2009

Semantic-Aware Delivery of Multimedia

Joseph Thomas-Kerr
University of Wollongong, jak09@uow.edu.au

Christian Ritz
University of Wollongong, critz@uow.edu.au

Ian Burnett
Royal Melbourne Institute of Technology, ianb@uow.edu.au

Semantic-Aware Delivery of Multimedia

Abstract

This paper describes a system that is able to take arbitrary semantic metadata, and utilize it in the multimedia delivery decision-making process. Format independence is achieved using schema languages to describe the details of any given content or metadata, so that declarative mapping rules can be specified for translating from format-specific data points to format-independent concepts that are directly used by the framework. The system utilizes the criterion of "semantic-distortion", as an extension of Rate-Distortion Optimization based multimedia delivery. Several short video clips were encoded using H.264/SVC scalable video coding, and Scalable-To-Lossless (SLS) audio coding and adapted to four target bit rates. Subjective tests found a 72% preference for those clips which had been adapted so as to devote more bandwidth to the semantically important parts of the content when compared with standard objective-based bit-rate adaptation.

Disciplines

Physical Sciences and Mathematics

Publication Details

J. Thomas-Kerr, C. H. Ritz & I. S. Burnett, "Semantic-Aware Delivery of Multimedia," in Proceedings of the 9th International Symposium on Communications and Information Technologies, 2009, pp. 1498-1503.

Semantic-Aware Delivery of Multimedia

J. Thomas-Kerr[1], *C. Ritz*[1], *I. S. Burnett*[2],
[1]School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong NSW Australia 2522
{jak09, critz}@uow.edu.au

[2]School of Electrical and Computer Engineering
Royal Melbourne Institute of Technology, Melbourne VIC Australia 3000
ian.burnett@rmit.edu.au

***Abstract*— This paper describes a system that is able to take arbitrary semantic metadata, and utilize it in the multimedia delivery decision-making process. Format independence is achieved using schema languages to describe the details of any given content or metadata, so that declarative mapping rules can be specified for translating from format-specific data points to format-independent concepts that are directly used by the framework. The system utilizes the criterion of "semantic-distortion", as an extension of Rate-Distortion Optimization based multimedia delivery. Several short video clips were encoded using H.264/SVC scalable video coding, and Scalable-To-Lossless (SLS) audio coding and adapted to four target bit rates. Subjective tests found a 72% preference for those clips which had been adapted so as to devote more bandwidth to the semantically important parts of the content when compared with standard objective-based bit-rate adaptation.**

I. INTRODUCTION

Recent multimedia coding formats developed by MPEG and ITU-T such as Scalable Video Coding (SVC) [1] and Scalable-to-Lossless Coding (SLS) [2] 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 [3] or dynamic estimation of channel capacity [4]. However, users automatically associate many layers of meaning (semantics) to the content they consume.

Research in this field of multimedia semantic-recognition is extensive, and it remains a challenging problem. However, systems are now being devised to allow a computer to recognise semantic information within media content e.g. that a picture contains a landscape or a cityscape [5]. 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 popularly known as the Semantic Web—has provided languages such as the Resource Description Framework (RDF) [6] and Ontology Web Language (OWL) [7] which can be used to express concepts in such a way that "This picture has many buildings" may also imply "it is a cityscape", and "it contains man-made objects." Much of this current work is aimed at applications such as improving the relevance of multimedia search results.

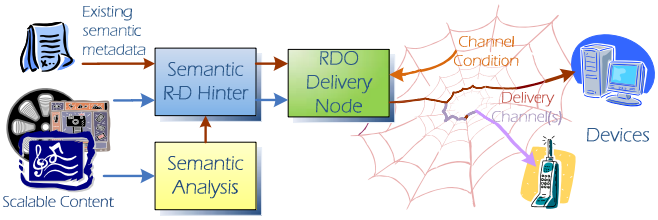

Fig. 1: A framework for semantic-aware multimedia delivery

While there have been some efforts to identify semantics for use in the delivery process [8-11], the framework recently proposed in [12] provides a solution for how to use this information within a practical system. This intelligent multimedia delivery system considers the contribution of the meaning of the content when deciding how best to deliver to a user. Such a system aims to ensure the user quality of experience, based on ensuring the intended meaning of the content, is maximized when making decisions on rate adaptation during media delivery.

