TOWARDS PERCEPTUALLY GUIDED RATE-DISTORTION OPTIMIZATION FOR HEVC

Kais Rouis[1,3]*, Mohamed-Chaker Larabi*[2]*, Jamel Belhadj Tahar*[3]

[1] National School of Engineering of Tunis, University of Tunis El Manar, Tunisia
[2] CNRS, Univ. Poitiers, XLIM, UMR 7252, Poitiers, France
[3] NOCCS Laboratory, National School of Engineering of Sousse, University of Sousse, Tunisia

ABSTRACT

This paper proposes a novel approach for perceptually guiding the rate-distortion optimization (RDO) process within the High Efficiency Video Coding (HEVC) standard. The reference codec does not consider effectively the perceptual characteristics of the input video and further, the particular perceptual sensitivity of each coding tree unit (CTU) inside a frame. The corresponding frame-level Lagrangian multiplier depends only on the quantization parameter. Inspired by the mechanisms of the human visual system, the proposed solution is a CTU-level adjustment of the standard Lagrangian value based on a set of complementary measured features. These measures rely on the spatial and temporal analysis of the current CTU in the frequency domain. Based on perceptual quality indices and Bjontegaard delta measurements, over several resolutions of tested video sequences, the proposed method demonstrates a promising coding performance according to the rate-distortion compromise.

Index Terms— HEVC, RDO optimization, Lagrangian adjustment, perceptual features.

1. INTRODUCTION

The significant progress in video coding technologies is a counterpart of the great development of multimedia communications. The increasing demand of data storage and transmission bandwidth capacity influenced this progress to meet the challenge required by the popularity and the large use of different services such as TV broadcasting, Internet video streaming, consumer electronics, etc. On the other hand, in most video applications, the end-user's quality of experience became a constant and hard constraint for content providers.

A new generation of video coding standard namely the High Efficiency Video Coding (HEVC) [1] was jointly released by the ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG). It permits to double the data compression ratio compared to the former H.264/MPEG-4 AVC for similar level of video quality [2]. A variety of new tools have been proposed for this standard such as intra and inter prediction mode algorithms, an effective optimization process and especially the quadtree decomposition of the largest coding unit (LCU) that could achieves the size of 64×64. Fig. 1 shows the adopted coding tree unit (CTU) and the use of prediction blocks, to provide an idea about the coding structure in HEVC.

To achieve higher coding efficiency, rate-distortion optimization (RDO) [3] is typically used at the encoder side to select the mode providing the best rate-distrotion (R-D) tradeoff. The aim of RDO is to minimize a distortion D at a target rate R_T, which can be described as :

$$\min\{D\} \quad \text{s.t.} \quad R < R_T \,. \tag{1}$$

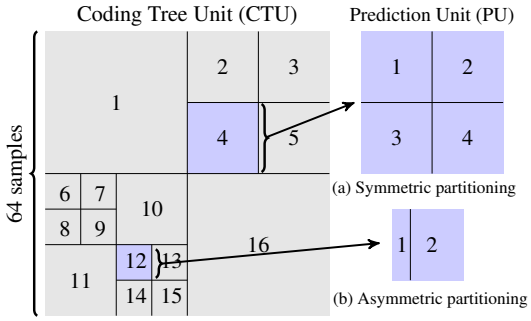

Fig. 1: Example of CTU partitioning into CUs and PUs for CTU of 64×64 together with PU partitioning (a and b). Asymmetric partitioning is used for inter prediction mode only.

The Lagrangian multiplier factor is employed to solve the aforementioned optimization problem by transforming it to an unconstrained form:

$$\min\{J\} \quad \text{where} \quad J = D + \lambda R \tag{2}$$

where J is the Lagrangian cost function and λ is the so-called Lagrangian multiplier. Usually, the value of λ can be experimentally determined as it represents the slope of the R-D curve. In [3], it can be practically determined by the quantization parameter (QP) used for encoding. The QP values are also refined according to encoding structures.

