https://doi.org/10.2352/ISSN.2470-1173.2019.11.IPAS-262

Parameter optimization in H.265 Rate-Distortion by single-frame semantic scene analysis

Ahmed M. Hamza; University of Portsmouth
Abdelrahman Abdelazim; Blackpool and the Fylde College
Djamel Ait-Boudaoud; University of Portsmouth

Abstract

The H.265/HEVC (High Efficiency Video Coding) codec and its 3D extensions have crucial rate-distortion mechanisms that help determine coding efficiency. We have introduced in this work a new system of Lagrangian parameterization in RDO cost functions, based on semantic cues in an image, starting with the current HEVC formulation of the Lagrangian hyper-parameter heuristics. Two semantic scenery flag algorithms are presented and tested within the Lagrangian formulation as weighted factors. The investigation of whether the semantic gap between the coder and the image content is holding back the block-coding mechanisms as a whole from achieving greater efficiency has yielded a positive answer.

Introduction

The current H.265/HEVC (High Efficiency Video Coding) standard contains several Rate-Distortion mechanisms within the encoding modules that are controlled by decision making processes of Coding Tree Unit (CTU) functions in the encoder. These choices, in the hierarchical fashion of the HEVC coding, ultimately determine which reference image segments are selected at each CTU partition, and how each Unit is partitioned, therefore playing a central role in the bitstream characteristics of the encoding signal.

The problem of parameter selection is of improving the overall efficiency of compression by optimizing these processes, where Lambda controls the balance of rate and distortion effects based on motion compensation cost functions that involve λ_{MOTION} and λ_{MODE}. Here, mode is the ultimate mode of block division for that Coding Unit (CU).

For each possible choice in lambda constituent cost-balancing parameters at any level (e.g., motion estimation, motion vector prediction) there is a corresponding, non-deterministic, effect on the higher level parameter-controlled functions that make the final coding decision:

$$\mathscr{J}_{MODE} = \mathbf{D}_{SSE} + \lambda_{MODE}\mathbf{R} \tag{1}$$

which in turn will affect partial distortion and rate of the overall picture, i.e., which blocks to derive residuals from for the transform modules of the encoding.

The simplified form derived for the Lagrangian multiplier *lambda* in [8] was:

$$\lambda(Q) = -\frac{d\mathbf{D}}{d\mathbf{R}} = c.Q^2 \tag{2}$$

with Q being the quantization value for the source. The relation is derived based on several assumptions about the source probability distribution within the quantization intervals, and the nature of the rate-distortion relations themselves (constantly differentiable throughout, etc.). The value initially used for c in the literature was 0.85. This was modified and made adaptive in subsequent standards including HEVC/H.265.

Our work here investigates further adaptations to the rate-distortion Lagrangian by semantic algorithms, made possible by recent frameworks in computer vision.

Adaptive Lambda Hyper-parameters in HEVC

Early versions of the rate-distortion parameters in H.263 were replaced with more sophisticated models in subsequent standards. Today, the Lagrangian is calculated based on several factors in addition to the QP/Q value, to take into account the frame and reference location in the Group of Pictures (GOP) being coded.

$$\lambda_{mode} = \alpha W_k 2^{(Q-12)/3} \tag{3}$$

A table of values for W_k and α is in the reference manual[5]. While this is dynamic, it is still a fixed-heuristic method that does not take into account the visual information semantics that describe the nature of the image and its parts. In this paper, we introduce a modification to the Lagrangian by replacing heuristic α values with our scene-based profiler values, which biases λ by visual content.

Parameter Optimization

Our scene understanding models aim to utilize high level semantic features in the parameter estimation of HEVC, in that the selection and refinement of rate-distortion parameters can be based on them and not on fixed heuristics alone.

The main intuition behind this is that different types of scenery imply different image characteristics, which can be informative to the rate-distortion optimization of groups of frames, single frames, and parts of an image.

Further to the semantic classification process is the task of tuning the control parameters chosen. The Lagrangian λ is necessarily linked to the quantization parameters due to the natural relationship between them, and the basic form obtained for HEVC as shown in Eq. 3. Since the quantization parameter is a transcoding factor linked to quantization modules in the encoder (which we do not seek to modify), our weight factoring α modification aims to reduce *on average* the bits needed for transmission of the CU collective across the entire sequence, at that set QP level.