Fig. 1 provides an illustration of the semantic-aware multimedia delivery system. Scalable content and semantic metadata (generated through semantic analysis or from existing knowledge) is fed into a semantic based Rate-Distortion (R-D) hinter. The outcomes of this hinter are used in the Rate Distortion Optimisation (RDO) stage, which considers both semantic information as well as objective criteria based on e.g. channel conditions to decide on how best to deliver the media content. To be effective in this environment, a semantic-aware delivery framework must support content that is encoded in any current, or future, format. This is provided through use of a format independent multimedia delivery language such as described in [13].

Sections 2, 3 and 4 of this paper provide a detailed description of the RDO delivery node, semantic R-D hinter and semantic analysis stages of Fig. 1, respectively. Section 5 describes experimental testing and results for evaluating the proposed system for adapting the delivery of video content to a user while conclusions are presented in Section 6.

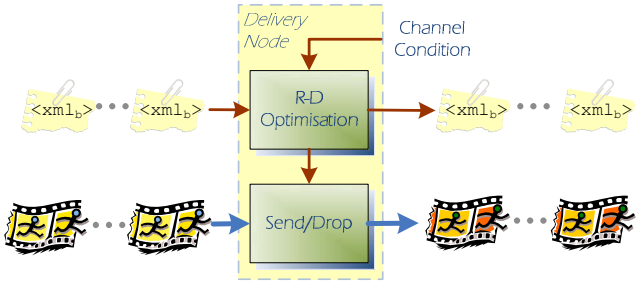

Fig. 2. A delivery node used content hints to perform R-D optimization.

II. RDO Delivery Node

The delivery node, illustrated in more detail in Fig. 2, is left only to decide whether and when to forward, drop or truncate each packet. That decision is made on the basis of some type of rate-distortion optimization algorithm, which takes as its inputs feedback about the channel condition, and metadata from the semantic hinter.

There are a number of rate-distortion optimization algorithms. Examples include the approaches by Chou [4], Chakareski [14], Eichhorn [15] and Cranley and Murphy [11]. Typically, these algorithms base rate-distortion optimization on error-probability cost functions, where errors are characterised in such a way as to encompass bit error rate, packet loss, and delay (such that the packet is too late to be useful). Furthermore, the formulation of distortion often considers 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.

Different algorithms perform better in particular scenarios and so the framework described in this paper does not prescribe one method over another. Instead, the framework allows the most suitable algorithm(s) to be implemented on any given delivery node. Hence, the proposed system simply prescribes the derivation of metadata about the media, including dependencies, packet sizes and distortion increments that can be utlised in a chosen R-D optimization algorithm along with information about the channel characteristics.

III. Semantic R-D Hinter

As proposed by Chakareski [14], the Rate-Distortion Optimization (RDO) is performed offline by a hinter, minimizing the amount of computation that must be done by the real-time delivery node. Fig. 3 depicts the proposed architecture of a semantic-R-D hinter, based on [12]. The output of the R-D hinter is metadata and media segments that are input to the delivery node for RDO. This metadata can be stored in a file (such as an ISO [16]or Quicktime [17] container) for later use, or transmitted with the content to a local or remote delivery node. The hinter itself is composed of elements that analyze the semantics and the syntax of the content.

