to Semantic Distortion that this chapter is primarily concerned with is that contributed by the delivery process, however the approach may also potentially be useful for other aspects of multimedia processing.

[7] In contrast, when Section 4.2.1 discussed mapping *BSDL Schema* to an ontology, there was no such imprecision. BSDL has already restricted the expressivity of XML Schema in order to guarantee an unambiguous mapping between the binary and XML domain and back again. As a result, mapping BSDL into the semantic domain is also unambiguous.

Clearly, this notion of Semantic Distortion is highly subjective (as indeed are many of the semantics of any given piece of media content). However, even approximations of Semantic Distortion as perceived by parties on the server-side of the process possess substantial value for optimising the delivery of the content semantics, as shown in Section 5 on page 19.

Given this definition of Semantic Distortion, it is possible to define a series of rules that map from concepts expressed in semantic metadata to a quantitative measure of SD. Although the content of mapping rules for SD will differ from those of syntactic analysis (see 4.2.1 on page 8), they have the same range of options for specification: directly within a (binary) schema, in the XML domain, or in the ontological domain. In contrast to the aforementioned syntactic mappings, SD rules translate readily into SWRL, such as Listing 5[8] which states that if there is an instance of *Communicating* during a certain time interval that uses the *English Language*, then the magnitude of the Semantic Distortion for that interval is doubled[9]. This rule covers both spoken communication (in which case the SD is associated with the audio track(s)), and visual communication (e.g. subtitles; where the SD is applied to the video).

Combination of Semantic Distortion with sample distortion

This is pivotal to the correct operation of the R-D optimisation algorithm. Chou (2006) considers Sample Distortion to be *additive*, that is, the overall distortion $D(\pi)$ is a large initial value D_0 less the sum of the distortion of all packets D_l *received and useful* (which are computed by the product sequence):[10]

$$D(\pi) = D_0 - \sum_l \Delta D_l \prod_{l' \preceq l} (1 - \epsilon(\pi_{l'})) \qquad \ldots(6)$$

However, the sample distortions used in Equation 6 are all measured according to a single algorithm, and hence have the same scaling and are directly comparable. This is not usually the case for Semantic Distortion, and is certainly not so when comparing Semantic Distortion with Sample Distortion. Instead, it is proposed that Semantic Distortion be considered to be *multiplicative*; that is, that SD represents a weighting factor that may be applied to a value of sample distortion for a packet, or group of packets. There are several motivations for this:

➢ First, multiplicative combination obviates the need for normalisation based on potentially unknown response curves for distortion algorithms (both sample and semantic). For example, say a Data Unit has a Sample Distortion with a magnitude of 0.3 dB, and a Semantic Distortion (based on the language of the communication) of magnitude 2. It is clear that these values cannot be combined additively without ensuring that they are first normalised to the same scale. However, while sample distortion uses objective measurements such as (P)SNR, the same cannot in general be said of

[8] where cyc: refers to the CYC upper-level ontology (Matuszek et al., 2006), time: to the OWL-Time ontology (Pan & Hobbs, 2004), dolce: to the DOLCE ontology (Gangemi et al., 2002), and rdo: to the ontology defined in 4.3 on the facing page. Note that the use of Cyc, OWL-Time, and DOLCE are not intended to be normative, they are used merely as an example of how to define mapping rules for SD.

[9] Note that the factor of 2 is in this case relatively arbitrary—yet still shown to be useful (Section 5)—see Section 6.1 on page 23 for a discussion about possible future work on methods for evaluating Semantic Distortion.

[10] where $l' \preceq l$ selects all packets l' that are ancestors of l as well as l itself, π is the vector of packet transmission policies, and ϵ the error/delay probability distribution for any given policy

subjective measurements of Semantic Distortion. Even though there are numerous quantitative measures (see for example ITU-T (1998)), comparison between data from different quantitative tests is challenging. Furthermore, Semantic Distortion is intended to encompass a wide range of data, as discussed (Section 2), beyond formal subjective testing. Combining disparate data sets multiplicitively avoids this need to normalise.

- Combination of several *Semantic Distortion* data-sets relating to a piece of content typically has similar issues relating to normalisation. Consider, with the above example, the addition of a second Semantic Distortion data set computed from Temporal Information (TI). The SD for the Data Unit in question has a magnitude of 0.7. Combining this datum with the others is straightforward as a scaling factor (i.e. multiplicative).
- Finally, multiplicative combination retains a known zero point. This is important if either sample or any Semantic Distortion has a magnitude of zero; in the first case, this indicates that the packet has no effect on the reconstruction of the signal; in the second, that it does not convey any semantics. Either way, these features must be transmitted to the output distortion value.

4.3 An Ontology for Semantic Distortion

The mapping process described in Section 4.2.2 on page 12 requires the definition of appropriate concepts to be used as the destination of the rules. These concepts fall into two categories: the formal definition of a Data Unit in so far as it pertains to R-D optimisation, and the definition of Semantic Distortion itself. These are described below in Sections 4.3.1 and 4.3.2, respectively. These definitions and their associated concepts are attached to the DOLCE (Gangemi et al., 2002) upper-level ontology, because of its precise separation of fundamental concepts[11]. Figure 8 on the following page depicts the Semantic Distortion ontology (prefixed by sd) along with its DOLCE ancestors (prefixed by dolce). Refer to (Gangemi et al., 2002) for a full treatment of DOLCE; the following description should suffice for this work.

