

Building MPEG-7 Transcoding Hints from Intrinsic Characteristics of MPEG Videos

Chia-Chiang Ho, Student Member, IEEE, Jin-Hau Kuo, and Ja-Ling Wu, Senior Member, IEEE

Abstract —MPEG-7 transcoding hints are defined to provide better quality or reduced complexity of transcoding applications. This paper describes application scenarios, hierarchies, and generation methods of transcoding hints that are based on intrinsic characteristics of MPEG videos.¹

Index Terms — MPEG-7 Transcoding Hints, MPEG characteristics.

I. INTRODUCTION

MPEG-7 defined media transcoding hints aiming at reducing complexity and improving the quality of general transcoding processes [1]. The basic assumption is that transcoding hints are generated in advance, and are stored together with the host video itself (or stored separately but with a proper linkage to the video). The possible applications of transcoding hints include but not limited to constant bitrate (CBR) to variable bitrate (VBR) conversion, encoding mode decision, frame rate and bitrate control, complexity reduction for motion re-estimation and video skimming and browsing.

Collected as one MPEG-7 description scheme (MediaTranscodingHint), transcoding hints defined in MPEG-7 include:

(i) Motion Hints:

MotionRange, consisting of four values (*xLeft*, *xRight*, *yDown* and *yUp*), indicates the recommended integer pel search range for conducting motion estimation to the left, right, bottom and top of a video frame, respectively;

Intensity describes the motion intensity, specifically, the variance of motion field, of a video segment;

Uncompensability describes the amount of new content in a video segment.

(ii) Shape Hints:

ShapeChange describes the amount of shape change (of a moving region) in a video segment;

NumOfNonTranspBlocks describes the average number of 16×16 blocks per frame containing at least one pixel with a non-zero alpha-map value. (Since shape hints are metadata used for object-based video coding and thus are not considered in this paper.)

(iii) Coding Hints:

AvgQuantScale describes the average quantization scale used to compress a video segment;

IntraFrameDistance describes the distance between intra-coded frames (I-frames);

AnchorFrameDistance describes the distance between anchor frames (I- or P-frames);

Difficulty describes the transcoding difficulty of one video segment, which is a normalized value (between 0 and 1), by taking video segments into consideration;

Importance describes the importance of one media segment with respect to the whole semantic media content. It is also a normalized value as the *Difficulty* hint does;

SpatialResolutionHint describes the maximum allowable spatial resolution reduction factor for preserving perceptibility.

Generality and flexibility are two central goals of MPEG-7 [2]. In MPEG-7, both generation and consumption of transcoding hints are not formally defined. The definition of “video segment” is also left for real applications. Focusing on MPEG videos, in this paper, we (1) clarify multimedia application scenarios within which transcoding hints are expected to be generated; (2) define practical transcoding hint hierarchies from both structural and semantic perspectives; (3) propose a compressed domain transcoding hint generator which takes advantages of intrinsic, low-level characteristics of MPEG videos. Our work provides a clear roadmap for implementations embedded on consumer electronics or multimedia softwares.

II. APPLICATION SCENARIOS AND HIERARCHIES OF TRANSCODING HINTS

Before devising hierarchies and generation schemes of transcoding hints, we would like to look into today's multimedia applications and figured out scenarios within which transcoding hints could be generated dependently. The success of MPEG-series standards makes MPEG videos scattering around the world. These videos reside in content archives of content providers, VCD and DVD disks, and personal collections (e.g., home videos), etc. Moreover, due to the widespread and cost reduction of digital camcorders, more MPEG videos are generated day by day. The generation of corresponding transcoding hints makes distribution and sharing of MPEG videos more easily and efficiently, even in today's heterogeneous network and computation surroundings. We classify MPEG video related applications into three different scenarios described as follows.

¹ This work was supported in part by the National Science Council, R.O.C., under the contract NSC91-2622-E-002-040 and 89-E-FA06-2-4-8.