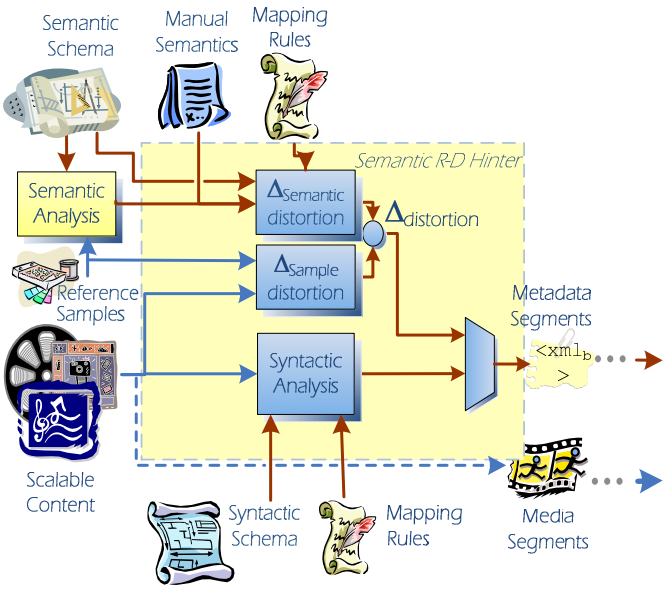

Fig. 3. The semantic hinter computes R-D metadata based on content syntax and semantics

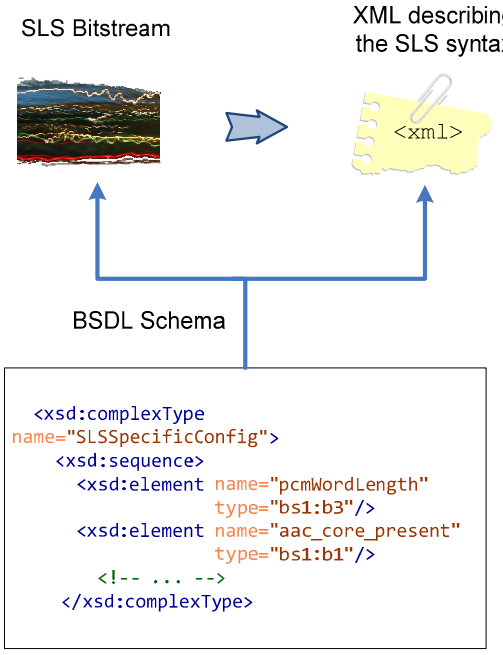

Fig. 4. A binary schema exposes the bitstream structure

The former (semantic analysis) is described further in Section IV. Format impendence is provided by the syntactic analysis stage. The hinter in Fig. 3 provides for Semantic Distortion to be combined with sample distortion, where decoded samples are compared to the samples that were originally encoded, using a measure such as Peak Signal to Noise Ratio (PSNR).

A. *Syntatic analysis*

Syntactic analysis extracts the interdependency, temporal and scalability metadata that are direct parameters of the compressed bitstream. The underlying syntactic structure of the content is exposed so as to provide access to the internal

data fields. In this work, the Bit Stream Syntax Description Language (BSDL) [18] is used to achieve this functionality. The BSDL schema describes the structure of an XML representation of the syntax of the binary bitstream. A simple example is illustrated in Fig. 4, which shows part of the BSDL schema representation of a SLS audio bitstream. A detailed description of BSDL is beyond the scope of this paper and further information can be found in [18].

The metadata exposed by using binary schema will be specific to a particular format (e.g. SLS, Flash, SVC). In order to use this metadata in a format independent 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 metadata needed by the RDO algorithm being used. The list of metadata required will vary depending on the particular RDO algorithm. Here, metadata is generated to describe three key items: 1. unique segments defined here as Data Units (e.g. a Picture Parameter Set (PPS) of a compressed video bitstream); 2. Interdependencies between Data Units (e.g. due to motion compensated frames of a compressed video bitstream), and; 3. Temporal relations between Data Units (e.g. timing information for a compressed video frame). Here, the Data Units, used as the atomic unit for R-D optimization, are defined similarly to Chou [4], where a packet on a network can contain at most one data unit. Such metadata is typically used in the RDO algorithms described in Section II.

IV. SEMANTIC ANALYSIS

The aim of semantic analysis is to generate metadata that can subsequently be reasoned on to compute Semantic Distortion.