The fundamental distinction in DOLCE is between *enduring* and *perduring* entities (Figure 8). The precise philosophical definition of these terms is complex and also somewhat controversial (Gangemi et al., 2002), but for the purpose of this chapter it will suffice to say that the former are entities that *exist* in some region of time (and possibly space), whereas the latter are events that *occur* during a region of (space-)time. Both endurants and perdurants have *qualities*, and a distinction is made between a quality (e.g. color, temperature) and its *quale*—a region defining the "value space" of a particular quality (e.g. red, 298K). This is partly inspired by the fact that an endurant individual will permanently have particular quality individuals (i.e. it will always have *a* color), but the value of those qualities may change over time. Quales belong to the class of all *abstract* concepts that are neither endurant nor perdurant.

While DOLCE includes the abstract notions of a temporal quality and a temporal region, RDO requires a more concrete conceptualisation of time in order to be able to synchronise semantic metadata with the underlying media Data Units. Furthermore, the metadata that a semantic-aware delivery framework must assimilate will have a large variety of fundamentally different representations of time:

- MPEG-7 (ISO/IEC, 2002a);
- SMPTE (of Motion Picture & Engineers, 1999);

[11] As described previously (Section 4.2.2 on page 12), the choice of DOLCE is not normative but rather a preferred embodiment