Considering the bitrate constraints, most of video compression applications tend to provide an optimized perceptual quality as a substitute of reducing the factual distortion between compressed and original frames. For instance, a measurement was performed in [4] to assess the sensitivity of each CTU and each frame to guide the bit allocation, relative to the regions' perceptual sensitivity. Other proposed compression schemes are based on texture analysis and synthesis approaches as in [5]. A perceptually adaptive Lagrangian multiplier was proposed [6] based on perceptual characteristics of the video content, where the Lagrangian is determined according to the temporal activity and the spatial energy factors. A similar approach in [7] was recently proposed to adapt the Lagrangian multiplier for each CTU, using a set of extracted perceptual features. This latter will be used for comparison with our proposed scheme.

In this work, we propose to explore the spatio-temporal visual characteristics of each CTU (See Fig. 1) in a frame, and taking into account salient objects under different scales and angular directions. The structural and salient information, which mimic the human visual system (HVS), are estimated using a set of accurate measurements. The HEVC standard Lagrangian multiplier is accordingly adjusted at the CTU level.

The rest of the paper is organized as follows. In Section 2, we describe the different components of the proposed adjustment scheme. We discuss in Section 3 the adopted quality assessment process which would be useful for coding efficiency validation. The experiments described in Section 4 help to evaluate the R-D compromise between the rate and video quality. Finally, the last section provides some conclusions about the proposed work.

2. PROPOSED R-D LAGRANGIAN ADJUSTMENT

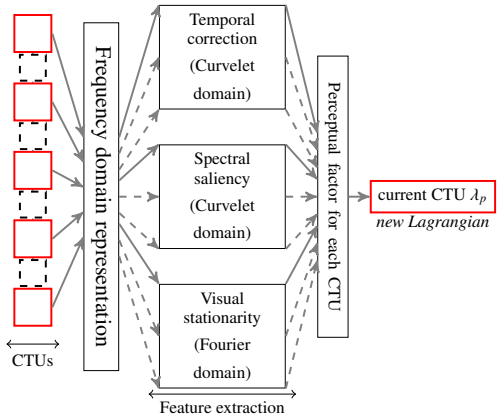

Fig. 2: Proposed Lagrangian adjustment scheme.

The proposed solution is based on a set of features that will be used to adjust the HEVC Lagrangian multiplier. Fig. 2 presents the general scheme of the adjustment process. As shown in this figure, the coding tree units are transformed in the frequency domain. The HEVC encoder operates on quadtree coding blocks going from 64×64 to 8×8 samples. As a result, the contained textural patterns are captured under different resolution depths. Dealing with the CTU as the largest unit does not prevent from investigating these patterns at different scales and angular directions. It is known that the curvelet transform uses parabolic tuned functions that ensure an accurate representation of the signal singularities. So, in the context of this work, we use a convenient discrete version of the curvelet transform called UDCT (uniform discrete curvelet transform) [8]. It provides a set of angular windows along a particular scale resolution. In our implementation, six directional subbands over two scales are extracted. Besides, we rely on a robust use of the CTUs power spectrum according to the Fourier shift property [9]. The displaced areas across frames could be detected through a correlation function, which describes the sharpness between two co-located blocks.

The RDO decisions will be guided separately for each CTU in the current frame. A new CTU-level Lagrangian value is then defined by combining the following three complementary information having an important impact on the visual quality. 1) *Temporal correction* (T_c): The new coding structures of the HEVC codec raises a strong dependency between coding blocks within successive frames. Hence, it appears reasonable to estimate the structural information of the current CTU based on the co-located one of the previous frame. The latter has been evidently altered by compression distortions. 2) *Spectral saliency* (S_l): Predicting salient information is effective to investigate whether the current CTU contains visual objects that attract the human perception. This information is very important in the RDO process in order to prevent significant distortions to the salient CTUs. The saliency has been predicted in the frequency domain using an adaptation of the method described in [10]. 3) *Visual stationarity* (P_c): The stationarity of visual scenes is basically related to the displaced areas/objects between successive frames. In other respects, the aim is to identify the displacement accorded to the relevant CTUs. It is a frame-level feature, measured in the Fourier domain and following the phase correlation concept [11].