B. 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 news story). This remains a challenging problem, with many efforts directed toward approaches that can expose various specific semantics of media content. For the purposes of evaluating the system proposed in this paper, the semantic metadata has been obtained through an asynchronous process that analyses the uncompressed media content through manual annotation. However, the system is designed to allow processing of any desired metadata, e.g. Flickr/Youtube tags, iTunes song ratings or existing binary formats such as ID3 [19], for instance.

C. 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 "error" 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).

Clearly, this notion of semantic distortion is highly subjective (as indeed are most 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 optimizing the delivery of the content semantics, as demonstrated through the subjective testing described in Section 5 as well as [8][11].

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. For example, 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 (assuming users are native English speakers). This rule covers both spoken communication (in which case the SD is associated with the audio track(s)), and visual communication (eg subtitles; where the SD is applied to the video). In this paper, a series of simple rules were determined for use in the system evaluated in Section V.

D. Combination of Semantic Distortion with sample distortion

This is pivotal to the correct operation of the R-D optimization algorithm. Chou [4] considers sample distortion to be additive, that is, the overall distortion is a large initial value less the sum of reductions in distortion due to receiving a set of L packets. This is described in (1).

$$D(\boldsymbol{\pi}) = D_0 - \sum_l \Delta D_l \prod_{l' \le l} (1 - \varepsilon(\pi_{l'})) \quad (1)$$

In (1), π_l is the transmission policy for a data unit l, $\boldsymbol{\pi}=(\pi_1,..,\pi_l)$ is the vector of transmission policies for each data unit, D_0 is the initial distortion, ΔD_l is the reduction in distortion due to receiving data unit l, $\varepsilon(\pi_l)$ is the probability that data unit l does not arrive.

However, the sample distortions used in (1) 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. Here 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 normalization based on potentially unknown response curves for distortion algorithms (sample and semantic). This may also be the case when multiple rules (potentially from independent sources) match a segment of content, leading to a need to aggregate further values that could have differing scaling. Finally, multiplicative combination retains a known zero point. This is important if either sample or semantic distortion has a zero value; 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.

V. SUBJECTIVE EVALUATION OF THE PROPOSED SYSTEM

This section describes the methodology, tested system and subjective results used to evaluate the proposed system.

A. Methodology

Double-blind, randomized subjective testing was used to validate that the proposed system successfully utilised Semantic Distortion to improve the quality of multimedia delivery. The scenario used for these tests was a mobile environment where channel characteristics are often highly variable, and also handset capabilities mean that audio and video require relatively similar bandwidth. As such, the source material was encoded at a sampling rate 22.05 kHz 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 (eg 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 SD 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 [44], including the conditions prescribed in table 412. Pairwise Comparison (PC) was used to evaluate the hypothesis that: "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." There were four news clips plus an initial (hidden) training clip. 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. These clips were chosen as they provide a range of semantic variability e.g. scenes corresponding to the studio introduction by the anchor and other scenes with footage of an event (often with commentary overlaid on audio from the event)

The audio from each clip was encoded using Scalable to Lossless Coding (SLS) [1] with an AAC base layer of 6kbps to provide a large scalable range. Scalable Video Coding (SVC) [1] 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 optimization (see, for example Cranley [11]) but it was decided to limit the sources of variability for the present experiment. In that regard, no attempts were made at error concealment, even though this would have an impact on a user's perception of a real world system employing SD.

B. Tested system

Semantic analysis for each clip was conducted manually to provide semantics indicating the language of communication (spoken or written), among other things. Mapping rules were created for these to describe how particular semantics relate to SD. Syntactic analysis was conducted using a BSDL Schema for SLS and another for SVC, then an XSLT stylesheet to expose the necessary semantics. Delivery optimization 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 optimization routine. Essentially, the algorithm used was:

1) Segment the clip into regions of constant SD (note that more than one rule may be matched at once);
2) For each section:
   a) Aggregate SD separately for audio and for video, according to the behavior;
   b) Apportion the target bandwidth between the audio and video stream according to the aggregated SD;
   c) Truncate each SLS frame so as to achieve the apportioned bit-rate; and
   d) Drop SVC NALUs to most closely approximate the target 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. Each set of clips was encoded with different ranges, resulting in total average bit rates ranging from approximately 24 kbps to 95 kbps over all 12 clips. For each semantic-aware clip produced using this algorithm, a reference clip was created with the same average bit rate. 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 Fig. 5 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. Regions of high SD for the audio relative to the video results in higher audio bit rates, while regions of lower SD for the audio relative the video result in lower audio bit rates. It can also be seen that both audio tracks have the same total average rate.