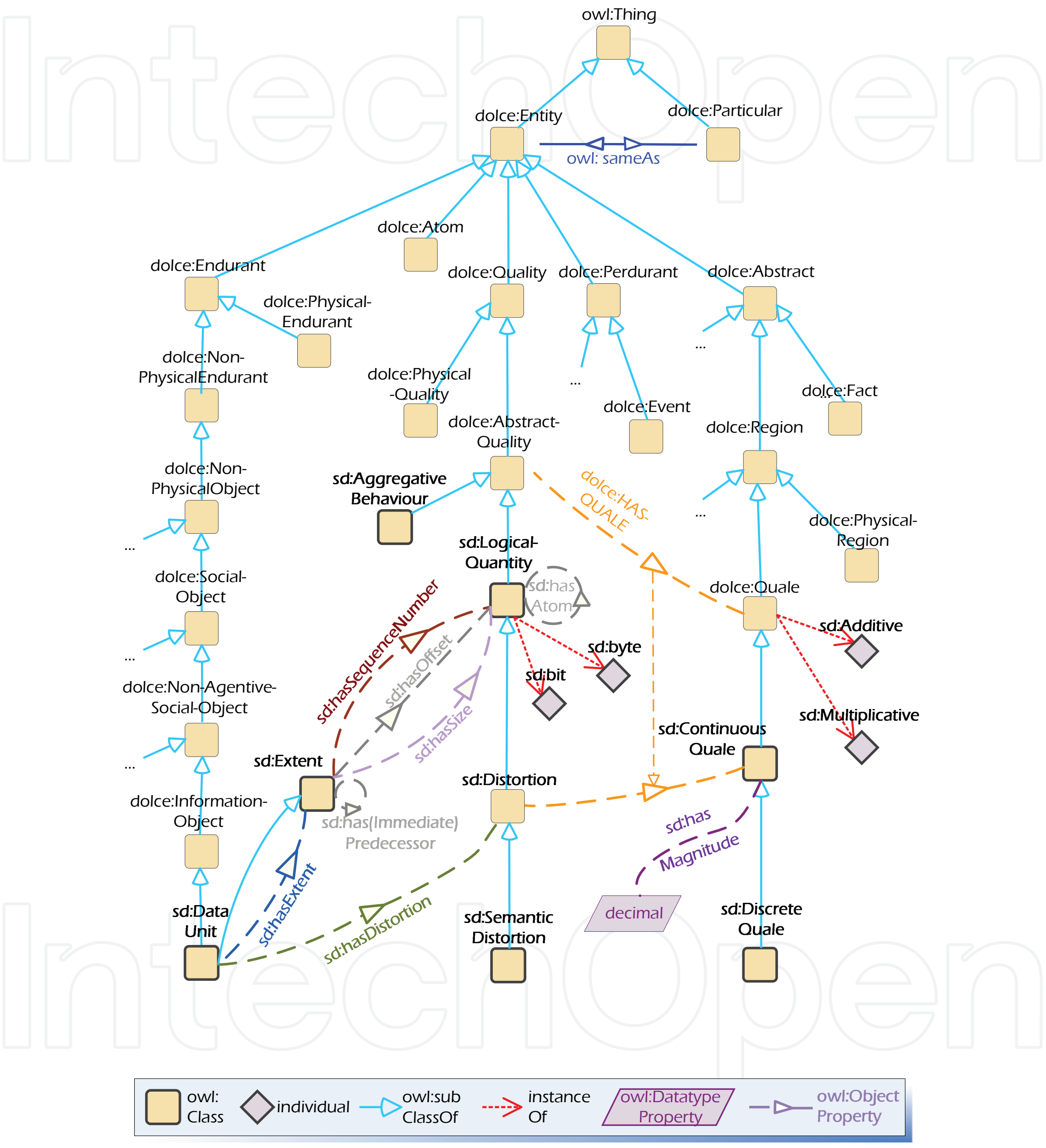


Fig. 8. An ontology for Semantic Distortion

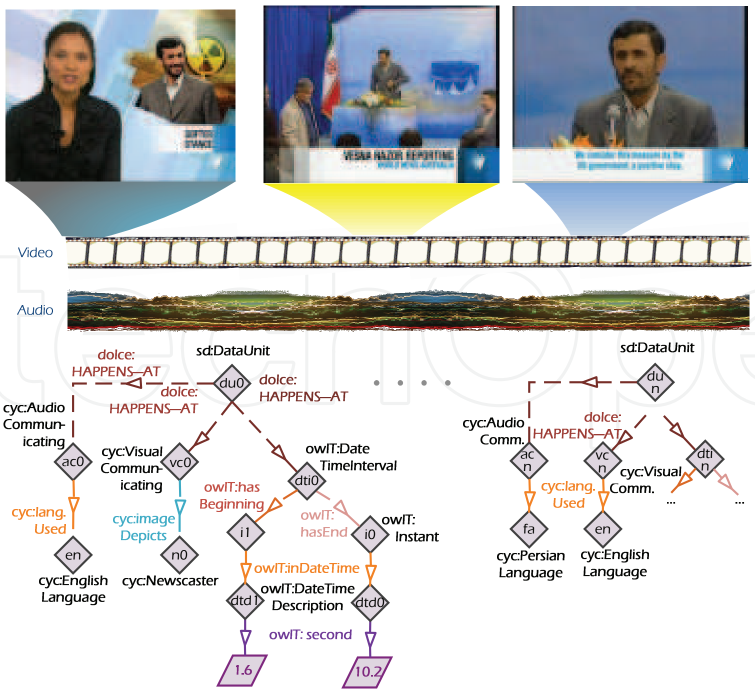


Fig. 9. Example of the semantic annotation of an Audio-Visual Clip

➢ XML Schema (*XML Schema Part 2: Datatypes Second Edition*, 2004);

➢ OWL Time (Hobbs & Pan, 2006); as well as

➢ the innumerable binary syntaxes used by media formats.

Each representation uses a different syntax to represent time. If these are to be reasoned on as part of semantic-aware delivery, methods are required to translate from one to another.
Throughout the proceeding discussion, the example shown in Figure 9[12] will be used to illustrate each concept. The example consists of a short Audio-Visual clip that forms part of a news article on events in the Middle-East[13]. The first part of the clip depicts a studio presenter introducing the story (the temporal interval containing this section is described by an owlT:DateTimeInterval, and the visual and aural communication features with cyc:(Visual/Audio)Communicating). Subsequently, contextual footage is shown of the subject walking to the podium while an off-screen narrator continues the story. Finally, the subject speaks in Persian with English subtitles appearing below (using similar owlT:DateTimeInterval and cyc:(Visual/Audio)Communicating instances). These features are annotated via CYC (Matuszek et al., 2006) classes and properties and then reasoned on using mapping rules (Listing 4).

4.3.1 Data units

For the purposes of R-D Optimisation, Chou (2006) designates an atomic segment of data as a DataUnit, where each packet on the network may contain at most one Data Unit. Rule 2 on page 11 is an example of the use of DataUnit, showing how it enables format-independence

[12] Individual IDs used in Figures 9–10 are used purely to differentiate individuals from each other. The general naming scheme used for these IDs is to abbreviate the *type name* of the individual and append a number which increments from 0, 1. . . n for each type. For example, the first instance of the Owl-Time class DateTimeInterval in Figure 9 has the ID dti0.

[13] This clip was used by permission from SBS World News Australia.

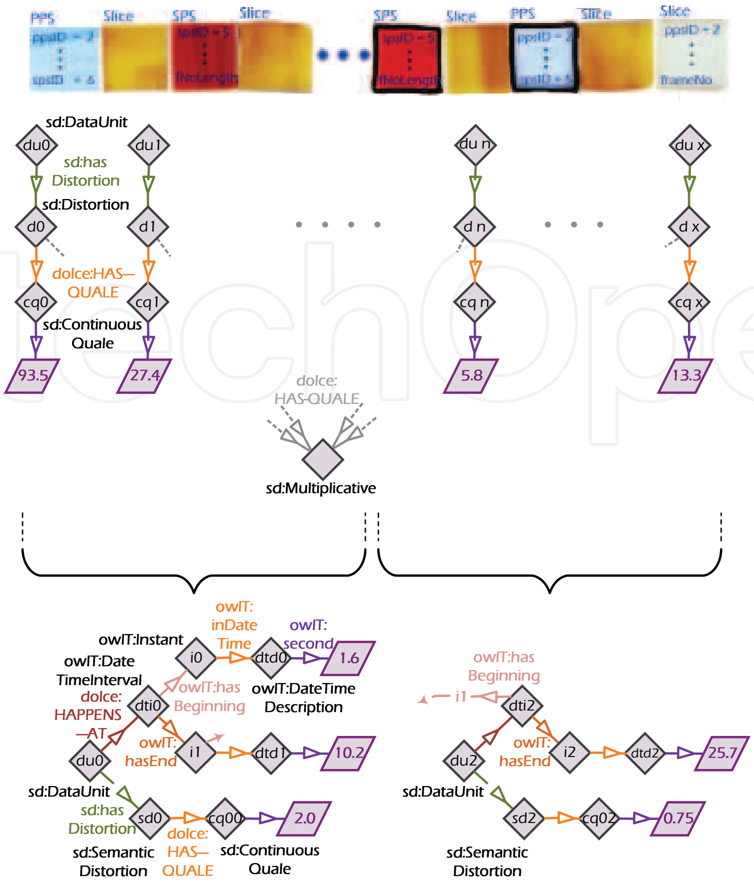


Fig. 10. Example instances of SD classes describing the distortion of a H.264/AVC bitstream

by mapping from arbitrary format-based structures to a single interface for RDO. Figure 10 depicts an example H.264/SVC bitstream, along with instances of the Semantic Distortion classes that deliniate the Data Units in the stream.

4.3.2 Distortion

Distortion is the other central concept in RDO. It is a type of Logical Quantity that measures "the amount by which the distortion at the receiver will decrease if the Data Unit is decoded (on time)" (Chou, 2006). Distortion has at least two distinct types: sample and semantic, which were described earlier (Section 4.2.2 on page 12). Distortion in general has a continuous value-space (continuousQuale), which in turn has a data-type property hasMagnitude. Distortion also has an hasAggregativeBehaviour property (this relation is not shown), which may be Additive or Multiplicative. It is left to the user to decide which behaviour(s) to assign.