These aforementioned complementary features, described in the next sections, are combined and used in order to refine the standard λ_{HEVC} in the RDO process as shown by the following equation:

$$\lambda_p = \underbrace{\left(\frac{P_c \,.\, T_c}{1 - \alpha S_l} \right)}_{P_f} \times \lambda_{HEVC} \,. \tag{3}$$

The perceptual factor (P_f) is used to reflect the perceptual importance of the CTU. In order to adjust the effect of P_f on the refinement of λ_{HEVC}, we use α as a weighting factor. For the experiments of this paper α is set to 0.5.

2.1. Temporal correction

Object shapes and curves compose the visual structural information that could be affected when increasing the QP levels. Therefore, we suggest to constrain the compression of the current CTU by that of the previous reconstructed frame (co-located). The co-located CTU is decomposed using the UDCT defined by shaped basis functions.

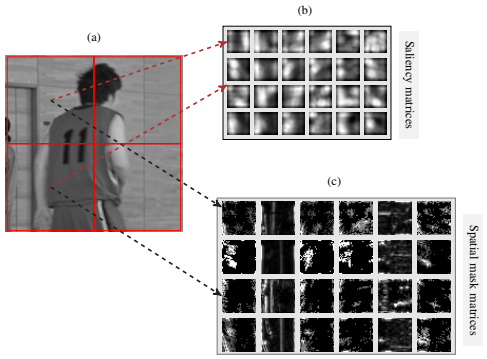

Fig. 3: Extracted matrices from the subbands' coefficients within the first and second scales of the UDCT decomposition (six orientations) [8]. Each row (right) corresponds to 64×64 block samples from a frame region example (left).

Fig. 3(a) shows four neighboring CTUs of a sequence frame (Luma component). The UDCT subband coefficient magnitudes are shown in Fig. 3(c), which will be used to extract a spatial mask. Each row corresponds to a single CTU (as indicated by the black dashed arrows), presenting the six orientations of the applied decomposition. The coefficient magnitudes' scatter is displayed from each subband matrix (32×32 coefficients).

Thus, we define a spatial mask as the maximum of six coefficient magnitudes at the same location:

$$M(x,y) = \max \left\{ \left| B^{\theta}(x,y) \right|,\ \theta = [1, ..., 6] \right\} \tag{4}$$

where B^{θ} is a directional subband at the *finest scale*. The mean information T_m is obtained as follows:

$$T_m = \text{mean}\Big\{ M(x,y),\ \forall\, x,\ y \in [1,2,..,32] \Big\}. \quad (5)$$

This is further performed for each CTU of the previously reconstructed frame. The proposed feature T_c describes the structural information contained in the considered CTU to be used to have an impact on the compression decisions of the current coding unit. It is given by:

$$T_c = \frac{T_m(i)}{\overline{T_m}} \quad (6)$$

where i is the index of the co-located CTU, $\overline{T_m} = \frac{1}{K}\sum_{j=1}^{K} T_m(j)$ is the mean T_m over the K CTUs in the frame.

2.2. Spectral based saliency

In complex scenes, objects could be presented under different scales and saliency strengths. In order to account for this diversity, the current CTU is decomposed using the UDCT basis functions. With the aim to define the saliency feature in the frequency domain, we only keep the matrices of the *coarsest scale* provided by the six angular windows.
The log-spectrum representation is obtained in a similar fashion to the model in [10]:

$$L(f) = \log\ (\ 1 + \left|\ \mathfrak{F}_w^{\theta}\ \right|\ ) \quad (7)$$

where $\mathfrak{F}_w^{\theta}$ represents the transformation of the current CTU, according to an angular window of direction θ. $|.|$ is the absolute value operator.