E. Results

The subjective test results are shown in Fig. 6. In total, 72% of the semantic-aware clips were preferred by subjects when compared to the average-rate reference clip. Of the twelve pairs, one semantic-aware clip was rated as worse than its average-rate partner. Another two were voted as no better and no worse, and the remaining semantic-aware clips were

preferred 84% of the time. This demonstrates that SD is of significant benefit in the multimedia delivery process. Moreover, the proposed system is effective in processing SD and R-D optimization-related metadata. In contrast, however, the results also suggest that the use of semantic distortion to optimize the apportionment of bandwidth between audio and video streams may not be beneficial for a minority of content, at least without more sophisticated optimization algorithms. However, while the modal trade-offs employed for this content fails to yield an improvement, it is quite possible that other uses of SD may give the desired results. Further investigation of this is left to future work.

VI. Conclusions

This paper 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 minimizing distortion of the intended meaning of the content. At the same time, the framework provides the flexibility to incorporate new semantics, optimization algorithms, and content formats as they become relevant. This process can operate largely without the addition of new software or hardware components, 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 band-width 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 as worse than its average-rate partner. Another two were voted as no better and no worse, and the remaining 9 semantically adapted clips were preferred 84% of the time. This demonstrates that SD is of significant benefit in the multimedia delivery process.

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 as well as ontological representations of semantic distortion.

Acknowledgment

This work was partially funded by the CRC for Smart Internet Technology. The test material used in the experimental work was used with permission from SBS World News Australia.

References

[1] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable video coding extension of the H.264/AVC std.,” *IEEE Trans. Circuits and Sys. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[2] Rongshan Yu, Rahardja, S., Lin Xiao, Chi Chung Ko, "A fine granular scalable to lossless audio coder," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.14, no.4, pp.1352-1363, July 2006.

[3] C. Timmerer et al., “Digital Item Adaptation - Coding Format Independence,” in *MPEG-21*, I. Burnett et al., Eds. Wiley, Chichester, UK., 2006.

[4] P. Chou, “Rate-distortion optimized streaming of packetized media,” *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 390–404, 2006.

[5] M. Naphade et al., “Large-scale concept ontology for multimedia,” *IEEE MultiMedia Magazine*, vol. 13, no. 3, pp. 86–91, 2006.

[6] D. Beckett, “RDF/XML Syntax Specification (Revised),” http://www.w3.org/TR/rdf- syntax-grammar/, 2004.

[7] M. Dean and G. Schreiber, “OWL Web Ontology Language Ref.,” http://www.w3.org/TR/owl- features/, 2004.

[8] M. Bertini et al., “Semantic adaptation of sport videos with user-centred performance analysis,” Multi- media, IEEE trans. on, vol. 8, no. 3, pp. 433–443, 2006.

[9] M. Xu et al., “Event on demand with MPEG-21 video adaptation system,” in Multimedia, 14th ACM intl. conf. on , 2006, pp. 921–930.

[10] M. Baba et al., “Adaptive multimedia playout method based on semantic structure of media stream,” in Communications and Information Technology, IEEE Intl. Symp. on, 2004, pp. 269–273.

[11] N. Cranley and L. Murphy, “Incorporating User Perception in Adaptive Video Streaming Systems”, *Digital Multimedia Perception and Design,* G. Ghinea and S. Chen, eds.,, pp. 242–263, Idea Group, 2006.