Figure 10 shows an example of the distortion instances for a bitstream. In this example, each NAL unit in a H.264 bitstream is given a sample-based distortion according to its contribution to a scheme (for example that proposed by Chou (2006))—these are the DataUnit, Distortion and Quale individuals toward the top of the figure. In this instance, SemanticDistortion applies to more coarsely-grained Data Units. These are specified using OWL-Time intervals (by mapping from the intervals previously shown in Figure 9 on the preceding page). Each instance of Semantic Distortion intersects many Data Units in the actual media. Assigning Semantic Distortion to individual media Data Units is done by the mapping rules associated with syntactic analysis (Section 4.2.1, above).

4.4 Summary

In summary (refer to Figure 1 on page 2), there are three primary components to the proposed semantic-aware multimedia delivery framework:

- ➢ a *hinter*, which computes all of the content-based metadata offline;
- ➢ a *delivery node* which is left with as little work to do as possible, since the hinter has already performed much of the necessary computation. The delivery node simply forwards or drops each packet as it arrives, based on some optimisation algorithm; and
- ➢ *semantic analysis*, used to provide the hinter with metadata that can be used to compute *Semantic Distortion*.

The novelty of this work consists in tying semantic analysis and semantic metadata to the R-D hinter, as well as an architecture for this hinter so that it operates in a format-independent manner. More specifically, the work contributes:

(a) the definition of *Semantic Distortion* (SD) (Section 4.2.2 on page 12);

(b) an ontology for RDO concepts and for SD that enables it to be inferred from arbitrary semantics, and then combined with sample distortion (Section 4.3 on page 15);

(c) extension of the concept of an R-D Hinter to encompass SD (Section 4.2 on page 7);

(d) a format-independent architecture for the semantic hinter, which operates by extracting all format-specific details into declarative data (schemata and mapping rules). It is argued that this is imperative to allow the increasingly diverse range of formats and devices to interoperate (Section 4.2.1 on page 8); and

(e) a semantic-independent architecture for the hinter, again accomplished by using schemata (ontologies) and mapping rules (Section 4.2.2 on page 12);

5. Subjective testing

5.1 Methodology

Double-blind, randomised subjective testing was used to validate the hypothesis that the use of Semantic Distortion can improve multimedia delivery. The scenario used for these tests was a mobile environment—where channel characteristics are often highly variable, and handset capabilities mean that audio and video require relatively similar bandwidth. As such, the source material was encoded at 22.05kHz for the audio, and the video at QVGA resolution and 15 frames per second. Initial trials were conducted using a mobile (cellular) handset, but it was decided that this introduced a significant number of variables (e.g. the particularly small screen size, problems with controlling playback, and uncertainties about the quality of the audio rendering hardware) without lending any additional credence to the experiment per se (as opposed to conducting the trials using a notebook, but using mobile-ready content). Consequently, respondents evaluated video displayed on the screen of a Compaq nc4000 notebook (1024x768 total resolution, 12″ screen), and listened through Sony MDR-V500 headphones. Respondents were free to adjust volume and viewing distance as desired, with the latter ranging from 8 to 16H (the QVGA image measured 75mm W × 58mm H). The testing was conducted according to ITU-T P.911 (ITU-T, 1998), including the conditions prescribed in Table 4[14]. Pairwise Comparison (PC) was used to evaluate the hypothesis that

[14] screen luminance, ratios & chromitacity, background illumination and noise level

```
 cyc:VisualCommunicating(?vid) ∧ t:DateTimeInterval(?dti) ∧ dolce:HAPPENS−AT(?vid, ?dti) ∧
 cyc:imageDepicts(?vid, cyc:StillImage)∧swrlx:makeOWLThing(?sd, ?dti)∧swrlx:makeOWLThing(?sdq, ?dti)
          →
rdo:hasDistortion(?vid, ?sd) ∧ rdo:SemanticDistortion(?sd) ∧ dolce:HAS−QUALE(?sd, rdo:Multiplicative) ∧
   dolce:HAS−QUALE(?sd, ?sdq) ∧ rdo:ContinuousQuale(?sdq) ∧ rdo:hasMagnitude(?sdq, 0.5)
```

Listing 7. A rule from the test data asserting that still images have a (relative) SD of 0.5

> *"use of Semantic Distortion in multimedia delivery improves the communication of the meaning/semantics of the content."*

To this end, the nineteen respondents were asked to decide which clip (A or B) "best conveys the gist of the news article to you." Respondents were not skilled in the arts of multimedia delivery, or subjective testing. An approximately equal number of each gender was chosen, and participants ranged in age from 16 to 70. Levels of familiarity with digital media varied as may be expected within the stated age range. Respondents had normal or corrected-to-normal eyesight, and normal hearing (with the exception of two participants who had mild age-related high-frequency hearing loss).

There were four clips plus an initial (hidden) training clip, all exhibiting some or all of the characteristics depicted in Figure 2 on page 2. Three were news footage, and the fourth part of an interview between an English interviewer and a Japanese interviewee, all between 25 and 45 seconds in length. The audio from each clip was encoded using Scalable to Lossless Coding (SLS) (ISO/IEC, 2005b) with an AAC base layer of 6kbps to provide a large scalable range. Scalable Video Coding (SVC) (ISO/IEC, 2007) was used for the video with 8 coarse-grained scalability (CGS) SNR (quality) layers (with LQP at 30, 34, 38, 42, 45, 48, 51, 54 for layer 0 to 7 (respectively), and $RQP = LQP + 2dB$) and 4 medium-grained SNR layers. Spatial and Temporal layers can be beneficial to semantic-aware optimisation (see, for example Cranley & Murphy (2006)) but it was decided to limit the sources of variability for the present experiment. In that regard, no attempts were made at error concealment[15], even though this would have an impact on a user's perception of a real world system employing SD.