For a given position (x, y) and a given direction θ, the spectral saliency is defined as follows:

$$S^{\theta}(x,y) = \Big|\ \mathfrak{F}^{-1}\Big(\ \exp\Big[\ D(f) + i\,\angle\mathfrak{F}_w^{\theta}\ \Big]\ \Big)\Big|. \quad (8)$$

$\angle\mathfrak{F}_w^{\theta}$ is the phase distribution and $\mathfrak{F}^{-1}$ is the inverse Fourier transform. $D(f)$ is the residual information obtained using a local average filter $h(f)$ and defined by:

$$D(f) = L(f) - h(f) * L(f). \quad (9)$$

At this stage six saliency matrices are obtained corresponding to the six adopted directions $\theta = [1, ..., 6]$. These matrices are illustrated in Fig. 3(b). The spectral based saliency feature S_l, used in the refinement of λ_{HEVC} is defined as the average over (x, y) of the maximum saliency values over the six directions, as given by Eq.10.

$$S_l = \frac{1}{16}\sum_{x,y}\max_{\theta}\Big(S^{\theta}(x,y)\Big). \quad (10)$$

2.3. Visual stationarity

The properties of phase correlation are appropriate to describe the displacement of coding blocks between adjacent frames. Let consider two CTUs namely t and c, where c is a replica of t, but shifted by x_0 and y_0 such that $f_c(x, y) = f_t(x - x_0, y - y_0)$, where f_c and f_t are two integrable functions. The shift property of the Fourier transform states that $F_c(u, v) = F_t(u, v)e^{(-j2\pi wd)}$, $d = (x_0, y_0)$, $w = (u, v)$. Extracting the power spectrum phase gives the normalized cross spectrum $R(u, v) = e^{(-j2\pi wd)}$. Typically, the phase correlation is then defined as the IFT of $R(u, v)$ and the peak location will be $\text{argmax}\{\mathfrak{F}^{-1}[R(u, v)]\}$. The respective Fourier transforms of these blocks are identical in magnitude in the shifted coordinate points, but they differ in phase which is a function of the relative translation.

With the aim to formalize the visual stationarity feature P_c, the displacement between block c and its co-located t (original samples) is measured as given in Eq. 11. The phase correlation is calculated by transforming both blocks into the Fourier domain and utilizing the shift property:

$$\Psi = fshift\Big|\mathfrak{F}^{-1}\left(\exp^{j(\angle F_t - \angle F_c)}\right)\Big| \quad (11)$$

where F_c and F_t are respectively the Fourier transformed blocks of the current CTU and its temporal equivalent and $\angle$ is the phase of the corresponding coefficients. $fshift$ means the Fourier shift operation. Here our interest is focused on the maximum (peak) of the phase correlation matrix defined as $P_m = \text{argmax}_{(x,y)}(\Psi)$.

Some CTUs will have notable values in comparison to the remaining ones that are almost uniform. To take this observation into account, we split the frame CTUs into two sets : the first is composed of the 20% highest P_m values and the other with the remaining 80%. Consequently, P_c is defined as:

$$P_c = \frac{\overline{P_{mh}}}{\sqrt{\overline{P_{ml}}}}. \quad (12)$$

$\overline{P_{mh}} = \frac{1}{K_1}\sum_{i=1}^{K_1} P_m(i)$ is the mean over the K_1 CTUs of the first set (20% of highest P_m values), and $\overline{P_{ml}} = \frac{1}{K_2}\sum_{j=1}^{K_2} P_m(j)$ is the mean over the K_2 remaining CTUs.

3. PERCEPTUALLY DISTORTION MEASUREMENT

It is important to describe the approaches chosen for measuring the perceptual quality of the processed video sequences in our experimental part. We opted for the perceptual weighted-mean squared error (PW-MSE) quality metric [12]. This metric was proposed particularly to consider compression distortions. It uses the contrast sensitivity function with the aim to remove the imperceptible errors lying in the high frequencies of the error signal. Besides, the masking effect of HVS is predicted using a randomness map considering spatial correlations. PW-MSE demonstrated a higher performance over several efficient metrics.

Although the traditional PSNR metric is based on the MSE, it may not be accurate to reveal these factors and is not able to perceptually estimate the quality degradation resulting from frame rate reduction. Note that quality indices of the PW-MSE, named in the following by PMSE, are converted to decibel units for 8-bit data. In the corresponding paper [13], the log-function was applied on the average of the modulated squared error SE_M. We used similarly the following equation:

$$PMSE = 20\,.\,\log\left(\frac{255}{\sqrt{\overline{SE}_M}}\right).$$

Additionally, the structural information impairments are measured by the commonly used and well-known SSIM metric [13]. It still have a high importance when dealing with the perceptual task.