[12] Thomas-Kerr, J., Burnett, I., Ritz, C., “Intelligent Multimedia Delivery? It's a question of semantics”, *7th Inter. Symp. on* Comms. and Info. Techs. (ISCIT2007), Sydney, Australia, pp. 473-478, October 16-19, 2007.

[13] Thomas-Kerr, J., Burnett, I., Ritz, C., “Format-Independent Rich Media Delivery Using the Bitstream Binding Language”, *IEEE Transactions on Multimedia*, vol.10, no.3, pp.514-522, April 2008.

[14] J. Chakareski et al., “RDhint tracks for low-complexity RDoptimized video streaming,” Proc. IntŠl Conf. Multimedia and Exhibition, vol. 2, pp. 1387–1390, 2004.

[15] A. Eichhorn, “Modelling dependency in multimedia streams,” *Proc. of the 14th ACM Inte. Conf. on Multimedia*, pp. 941-950, 2006.

[16] ISO/IEC, “14496-12, IT - Coding of audio-visual objects - Part 12: ISO base media file format,” 2005.

[17] Apple, “QuickTime File Format,” developer.apple.com/reference/QuickTime/, 2001.

[18] J. Thomas-Kerr et al., “Is That a Fish in Your Ear? A Universal Metalanguage for Multimedia,” IEEE MultiMedia, vol. 14, no. 2, pp. 72–77, 2007.

[19] M. Nilsson, “ID3 tag version 2.4.0 - Main Structure,” , 2000.

[20] U. Niedermeier et al., “An MPEG-7 tool for compression and streaming of XML data,” in Multimedia and Expo, IEEE Intl. Conf. on, 2002, pp. 521–524.