Semantic Segmentation Basis

We experimented with several state-of-the-art frameworks as semantic inputs to our methods. With the success of convolutional "deep" nets at object and face recognition tasks, several frameworks such as these for image *segmentation* have gained traction as robust vision tools on a multitude of image contexts. The segmentation task is more difficult than just object recognition, which is a prerequisite. Segmentation implies a shape mask of objects detected, and additionally, *semantic* segmentation implies a differentiation between objects of the same class within the image.

MASK R-CNN [4] and SegNet [6], including Bayesian extensions to SegNet [1], are Convolutional Neural-Net based frameworks that have appended region-suggestion, entity classification, pixel-wise segmentation and entity differentiation layers of reasoning on top of successful object detection models in machine vision.

Modeling alpha on image semantics

The pixel-wise segmentation of images allows us to produce models for the hyper-parameters $alpha$ by meaningful visual interpretation in real-time.

In the current implementation [5] both α and W_k are weighting factors that never exceed 1.0. The values are carefully designed to be set at certain empirically-derived limits from the HEVC testing, and set depending on the coding mode used.

Alpha, from the reference manual, is set thus:

$$\alpha = \begin{cases} 1.0 - 0.05 \times \text{NumBPictures} & \text{If referenced} \\ 1.0 & \text{Non-referenced} \end{cases} \quad (4)$$

with the upper value clipped between 1.0 and 0.5. In other words, it will only decrease Lambda, and at most by half. The greater the number of referenced frames, the less emphasis on the decision bit-cost the Lagrangian cost function.

Our approach is to allow the alpha hyper-parameter to expand beyond the clipped 0.5-1.0 weight limits, temporally learning its value as a function of both it's normal setting and what we call the semantic flag values (which are all between 0 and 1.0).

$$\alpha_m = w_0 F_0 + w_1 F_1 \quad (5)$$

where the w_i values are learned for all semantic scene flag parameters F set during the scene discovery phase. A statistical learning process (we use a simple linear network) converges on the appropriate weighting scheme per scene model.

The task therefore consists of a) producing a semantic flag output for classes of image or image region defined, and b) allowing a pre-encoding step to learn the appropriate weighting of these contributions to $alpha$ for each 64x64 CTU.

The adaptive optimization process is based on adaptive elements from [2], where reinforcement learning with simple neuron-like linear models are used on the control problem. Our choice of reinforcement learning method was dictated by the fact that we cannot phrase the problem as purely supervised learning, in the sense that "correct" values are not known for individual learning elements at each discrete time step, but a reinforcement error signal can be obtained indirectly from the learning environment as a whole: in this case from encoding bit-rate/size values per frame.

The weight update mechanism is as follows:

$$w_i(t+1) = w(t) + \delta r(t).e_i(t) \quad (6)$$

where r is the reinforcement signal, δ is a positive real-valued constant for rate of change and e is the eligibility function, reduced to 1 in our case (as opposed to the trace function in [2]).

This is because in the set of experiments presented below, the potential input signal changes (i.e., flag inputs) happen at the per-frame level, which is when the reinforcement signal is generated, so no additional granularity of eligibility of update is needed. If we were updating weights per CTU/block however (using additional CU-level semantic flags), the bursts of eligibility windows may be needed.

Semantic Scene Models

We present here two full-picture descriptor models as semantic flags in Eq.5.

The first, F_0 is an indoor/outdoor confidence flag. This is set simply according to the percentage of segmented pixels falling into either class-group. This way, even if a scene is misclassified, the properties of respective entities are captured.

The second flag confidence value is based on urban landscape imagery having distinct image features that may inform the encoder process choice for lambda. This is detailed as follows.

Algorithm 1 Urban-outside flag determination process.

```
Ent ← SegmentationRun()
Confidence ← 0.0
flag_0 ← RunAlgorithm_0() //in/out-door flag procedure
if flag_0 < 0.3 then return -1.0 //Rule out strictly indoor scenes
end if
for all entities e in Ent do
    if e ∈ UrbanStructures() then
        rc ← rc + pixelCount(boundingRegion(e))
    else
        rc ← rc − pixelCount(boundingRegion(e))
    end if
end for
    return rc/getTotalViewPixels()
```

That is, if the overall majority of scene content belongs to an urban type-class, and the outdoors flag is set above 0.3 threshold (i.e., this is most likely outdoors), but with minimal human-or-animate presence, we can classify the scene to be urban large-scale.