The authors are with the Communications and Multimedia Laboratory, National Taiwan University, Taipei, Taiwan (email: {conrad, david, wjl}@cmlab.csie.ntu.edu.tw).

Scenario 1-- Raw Input, Compressed Output:

This scenario refers to those applications involving MPEG encoding. Examples are professional studio encoders, personal video recorders (PVR) and software MPEG encoders, etc.

Scenario 2-- Compressed Input, Compressed Output:

This scenario refers to those applications involving MPEG video processing. Examples include video post-processing, video editing and video content analysis for indexing, browsing, and/or searching. These applications take MPEG videos as inputs, and generate metadata and/or modified MPEG videos as outputs.

Scenario 3-- Compressed Input, Raw Output:

This scenario refers to those applications performing MPEG video decoding. Both hardware and software decoders, which decode MPEG videos coming from network or local storages, belong to this scenario.

With such application classifications in mind, we defined generic transcoding hint hierarchies for MPEG videos in an entity-relationship fashion, as shown in Fig. 1. Importance hint and spatial resolution hint are semantically determined through user interactions. The transcoding hint generator is responsible for generating other hints. Two styles of video representation are proposed. The *sequence-scene-frame* style coincides with human understanding of videos, and is the semantic one. The *sequence-GOP-frame* style coincides with internal structure of MPEG bitstreams, and is thus structural one. These two styles of video representation may help for improving efficiency and accuracy of transcoding in both systematic and user-oriented ways. The availabilities of different hints depend on the generation methods used in our transcoding hint generator, which will be illustrated later.

Another way to look at these hierarchies comes from the classification of hints, as given in Table I. Such classification affects the relationship between transcoding hints and video

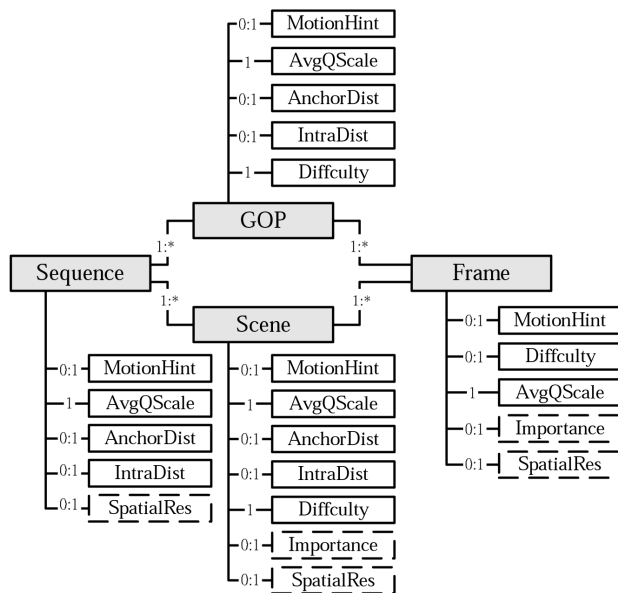


Fig. 1: The proposed transcoding hint hierarchy for MPEG videos. The names of hints have been abbreviated for clarity.

representations. For example, motion hints present both structural and semantic meanings. Associating motion hints with the structural video representation helps for those transcoders requiring re-estimation of motion vectors, while associating motion hints with semantic video representation helps for transcoders thinking semantically. This is because motion hints themselves somewhat stand for intrinsic importance of each individual scene, rather than those importance values set externally by the user (maybe the content provider).

The proposed video representations and associated transcoding hint hierarchies can be easily implemented in XML forms and by using the MPEG-7 defined binary representations.

III. THE PROPOSED COMPRESSED DOMAIN TRANSCODING HINT GENERATOR

Applications belonging to the aforementioned scenarios 2 and 3 process the MPEG compressed videos for various usages, such as playback, browsing, editing and transcoding. For these applications, we proposed a compressed domain transcoding hint generator that is built upon an MPEG decoder, as shown in Fig. 2. All information required for generating video representations and transcoding hints is available just after performing the VLC decoding module. That is, no de-quantization, IDCT or motion compensation operations are involved, and thus the generator operates very efficiently.