5.1.1 Semantic analysis

Semantic analysis for each clip was conducted using classes from the Cyc (Matuszek et al., 2006) ontology, to provide semantics indicating the language of communication (spoken or written), among other things. The choice of Cyc for this task was purely as an example, the semantic-aware delivery framework places no constraints on specific metadata ontology(s) (as discussed in Section 3 on page 3). Mapping rules were created for these classes (Listing 4 is one of these, and Listing 7 another[16] describe how particular semantics relate to SD. Temporal regions were specified using OWL-Time (Hobbs & Pan, 2006). Again, this choice is by way of example only, other temporal schemata may equally be used. Figure 10 on page 18 depicts example SD and OWL-Time instances.

[15] except for silencing of an SLS decoder bug observed at particularly low bit rates that that led to saturation of the signal in sections where there should be silence. This bug was observed equally on both clips in a pair, and would otherwise have caused discomfort for test subjects.

[16] Full rules are available in Appendix I of Thomas-Kerr (2009)

5.1.2 Syntactic analysis

Syntactic analysis was conducted using a BSDL Schema for SLS and another for SVC (see Appendix I in Thomas-Kerr (2009)), then an XSLT stylesheet to map SLS & SVC fields to the necessary RDO metadata (as per Section 4.2.1 on page 8). Delivery optimisation was performed using a very simple algorithm, so as to limit (as much as possible) the testing to the Semantic Distortion concept, rather than introduce a second independent variable in a sophisticated optimisation routine. Essentially, the algorithm used was

1. Using the rules generated in Section 5.1.1, compute SD values for the entire clip;
2. Aggregate the SD values separately for audio and for video, according to the behavior specified (see Sections 4.2.2 and 4.3);
3. Segment the clip into regions so that each region has a constant SD for audio and a constant SD for video;
4. For each region
   (a) Apportion the target bandwidth between the audio and video stream according to the aggregated SD for each component;
   (b) Truncate each SLS frame so as to achieve the apportioned audio bit-rate; and
   (c) Drop SVC NALUs to most closely approximate the target video rate (while respecting the discardable flag).

Each clip was encoded to three different bit rates using this method, for a total of twelve clips, plus the hidden training clip. For each semantic-aware clip produced using this algorithm, a reference clip was created with the same average audio and video bit rates as the semantic-aware clip (by truncating the SLS and dropping SVC NALUs). This means that the semantic-aware clip devotes more of the available bandwidth to that part (in this example, audio or video) that carries more of the semantics of the content, whereas the reference sample uses the same total bandwidth, but has a static ratio between audio and video. This is illustrated in Figure 11 which shows the semantically-adapted and equivalent average rate series for the audio tracks of the high-bitrate “iran” sequence. The video tracks are not shown since the coarser granularity of the video scalability means that variance is too great to discern average trends. Nonetheless, the audio tracks clearly show how the adaptation algorithm responds to varying SD, and also that both audio tracks have the same total average rate.

5.2 Results

In total, 72% of the semantic-aware clips were preferred by subjects when compared to the average-rate reference clip (as shown, Figure 12), with a variance of 20.57%[17] and a 95% confidence interval of ±5.74%.

Of the twelve pairs, one very low-rate semantic-aware clip was rated as worse than its average-rate partner. It is likely that this is due to the deliberate simplicity of the adaptation algorithm. A more sophisticated algorithm would be expected to deal with such outliers more effectively. Having said this, three respondents independently remarked that they preferred one particular low-rate non-semantic-aware clip because it accorded the speaker "more respect" by making his voice clear, even though they couldn’t understand it. Because

[17] variance is not considered to be particularly informative in this instance, due to the binary nature of each sample in a Parwise Comparison (the respondant picks one clip or the other). Because of this, every sample is relatively distant from the mean.