Experimentation

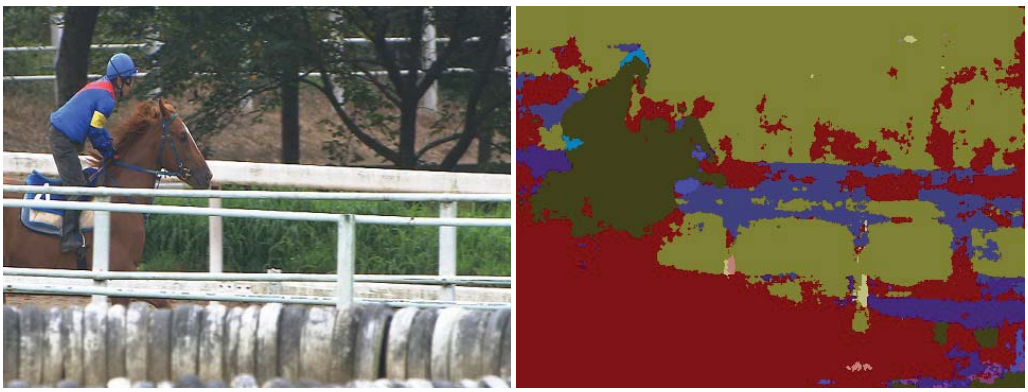
Our ultimate goal is coding efficiency so our test setup remains based on HEVC reference code[5] and test video sequences listed in Table 1.

We use several pieces of software in our experimental setup. For semantic and instance segmentation networks we use the SegNet[6, 1] deep network architecture trained on its own reference indoor and outdoor datasets and the CityScapes [3] data, so there is zero overlap between our own images and the training set.

Trained on the difficult indoor SUN RGB-D[7] semantic understanding benchmark suite, SegNet is able to segment well into

(a) Scaled input frame (b) Segmented image

Figure 1: A Kimono1 frame segmentation by SegNet. The class label assignments and region division present a successful differentiation of the frame regions with different visual characteristics. No image frames from HEVC test sequences were used to train this SegNet model.



(a) Scaled input frame (b) Segmented image

Figure 2: A representative frame from the Keiba HEVC test sequence, run through SegNet trained on the outdoor CamVid model. Not all class assignments match (CamVid classes are road-scene oriented) but the region division is successful.

the 37 object classes for our purposes, considering our overall task is less demanding than the pixel-wise boundary precision measures used in the vision challenges. Note that this particular indoor data set includes scene labeling data as well, but the semantic segmentation of SegNet and comparable architectures is geared towards segmenting entities, not scene understanding as a whole, and this is the output we work with.

Although image category precision results vary across weakly-trained and state-of-the-art methods, our generalized algorithms, with the purpose of *basic* scene understanding, are broad enough categorizations that tolerate pixel-level boundary precision hits to a larger extent than the challenges the training datasets (and their constituent parts) were developed for. For instance, even if 70% of an object's boundary pixels are correctly labeled, the overall effect is still positive for our algorithm in optimizing parameters.

Encoding Results

Semantically guiding the alpha parameter per frame leads to results shown below, averaged over 150 frames for all sequences. We use the Bjøntegaard delta bitrate (BDBR) method to measure the change in coding efficiency of the final quantized and transformed frames. Not shown here is the variation in bitrate gains and losses over each invididual frame in the sequence, which varies considerably by sequence.

Sequence Name	Resolution	Frame Rate
Traffic	2560×1600	30
BasketballDrive	1080p	50
BQTerrace	1080p	60
Cactus	1080p	50
Kimono	1080p	24
ParkScene	1080p	24
FourPeople	720p	60
KirstenAndSara	720p	60
RaceHorses	480p	30
BQMall	480p	50

Table 1: Test Video Sequence Details. We chose a variety of size/frame rate combinations containing indoor and outdoor scenarios for visual content.