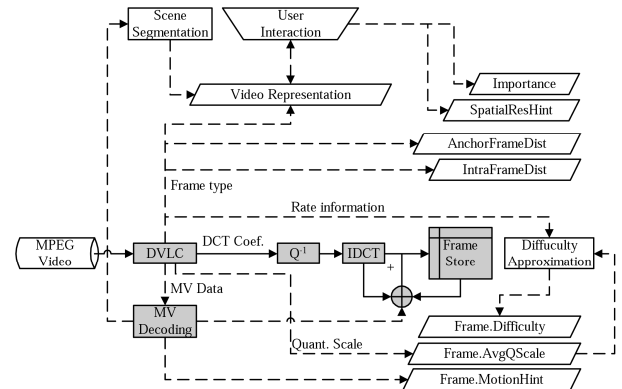


Fig. 2: The architecture of the proposed compressed domain transcoding hint generator. The gray-shaped modules constitute a typical MPEG decoder.

Table I: Classifications of transcoding hints

		Structural	Semantic
Motion Hints		Yes	Yes
Coding Hints	AvgQuantScale	Yes	Yes
	IntraFrameDistance	Yes	No
	AnchorFrameDistance	Yes	No
Difficulty		Yes	Yes
Importance		No	Yes
SpatialResolutionHint		No	Yes

The scene segmentation module divides the entire video sequence into separate scenes. We adopt a scene change detection scheme based on macroblock type statistics in B-frames [3]. The structural video representation is figured out by inspecting frame coding types and GOP information embedded in the bitstream. Then, transcoding hints belonging to individual frames, GOP's, scenes and the whole video are generated using the algorithms described below.

A. Algorithms for generating different transcoding hints

Motion range hint: For each predicted frame, we firstly check if its reference frames are in the same scene. If not, the motion range hint is not calculated for this frame. Otherwise, the maximal and minimal values of forward and backward motion vectors are figured out and denoted as Min_{x_F} , Max_{x_F} , Min_{y_F} , Max_{y_F} , Min_{x_B} , Max_{x_B} , Min_{y_B} , Max_{y_B} , respectively. For P-frames, Min_{x_B} , Max_{x_B} , Min_{y_B} and Max_{y_B} are set to zero. To account for different frame distances between the predicted frame and its reference frames, the above values are normalized as:

$$\begin{aligned} NMax_{x_F} &= Max_{x_F} / FD, \\ NMin_{x_F} &= Min_{x_F} / FD, \\ NMin_{y_F} &= Min_{y_F} / FD, \\ NMin_{y_F} &= Min_{y_F} / FD, \\ NMax_{x_B} &= Max_{x_B} / BD, \\ NMin_{x_B} &= Min_{x_B} / BD, \\ NMin_{y_B} &= Min_{y_B} / BD, \\ \text{and} \\ NMin_{y_B} &= Min_{y_B} / BD, \end{aligned} \quad (1)$$

where FD is the frame distance between the current frame and the forward reference frame, and BD is the frame distance between the current frame and the backward reference frame. The frame-level motion range hint is computed as:

$$\begin{aligned} xLeft &= -MAX(NMin_{x_F}, NMin_{x_B}), \\ xRight &= MAX(NMax_{x_F}, NMax_{x_B}), \\ yUp &= -MAX(NMin_{y_F}, NMin_{y_B}), \\ \text{and} \\ yDown &= MAX(NMax_{y_F}, NMax_{y_B}). \end{aligned} \quad (2)$$

For the whole video, one GOP or one scene, the corresponding motion range hint is figured out by taking the maximal value of available frame-level motion range hints of all frames in the range.