4. EXPERIMENTAL RESULTS

Comprehensive experiments were conducted in order to evaluate the feasibility of the proposed approach and its impact on the coding efficiency. The scheme was implemented on the top of the HEVC test model (HM16.12). Experiments were performed using the Random Access (RA) configuration (main profile) and the common test condition (CTC) specified by Joint Collaborative Team on Video Coding (JCT-VC) [14]. Moreover, we considered the Low Delay (LD) configuration (main profile) with an IBBB structure setting. Different sequences were used in the simulation, including resolutions ranging from WVGA (832×480) to WQXGA (2560×1600). Moreover, the RDO was enabled and the frame level QP was set to values between 22 to 37 with an interval of 5.

Table 1: Summary of the compression efficiency results with HEVC-HM16.12 as anchor.

Sequence	Resolution	Low Delay (%)		Random Access (%)	
		BD-Rate (PMSE)	BD-Rate (SSIM)	BD-Rate (PMSE)	BD-Rate (SSIM)
BQTerrace	1080p	-29.06	-24.35	-24.53	-22.21
BasketballDrive	1080p	-10.98	-11.96	-5.85	-7.97
KristenAndSara	720p	0.06	-3.89	-3.20	-4.24
FourPeople	720p	0.96	-4.45	-1.01	-9.76
RaceHorses	WVGA	-6.37	-6.24	-1.55	-2.12
PartyScene	WVGA	-0.56	-5.38	-3.47	-4.25
Traffic	WQXGA	-9.47	-19.47	-4.05	-11.35
Average	————	**-7.91**	**-10.82**	**-6.23**	**-8.84**

Considering the PMSE as the distortion measure, the BD-rate savings of the proposed method can be observed from Table 1. Our method performs consistently better under Random Access encoding structure than the anchor codec. Even though the coding efficiency slightly drops for 720p sequences using the LD structure, the average over the set of test sequences is up to 7.91%. We can clearly notice the particular results of BQTerrace sequence (1920×1080) with a BD-rate saving of up to 29.06% and 24.53% for LD and RA configurations, respectively. Fig. 4 shows the R-D curves in terms of PMSE, for BQTerrace and BasketballDrive sequences.
The same observations hold for the SSIM index, with the exception of better performance with regard to 720p sequences, and higher BD-Rate savings especially for Traffic and PartyScene. The variation between the results based on PMSE and SSIM could be explained by the fact that PMSE is an error based measure even though it is including the notion of contrast sensitivity and contrast masking. In turn, SSIM is based on a finer description of the perceptual features and focuses on the variation of the structural content as it would be perceived by the HVS. Overall, the results are relatively close based on both metrics which results in a reliable perceptual video coding.

In [7], a similar approach was proposed to adapt the Lagrangian multiplier for each CTU, using a set of extracted perceptual features. The authors validated their approach using the SSIM as a quality index. To settle a fair comparison with this work, we used the same test model (HM10.0) and test conditions [15]. Table 2 shows the superior efficiency of our method for the same test sequences (as in Table 1).

5. CONCLUSION

In this paper, we proposed an approach for the HEVC Rate distortion optimization aiming at adjusting the Lagrangian using a set of features based on the human perceptual mechanisms. The impairments that occur within a CTU could be controlled based on the underlying spatial and temporal correlations. The frequency domain analysis presents a convenient framework in this sens, from which we described the visual complexity and salient information. Guiding the RDO process was carefully completed when incorporating the extracted features to be properly assembled. The compression efficiency results were very convincing taking into account the achieved bitrate savings. When compared to a similar approach, our proposed method performed better.

Table 2: Comparison of the compression results with the approach of Yang et al. [7] using the same test set and HEVC-HM10.0 as anchor (average of the percentages over the tested sequences).