Table 2 shows higher BDRATE efficiency gains in the intra mode slices (usually the first frame in the GOP per HEVC design), where there is no temporal redundancy being exploited by the encoder, and the decisions affected are in the SATD and SSE cost functions for the prediction (rough) and final mode choice

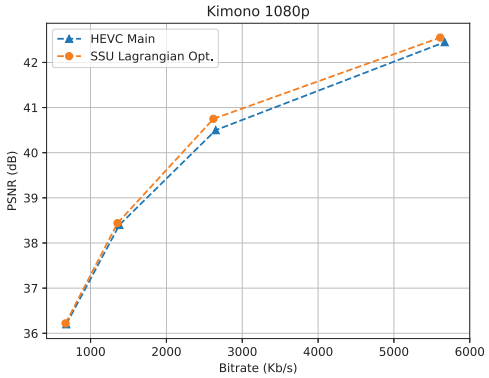


Figure 3: PSNR/bit-rate quality curve comparison between our optimized process with Scene Understanding Unit, and the reference HM, for the Kimono1 sequence.

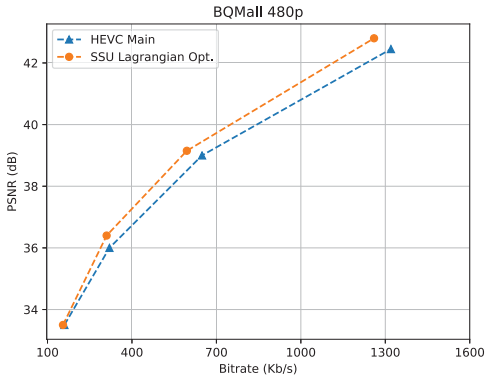


Figure 4: PSNR/bit-rate quality curve comparison between our optimized process with Scene Understanding Unit, and the reference HM, for the Kimono1 sequence.

Sequence Name	Intra BDBR (%)	Inter BDBR (%)
Traffic	-4.1	-3.0
BasketballDrive	-2.2	-0.8
BQTerrace	-2.1	-3.2
Cactus	-5.6	-5.1
Kimono	-5.2	-2.4
ParkScene	-4.7	-2.3
FourPeople	-2.6	-1.1
KirstenAndSara	-4.3	-6.0
RaceHorses	-4.3	-3.7
BasketballDrill	-6.4	-5.5

Table 2: Test video sequence encoding results. Both intra and inter (B-slice) coding benefits form training the composite, semantically weighted α, and its effect on the Lagrangian. Test points conducted at QP = 22, 27, 32, 37.

stages, respectively.

Discussion

An additional system of parameter optimization can be designed involving deeper granularity for lambda. In this case, we seek to modify our model even within the same picture, by region, or by individual CTU/coding block. This is quite unusual for hybrid coding schemes in the modern line of standards, which generally make RDO parameter decisions without variation in cost-function Lagrangians within the same image.

The rationale here is that different parts of the coded image carry different characteristics in terms of lighting and movement, therefore a semantic discovery of these regions can give grounds to different base weighting schemes, unlike the currently used heuristics by frame in HEVC, and our own modifications of them by semantic weighting of scene content in this work.

Our results have favored scene variation sequences that involve a large number of moving textures in non-background entities, especially outdoor scenes. For instance the FourPeople sequence showed little improvement overall in either INTRA or INTER modes.

Conclusion

We have introduced in this work a new system of Lagrangian parameterization in RDO cost functions, based on semantic cues in an image, starting with the current HEVC formulation of the Lagrangian hyper-parameter hueristics. The investigation of whether the semantic gap between the coder and the image content is holding back the block-coding mechanisms as a whole from achieving greater efficiency has yielded a positive answer: the rate-distortion regularization by a semantically-weighted Lagrangian λ does indeed improve performance over the fixed heuristics of image hierarchy level currently employed in reference standard encoders, albeit more so in intra coded frames.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[2] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, smc-13, September 1983.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.

[5] JCT-VC. HEVC reference software 16.0 [ONLINE]. `https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/`, 2016.

[6] Alex Kendall, Vijay Badrinarayanan, , and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[7] S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)*, 2015.

[8] Thomas Wiegand and Bernd Girod. Lagrange multiplier selection in hybrid video coder control. In *IEEE International*

Conference on Image Processing (ICIP 01), Thessaloniki, Greece, September 2001. IEEE.