Motion Uncompensability hint: For each predicted frame, we calculate the ratio of intra-coded macroblocks over total macroblocks. This is based on the general rule of typical MPEG encoders, that is, choosing intra mode when the least error of motion compensation is still too large. However, the definition of Uncompensability is not precise in the frame level. Owing to different frame distances and different coding modes, different frames have different “capabilities” of motion compensation in nature. So, we calculate the Uncompensability hint only for higher levels. For a GOP, a scene, or the whole video, the value of the Uncompensability hint is the average of the predefined ratio values of each predicted frame in the range.

Motion intensity hint: For each P-frame, we calculate the average intensity of forward motion vectors. For each B-frame, average intensities of both forward and backward motion vectors are calculated. Similar to the generation of the motion range hint, the frame distance normalization is also performed here. For B-frames, the average of normalized forward and backward average motion intensities is then calculated. The motion intensity of one GOP or one scene is the average of motion intensities of all frames in the range. Finally, a sequence-wide normalization is performed such that the motion intensity value of the most active GOP or scene is 1.0.

AvgQuantScale hint: The calculation of AvgQuantScale hint for each frame is straightforward. For each frame, the AvgQuantScale is set to be the average of quantization scale values of all coded macroblocks. For one GOP, one scene or the whole video, the AvgQuantScale value is the average of quantization scales of all frames in the range.

IntraFrameDistance and AnchorFrameDistance hint: These two hints are figured out for every GOP or scene, by checking frame coding patterns. If the coding pattern is not regular, they are just not available. If IntraFrameDistance or AnchorFrameDistance is not consistent over the entire video, it is not available in the sequence level.

Difficulty hint: We approximate coding difficulty of each frame as

$$Difficulty = Consumed_Bitcount * AvgQuantScale. \quad (3)$$

GOP-wide, scene-wide and sequence-wide normalizations are then performed to calculate Difficulty values of all GOP's and scenes.

IV. EXPERIMENTAL RESULTS

To validate the applicability of the proposed transcoding hint generator, we cascaded four novel test sequences, namely, “container”, “mobile”, “foreman” and “mother and daughter” into a single video sequence with totally 1200 frames (each scene is of 300 frames). Fig. 3 shows sample snapshots of these four scenes. Then MPEG-1 encoding was performed with the following settings: The number of frames in a GOP is 30 (so each scene spans ten GOP’s); the I/P frame distance is 3; the bitrate is 1,500,000 bps; the motion search ranges are ± 31 pixels. Let us call the generated MPEG video as “4Seq”. Table II shows the summarization of generated scene-level transcoding hints.



Fig. 3: Sample snapshots of the four scenes in the sequence “4Seq”.

Table II: The summarization of scene-level transcoding hints of the sequence “4Seq”.

		container	mobile	foreman	mother and daughter
Motion Hints	xLeft	31	31	31	31
	xRight	31	31	31	31
	yDown	31	31	31	31
	yUp	31	31	31	31
	Uncomp	0	0	0	0
	Inten	0.089	0.337	1.0	0.398
Coding Hints	avgQ	2.69	12.8	4.72	2.03
	intraF	30	30	30	30
	anchorF	3	3	3	3
Difficulty		0.199	1.0	0.368	0.170

As for the motion intensity hint, Table II reveals that the scene “foreman” has stronger motion activities than other scenes do, while the scene “container” has weakest motion activities. This coincides with the feeling of human observers. Fig. 4 shows the GOP-level motion intensity hints. The GOP with the highest motion intensity resides in the scene “foreman”, at the time that the man waved his arm to the right and the camera also moved rapidly.

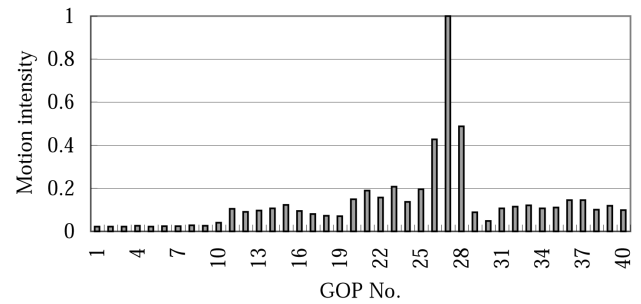


Fig. 4: The GOP-level motion intensity hints of the sequence “4Seq”.