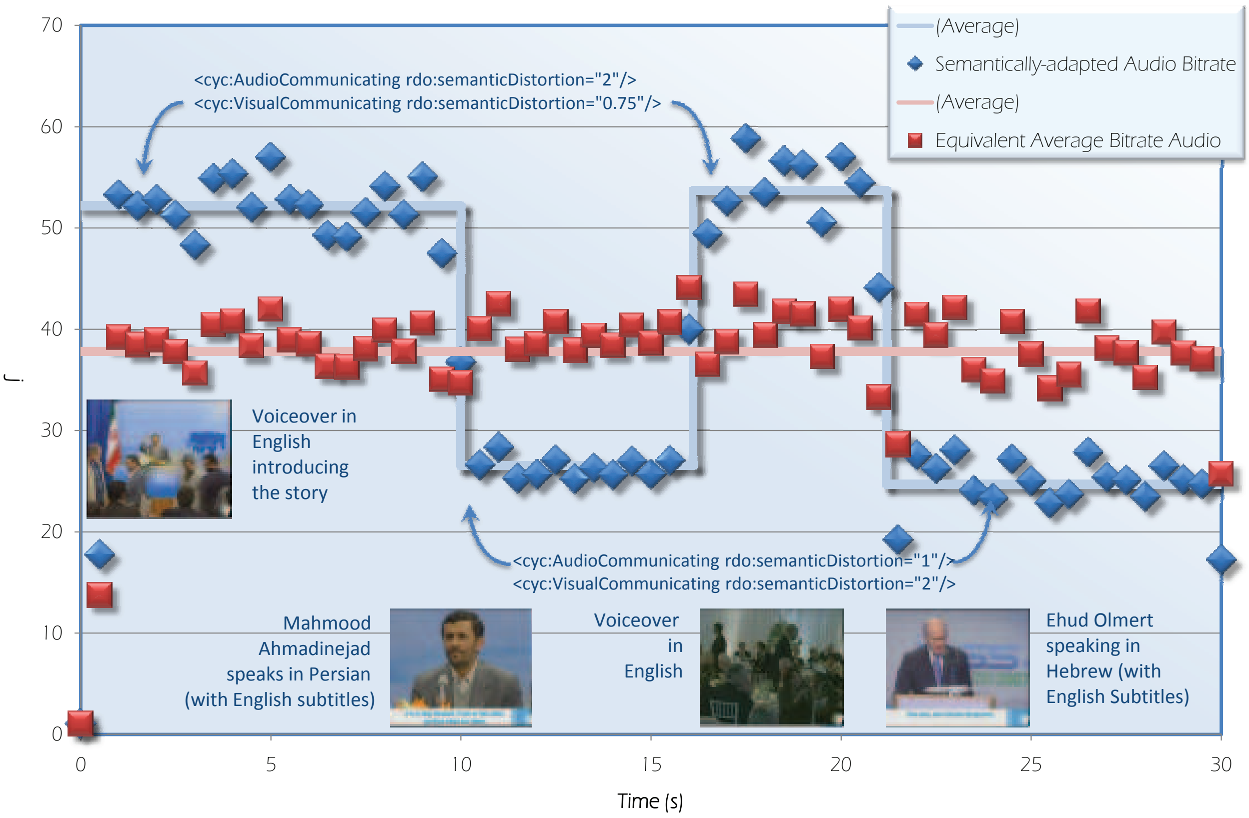


Fig. 11. Semantic adaptation apportions bandwidth according to the content *meaning*

the tests were double-blind, it is not known whether this comment corresponded to the clip in question.

Another two clips were voted as no better and no worse, and the remaining nine semantic-aware clips were preferred 84% of the time. This demonstrates that Semantic Distortion is of significant benefit in the multimedia delivery process. Moreover, the system proposed in Section 4 is effective in processing Semantic Distortion and R-D optimisation-related metadata in a way that meets the objectives identified in Section 3. In contrast, however, the result also suggests that the use of Semantic Distortion to optimise the apportionment of bandwidth between audio and video streams could possibly not be beneficial for a minority of content, at least without more sophisticated optimisation algorithms. However, while the modal trade-offs employed for these few cases fails to yield an improvement, it is quite possible that other uses of Semantic Distortion (see Section 4.1 on page 5) may give the desired results. Further investigation of this is left to future work.

6. Conclusion

This chapter describes a framework for incorporating semantics into the multimedia delivery process. It builds on existing work for exposing semantics in content and delivering media in a rate-distortion optimal way. In effect, this alters the conceptual end-points of the multimedia delivery chain. Instead of server-client, using semantics extends the process to (human) creator-consumer, by minimising distortion of the *intended meaning* of the content (see for example the news report in Figure 2 on page 2). At the same time, the framework provides the flexibility to incorporate new semantics, optimisation algorithms, and content formats as

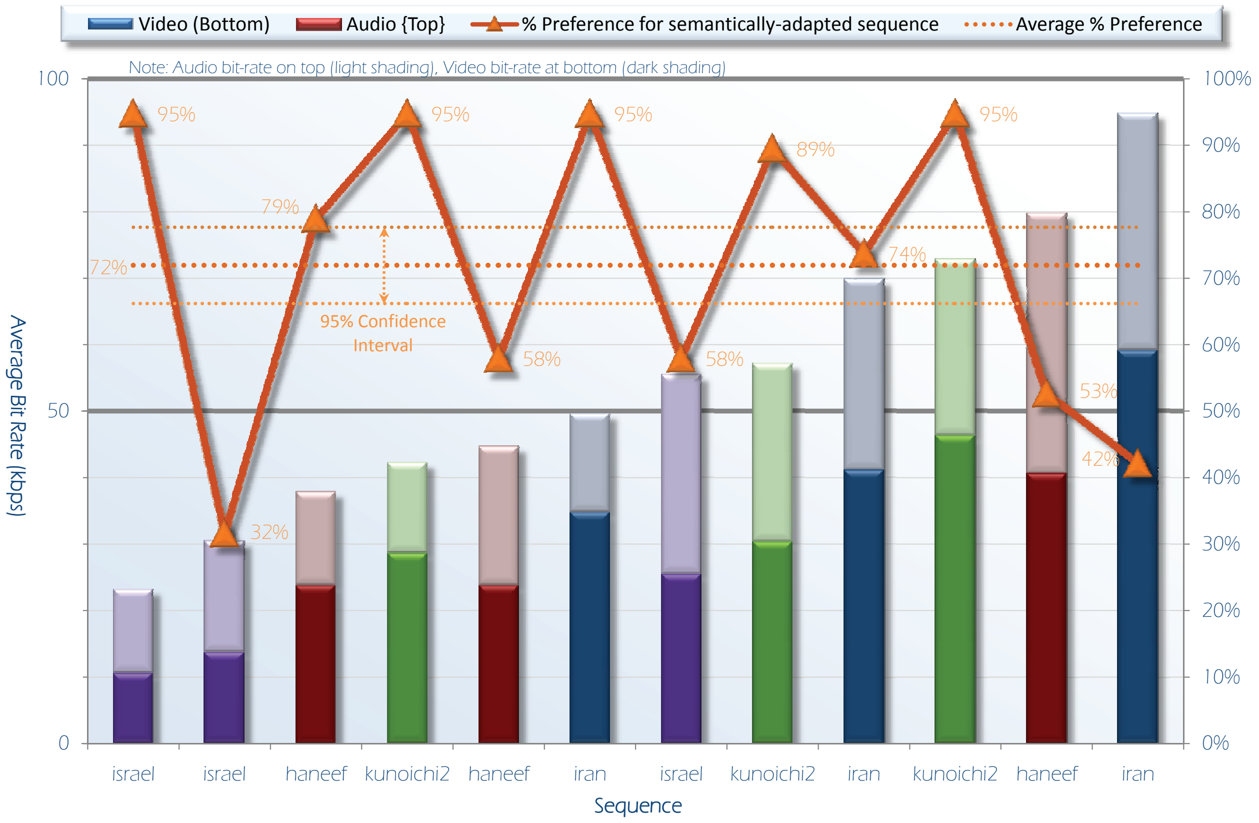


Fig. 12. Subjective testing shows a 72% preference for Semantically-aware media delivery