Author Biography

Ahmed Hamza is a PhD student in the School of Computing at the University of Portsmouth. He obtained a M.S in Computer Science from Georgetown University in 2010, where he worked on algorithms in chem-informatics. His interests are in Video Coding and optimization, algorithms, information theory, and natural systems. Ahmed is a member of IET and IEEE.

Abdelrahman Abdelazim is a Curriculum Manager Engineering at Blackpool College. He holds a BEng (Hons) degree in Digital Communication and a PhD degree in Engineering, both from the University of Central Lancashire (UCLAN), Preston, UK. Between 2008 and 2012 he worked as Lecturer in Electronics within the School of Computing, Engineering and Physical Sciences (CEPS). From 2012 to 2017 he was Associate Professor and Head of Department at the American University of the Middle East (AUM). His experience includes leading the implementation of a number of academic and industrial digital communication projects. His research interests are in the area of reducing the complexity of Video Coding Encoders in real-time Scalable and Multi-view applications and the area of teaching and learning in higher education. Abdelrahman is member of the IET since 2006, and he established the Kuwait community in 2013. He is a chartered engineer and a Fellow of the Higher Education Academy.

Djamel Ait-Boudaoud joined the University of Portsmouth in 2010 and is currently Professor and Dean of the Faculty of Technology. Before Portsmouth he was the head of the School of Computing, Engineering and Physical Sciences at the University of Central Lancashire for close to 10 years. His research interests are predominantly focused on the problems of optimisation with applications in 3D computer vision, video standards (H264) and solving combinatorial (ordered-sequence) problems using evolutionary algorithms. He gained a PhD in 1991 from the Department of Electrical and Electronic Engineering at University of Nottingham, UK., is a Chartered Engineer (CEng) and a fellow of the Institution of Engineering and Technology (FIET).

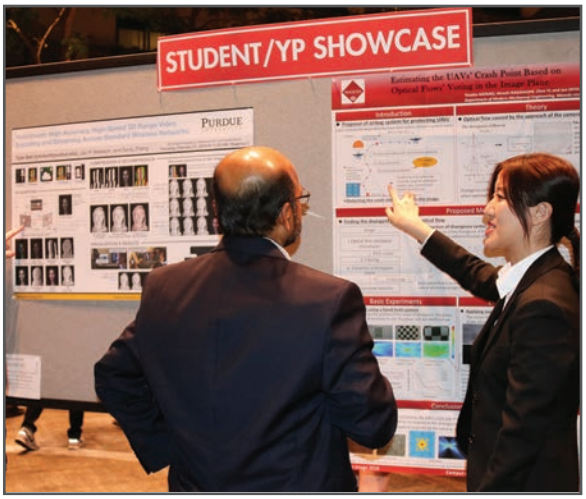
JOIN US AT THE NEXT EI!

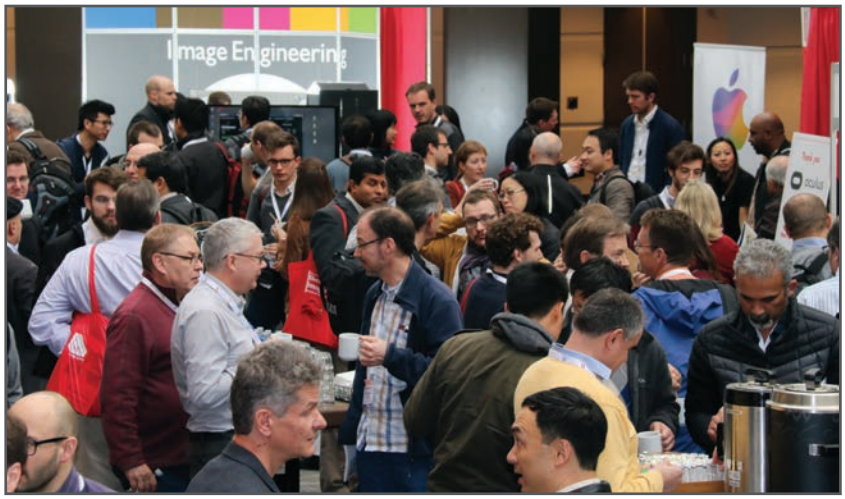
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org