The motion range hints of the four sequences were detected to be the same, i.e. 31, for all four directions. This means that at each scene, there is, more or less, some macroblocks with motion vectors exceeding the maximal allowable values (31 here). This can be verified by taking frame-level motion range hints of the scene “container” as an example, which is shown in Fig. 5.

The IntraFrameDistance and AnchorFrameDistance are correctly detected as 30 and 3, respectively. As for the average quantization scale, the scene “mobile” has the largest value while the scene “mother and daughter” has the least one. Since the adopted encoder utilizes a constant bitrate rate control algorithm, such diverse values of average quantization scales directly correspond to the values of the Difficulty hint (generated according to (3)). From Table II, it can be seen that the scene “mobile” is the most difficult to encode, while “container” and “mother and daughter” are comparably much more easier to encode. Fig. 6 shows GOP-level difficulty hints. It can be observed that GOP’s of the second half of the scene “foreman” have higher difficulty hint values than GOP’s of the first half do. This is because the second half of ‘Foreman’ is with larger motion and more complex textures.

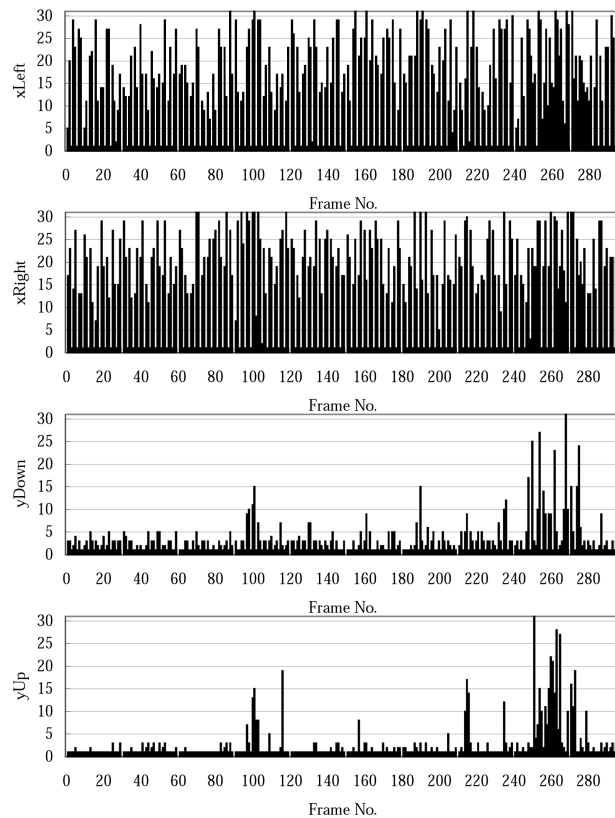


Fig. 5: The frame-level motion range hints of the scene "container" in the video "4Seq".

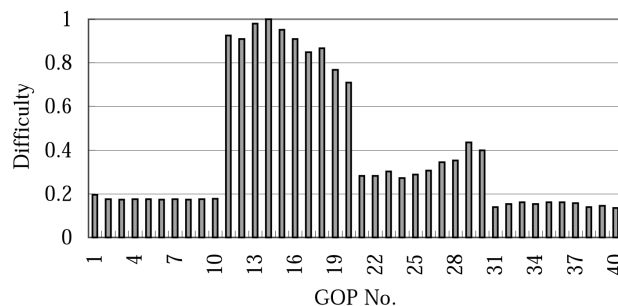


Fig. 6: The GOP-level difficulty hints of the sequence "4Seq".