they become relevant. This process can operate largely without the addition of new software or hardware, since format-specific details are provided in schemata rather than hard-coded. The framework has been validated via subjective testing that asked candidates to make a pairwise comparison between a video clip that had been semantically adapted (more bandwidth devoted to that mode carrying more of the content semantics) and one adapted to an equivalent constant average bitrate. In total, 72% of the semantically adapted clips were preferred by subjects when compared to the average-rate reference clip. Of the twelve pairs, one semantically adapted clip was rated worse than its average-rate partner. Two were voted as no better and no worse, and the remaining semantically adapted clips were preferred 84% of the time.

This result demonstrates that Semantic Distortion is of significant benefit in the multimedia delivery process. Furthermore, it validates the format-independent architecture proposed in Figure 3, as well as the simple algorithm used to semantically adapt content along its modal axis, across a range of bit-rates typical of current mobile communication channels. Nevertheless, the results suggest significant scope to develop higher performance semantic adaptation algorithms. Some possibilities for this are suggested in the following chapter.

6.1 Future work

The present work has focused predominantly on the format-independent semantic hinter. Future work may consider more closely the design of the semantic analysis and delivery node modules (see Figure 3 on page 5). In this work, the syntax of compressed media content was described declaratively (using schemata) to enable a generic hinter to extract data for the R-D process. Semantic analysis, on the other hand, is generally conducted using raw (uncompressed) media data, however here too numerous content formats are used. Furthermore, there are a wide range of low-level semantic features (e.g. color, texture, luminance) that are extracted from the raw data in order to infer higher-level semantics. Future work could therefore investigate declarative mechanisms for (a) describing how low-level features are computed from a given content format, and (b) mapping such features to high-level semantics.

Secondly, future work should consider how to describe an R-D optimisation algorithm using declarative language. This chapter has addressed the format-independent design of the RDO metadata and send/drop module (Figure 4 on page 6). However, it has not fully addressed a generic mechanism for describing the RDO algorithm itself. Such a mechanism would allow new RDO algorithms to be installed in a diverse variety of existing delivery nodes without requiring their hardware or software to be upgraded.

7. References

Apple (2001). QuickTime File Fmt., `http://developer.apple.com/reference/QuickTime/`.

Austen, J. (1813). *Pride and Prejudice*, T. Egerton, Whitehall.

Baba, M. et al. (2004). Adaptive multimedia playout method based on semantic structure of media stream, *Comms. and IT, IEEE Intl. Symp. on*, pp. 269–273.

Beckett, D. (2004). RDF/XML Syntax Specification (Revised), `http://www.w3.org/TR/rdf-syntax-grammar/`.

Bertini, M. et al. (2006). Semantic adaptation of sport videos with user-centred performance analysis, *Multimedia, IEEE trans. on* 8(3): 433–443.

Boag, S. et al. (2007). XQuery 1.0: An XML Query Language, `http://www.w3.org/TR/xquery`.

Bray, T. et al. (2008). Extensible Markup Language (XML), `http://www.w3.org/TR/xml/`.

Brightman, I. (2005). The trillion dollar challenge: Principles for profitable convergence, *Technical report*, Deloitte; Technology, Media Telecommunications. Available: `http://www.deloitte.com/dtt/cda/doc/content/UK_TMT_TrillionDollarChallenge_05.pdf`.

Chakareski, J. et al. (2004a). Low-complexity rate-distortion optimized video streaming, *Image Processing, Intl. Conf. on* 3: 2055–2058.

Chakareski, J. et al. (2004b). R-D hint tracks for low-complexity RD-optimized video streaming, *Proc. IntŠl Conf. Multimedia and Exhibition* 2: 1387–1390.

Chou, P. (2006). Rate-distortion optimized streaming of packetized media, *IEEE Transactions on Multimedia* 8(2): 390–404.

Clark, J. (1999). XSL transformations (XSLT), `http://www.w3.org/TR/xslt`.

Cranley, N. & Murphy, L. (2006). *Incorporating User Perception in Adaptive Video Streaming Systems*, Idea Group, chapter 12, pp. 242–263.

Cranley, N. et al. (2003). User-perceived quality-aware adaptive delivery of MPEG-4 content, *13th Intl. workshop on Network and operating systems support for digital A/V* pp. 42–49.

Dean, M. & Schreiber, G. (2004). OWL Web Ontology Language Ref., `http://www.w3.org/TR/owl-features/`.

Eichhorn, A. (2006). Modelling dependency in multimedia streams, *Multimedia, 14th ACM intl. conf. on*, ACM Press New York, NY, USA, pp. 941–950.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. & Schneider, L. (2002). Sweetening ontologies with DOLCE, *Lecture notes in computer science* pp. 166–181.

Hobbs, J. & Pan, F. (2006). Time ontology in owl, `http://www.w3.org/TR/owl-time/`.

Hong, D. & Eleftheriadis, A. (2002). XFlavor: bridging bits and objects in media representation, *Multimedia and Expo, IEEE Intl. Conf. on*, pp. 773–776.

Horrocks, I. et al. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML, `http://www.w3.org/Submission/SWRL`.

Hunter, J. (2001). Adding multimedia to the semantic web - building an MPEG-7 ontology, *Semantic Web Working Symposium (SWWS)*, pp. 261–281.

ISO/IEC (2002a). 15938-5 IT–Multimedia content description interface, MDS.