	Low Delay (%)	Random Access (%)
	BD-Rate (SSIM)	BD-Rate (SSIM)
[7]	-5.81	-5.13
Proposed	**-11.79**	**-8.34**

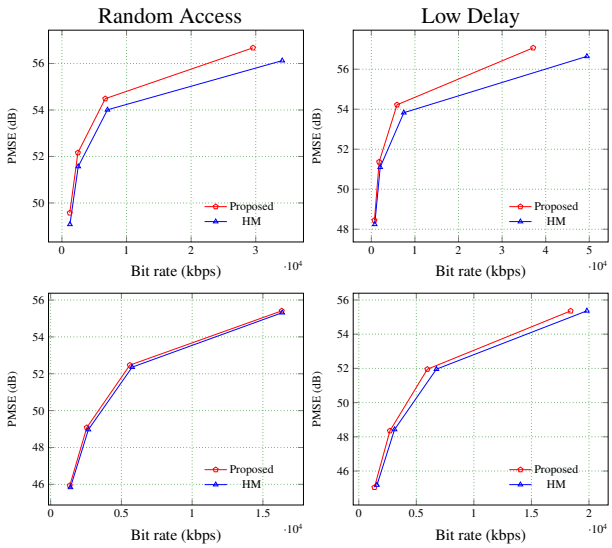

Fig. 4: R-PMSE curve comparison against HEVC anchor under the CTC of BQTerrace (top), and BasketballDrive (bottom).

6. REFERENCES

[1] Gary J Sullivan, Jens Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] J-R Ohm, Gary J Sullivan, Heiko Schwarz, Thiow Keng Tan, and Thomas Wiegand, "Comparison of the coding efficiency of video coding standardsincluding high efficiency video coding (HEVC)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684, 2012.

[3] Gary J Sullivan and Thomas Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[4] Huanqiang Zeng, Aisheng Yang, King Ngi Ngan, and Miaohui Wang, "Perceptual sensitivity-based rate control method for high efficiency video coding," *Multimedia tools and applications*, vol. 75, no. 17, pp. 10383–10396, 2016.

[5] Fan Zhang and David R Bull, "A parametric framework for video compression using region-based texture models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1378–1392, 2011.

[6] Huanqiang Zeng, King Ngi Ngan, and Miaohui Wang, "Perceptual adaptive lagrangian multiplier for high efficiency video coding," in *Picture Coding Symposium (PCS), 2013*. IEEE, 2013, pp. 69–72.

[7] Aisheng Yang, Huanqiang Zeng, Jing Chen, Jianqing Zhu, and Canhui Cai, "Perceptual feature guided rate distortion optimization for high efficiency video coding," *Multidimensional Systems and Signal Processing*, vol. 28, no. 4, pp. 1249–1266, 2017.

[8] Truong T Nguyen and Hervé Chauris, "Uniform discrete curvelet transform," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3618–3634, 2010.

[9] Pallab Kanti Podder, Manoranjan Paul, and Manzur Murshed, "Fast coding strategy for HEVC by motion features and saliency applied on difference between successive image blocks," in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2015, pp. 175–186.

[10] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[11] CD Kuglin, "The phase correlation image alignment method," in *Proc. International Conference on Cyber-netics Society*, 1975, pp. 163–165.

[12] Sudeng Hu, Lina Jin, Hanli Wang, Yun Zhang, Sam Kwong, and C-C Jay Kuo, "Compressed image quality metric based on perceptually weighted distortion," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5594–5608, 2015.

[13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] Ken McCann, C Rosewarne, B Bross, M Naccari, K Sharman, and GJ Sullivan, "High efficiency video coding (HEVC) test model 16 (HM 16) improved encoder description," *Jonit Collaborative Team on Video Coding, JCTVC-S1002, Strasbourg, FR*, 2014.

[15] F Bossen, "Common test conditions and software reference configurations, Joint Collaborative Team on Video Coding (JCT-VC) of ITUT SG16 WP3 and ISO," Tech. Rep., IEC JTC1/SC29/WG11, Doc. JCTVC-J1100, Stockholm, Sweden, 2012.