V. APPLICATIONS

Due to the wide-spreading of DVD and the trend of HDTV/DTV, most existing video contents are still encoded in MPEG-2 form. Therefore, the MPEG-2-to-New video standard (e.g. AVC/H.264) transcoder or vice versa, is believed to be an indispensable functional module of future video signal treatments. As shown in [4], the main functionality addressed by the Media Transcoding Hint DS is to provide transcoding hints to support the transcoder. This transcoding hints enable the transcoder to reduce the computational complexity and to assist the transcoder in (re-) encoding parameter decisions to preserve visual quality through the transcoding process. The corresponding applications include the image resolution reduction for heterogeneous client devices (e.g. making use of importance and spatial resolution hint), video transcoder and real-time transcoder server for video streaming.

VI. CONCLUSION REMARKS

The advantages of using transcoding hints have been confirmed by experimental reports revealed in MPEG meetings. It is interesting that all transcoding hints are defined as optional, and this fact shows that the applicability of transcoding hints is very different for different media formats and target applications.

Defining transcoding hints is not to solve obstacles of transcoding, but to increase quality and/or to reduce complexity. An intelligent transcoder should have two alternative configurations: one is without the transcoding hint and the other uses transcoding hints wisely. Generating metadata to aid the task of transcoding is of course a positive idea. It is our belief that generating media-format-dependent metadata may help more for the task of transcoding than those general transcoding hints may do. The previous work done in our laboratory had transferred this idea into a practical and useful system [5].

REFERENCES

- [1] Information Technology-- Multimedia Content Description Interface, Part 5 Multimedia Description Scheme, ISO/IEC 15938-5, 2001.
- [2] S.-F. Chang, T. Sikora, and A. Purl, "Overview of the MPEG-7 standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 688-695, June 2001.
- [3] J.-H. Kuo and J.-L. Wu, "An efficient algorithm for scene change detection and camera motion characterization using the approach of heterogeneous video transcoding on MPEG Compressed Videos," *IEEE 3rd Int. Conf. On Information, Communications and Signal Processing*, Oct 2001.
- [4] Anthony Vetro, Peter Kuhn, Teruhiko Suzuki, John R. Smith, Ana B. Benitez, Charilaos Christopoulos, "CE Report on the Media Transcoding Hint DS," *ISO/IEC JTC1/SC29/WG11 MPEG/M6168*, July 2000.
- [5] P.-K. Hsiao, Y.-S. Tung, Ja-Ling Wu, K.-L. Huang and H.-S. Chen, "A frame-based MPEG characteristics extraction tool and its application in video transcoding," *IEEE Trans. Consumer Electronics*, vol. 48, no. 3, pp. 522-532, Aug. 2002.



Chia-Chiang Ho received the B.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1996. Currently, he is a Ph. D student in the same department. His research interest includes video coding, video streaming and multimedia system design.



Jin-Hau Kuo received the B.S. degree in information and computer engineering from Tatung University, Taipei, Taiwan, in 1997. He is currently a Ph.D. student in the department of computer science and information engineering, National Taiwan University. His current research interests include data compression, digital image processing, content-based retrieval, code optimization, and multimedia systems.



Ja-Ling Wu received the B.S. degree in electronic engineering from TamKang University, Tamshoei, Taiwan, R.O.C. in 1979, and the M.S. and Ph.D degree in electrical engineering from Tatung Institute of Technology, Taipei, Taiwan, in 1981 and 1986, respectively. From 1986 to 1987, he was an Associate professor of the Electrical Engineering Department at Tatung Institute of Technology, Taipei, Taiwan. Since 1987, he has been with the Department of Computer Science and Information

Engineering, National Taiwan University, where he is presently a Professor. He was also the Head of the Department of Information Engineering, National Chi Nan University, Puli, Taiwan, from Aug. 1996 to July. 1998.

Prof. Wu was the recipient of the 1989 Outstanding Youth Medal of the Republic of China, the Outstanding Research Award sponsored by the National Science Council, from 1987 to 1992, and the Excellent Research Award from NSC, 1999.

Prof. Wu has published more than 200 journal and conference papers. His research interests include algorithm design for DSP, data compression, digital watermarking techniques and multimedia systems.