ISO/IEC (2002b). 15938 Information technology—Multimedia content description interface.

ISO/IEC (2004a). 14496-3/Amd.5 Scalable to Lossless Coding.

ISO/IEC (2004b). 14496 Coding of audio-visual objects—Part 1: Systems.

ISO/IEC (2005a). 14496-12, IT–Coding of A/V objects–Part 12: ISO base media file format.

ISO/IEC (2005b). 14496-3:2005/Amd 3, Scalable Lossless Coding.

ISO/IEC (2007). 14496-10:2005/FDAM 3 Scalable Video Coding.

ITU-T (1998). Subjective A/V Quality Assessment Methods for Multimedia Apps., Rec. P.911.

JEITA (2002). Exchangeable image file format (EXIF) for digital still cameras.

Klint, P. et al. (2005). Toward an engineering discipline for grammarware, *ACM Trans. on Software Engineering Methodologies* 14(3): 331–380.

Lehti, P. & Fankhauser, P. (2004). XML data integration with OWL: experiences challenges, *Applications the Internet, 2004. Proceedings. 2004 Int.l Symp. on* pp. 160–167.

Matroska (n.d.). Specification of the Matroska container, `http://matroska.org/technical/specs/index.html`.

Matuszek, C. et al. (2006). An Introduction to the Syntax and Content of Cyc, *Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, AIII Symp. on* pp. 44–49.

Naphade, M. et al. (2006). Large-scale concept ontology for multimedia, *IEEE MultiMedia Magazine* 13(3): 86–91.

Neve, W. et al. (2006). BFlavor: A harmonized approach to media resource adaptation, inspired by MPEG-21 BSDL XFlavor, *Signal Processing: Image Comms.* 21(10): 862–889.

Niedermeier, U. et al. (2002). An MPEG-7 tool for compression and streaming of XML data, *Multimedia and Expo, IEEE Intl. Conf. on*, pp. 521–524.

Nilsson, M. (2000). ID3 tag v2.4.0 - Main, `http://www.id3.org/id3v2.4.0-structure`.

of Motion Picture, S. & Engineers, T. (1999). SMPTE 12M-1999, Television, Audio and Film Ű Time and Control Code.

Paleari, M. & Huet, B. (2008). Toward emotion indexing of multimedia excerpts, *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*, pp. 425–432.

Pan, F. & Hobbs, J. (2004). Time in OWL-S, *Semantic Web Services, AIII Symp. on* pp. 29–36.

Thomas-Kerr, J. (2009). *Building Babel: Freeing multimedia processing and delivery from hard-coded formats*, PhD thesis, University of Wollongong.

Thomas-Kerr, J., Burnett, I. & Ritz, C. (2006). Enhancing Interoperability via Generic Multimedia Syntax Translation, *Proceedings of the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution* pp. 85–92.

Thomas-Kerr, J., Burnett, I. & Ritz, C. (2008). Format-Independent Rich Media Delivery Using the Bitstream Binding Language, *Multimedia, IEEE Transactions on* 10(3): 514–522.

Thomas-Kerr, J., Burnett, I. & Ritz, C. (2009). A system for intelligent delivery of multimedia based on semantics, *Communications and Information Technologies, 2009. ISCIT'09. International Symposium on*.

Thomas-Kerr, J., Janneck, J., Mattavelli, M., Burnett, I. & Ritz, C. (2007). Reconfigurable Media Coding: Self-Describing Multimedia Bitstreams, *2007 IEEE Workshop on Signal Processing Systems*, pp. 319–324.

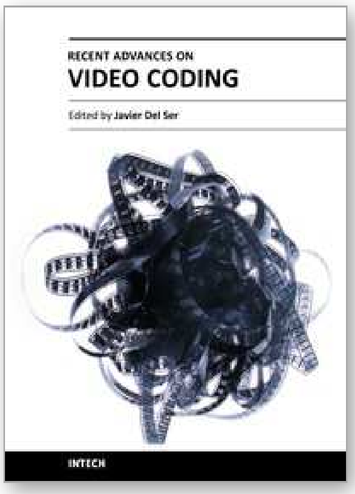
Thomas-Kerr, J. et al. (2007). Is That a Fish in Your Ear? A Universal Metalanguage for Multimedia, *IEEE MultiMedia* 14(2): 72–77.

Thompson, H. S. et al. (2004). XML Schema Part 1: Structures, `http://www.w3.org/TR/xmlschema-1/`.

Timmerer, C. et al. (2006). Digital Item Adaptation - Coding Format Independence, *in* I. Burnett et al. (eds), *MPEG-21*, Wiley, Chichester, UK.

XML Schema Part 2: Datatypes Second Edition (2004). `http://www.w3.org/TR/xmlschema-2/`.

Xu, M. et al. (2006). Event on demand with MPEG-21 video adaptation system, *Multimedia, 14th ACM intl. conf. on* , pp. 921–930.



Recent Advances on Video Coding
Edited by Dr. Javier Del Ser Lorente

ISBN 978-953-307-181-7
Hard cover, 398 pages
Publisher InTech
Published online 24, June, 2011
Published in print edition June, 2011

This book is intended to attract the attention of practitioners and researchers from industry and academia interested in challenging paradigms of multimedia video coding, with an emphasis on recent technical developments, cross-disciplinary tools and implementations. Given its instructional purpose, the book also overviews recently published video coding standards such as H.264/AVC and SVC from a simulational standpoint. Novel rate control schemes and cross-disciplinary tools for the optimization of diverse aspects related to video coding are also addressed in detail, along with implementation architectures specially tailored for video processing and encoding. The book concludes by exposing new advances in semantic video coding. In summary: this book serves as a technically sounding start point for early-stage researchers and developers willing to join leading-edge research on video coding, processing and multimedia transmission.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821