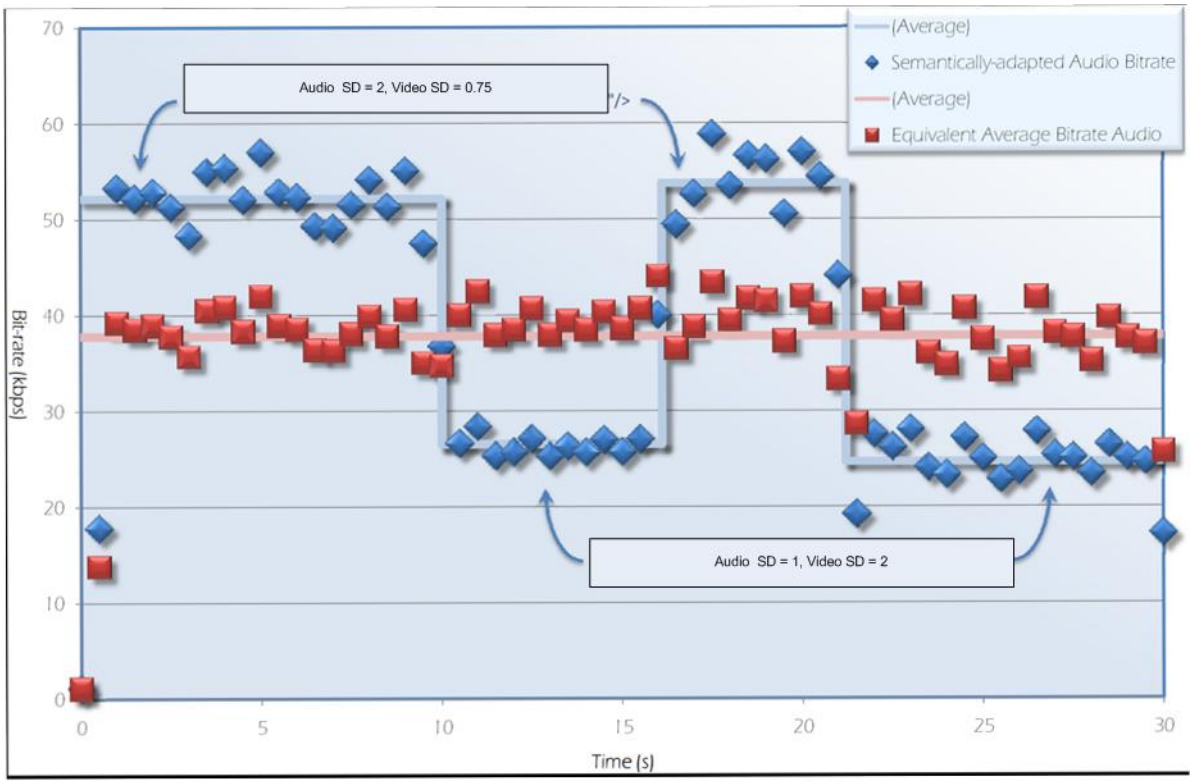

Fig. 5. Semantic adaptation diverts more bits to the portion of the content containing more of the meaning

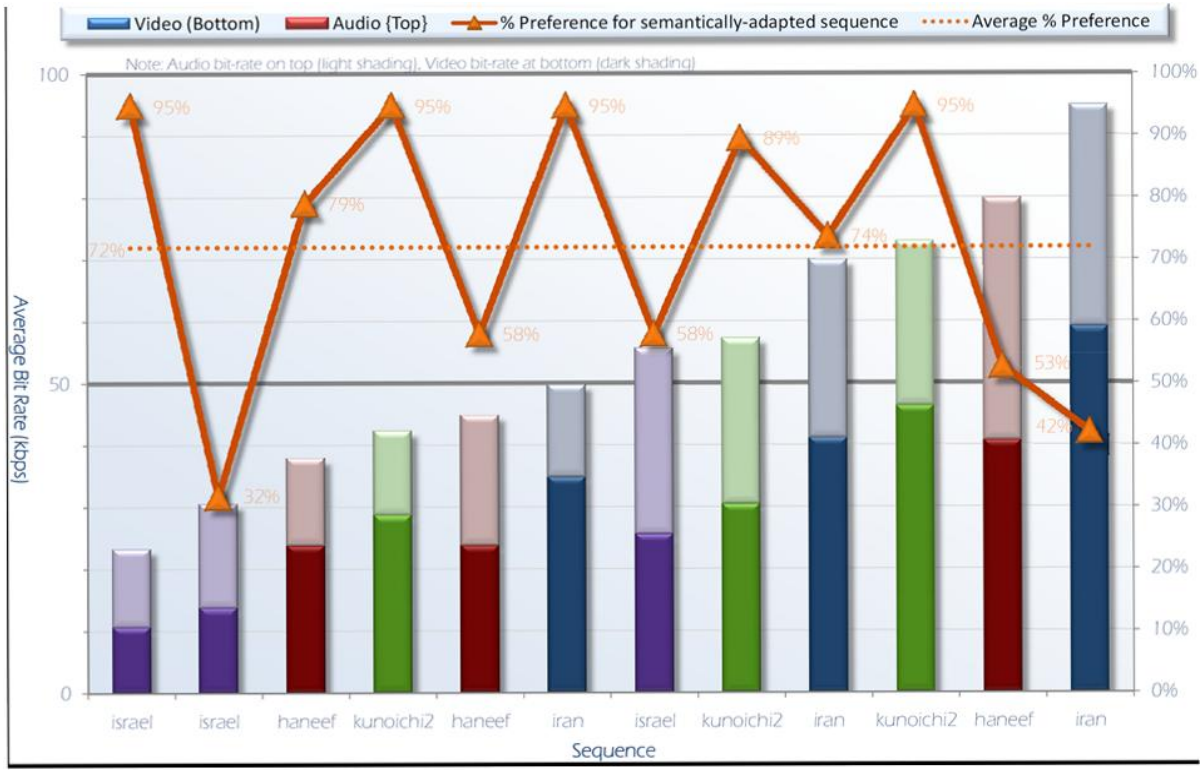

Fig. 6. Subjective testing shows a 72% preference for Semantically-aware multimedia delivery. On the histogram bars, audio bit-rate is represented by the top half (later shading), video bit rate is represented by the lower half (darker shading)