DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

HEVC saliency map computation

Marwa AMMAR, Mihai MITREA*, Ismail Boujelben*, Patrick Le Callet***

* *Institut Mines-Télécom; Télécom SudParis, ARTEMIS Department ; UMR 8145 – MAP5*

***Polytech'Nantes, IRCCyN*

Abstract

This paper investigates whether the information related to the human visual saliency is still preserved at the level of the HEVC compressed stream syntax elements. In this respect, a new saliency model, matched to the peculiarities of this emerging standard is defined. It consists of four elementary maps, describing the four main saliency features: intensity, color, orientation and motion. These maps are defined based on the energies of the luma and chroma coefficients, on the variations of the intra prediction modes and on the energy of motion vectors, respectively. They are fusioned according to 48 static and static-dynamic pooling formulas. The results are compared to three state-of-the-art uncompressed (pixel) domain as well as to the MPEG-4 AVC compressed domain saliency maps. It is brought to light that the HEVC saliency model outperforms (with singular exceptions) the state-of-the-art uncompressed domain and is as good as MPEG-4 AVC saliency model. We can thus state that, as its MPEG-4 AVC ancestor, although not designed based upon visual saliency principles, the HEVC compression standard preserves this human visual property at the level of its syntax elements.

1 Introduction

Visual saliency maps already proved their efficiency in a large variety of image/video communication, covering from selective compression and channel coding to watermarking.

Such saliency maps are generally based on individual saliency maps, computed from different characteristics (like color, intensity, orientation, motion, …) extracted from the pixel representation of the visual content. These individual maps are subsequently pooled into a global saliency map.

Some previous studies also took the challenge of extracting the visual saliency directly from transformed (MPEG ASP) or even compressed (MPEG-4 AVC) domains.

By addressing the HEVC (High Efficiency Video Coding) stream, the present paper resumes and extends our previous work [1], [2] devoted to the definition and evaluation of a saliency map solely extracted from the MPEG-4 AVC stream syntax elements.

The HEVC saliency map definition is structured at three levels.

First, the HEVC stream syntax elements are investigated according to their a priori potentiality to be connected to the visual saliency. Note that, in this respect, the extension from MPEG-4 AVC to HEVC is not straightforward. On the one hand, HEVC allows different block sizes to be defined; consequently the energy conservation theorem, invoked in the MPEG-4 AVC intensity and color map definitions, is reconsidered and adapted to this new applicative configuration. On the other hand, both intra and inter prediction modes are changed, thus imposing a detailed investigation on the orientation and motion maps. The inter prediction modes are now structured into two classes (AMVP – advanced motion vector prediction and merge modes) thus making a priori the motion saliency detection dependent on the encoding configuration.

Secondly, a total of 48 fusion formulas [1] (6 for combining static features: color advantage, orientation advantage, intensity advantage, mean, maximum, and multiplication; and for each of them, 8 to combine static to dynamic features: mean, maximum, multiplication, skewness, binary threshold, motion priority, dynamic weight and scale invariant) are benchmarked.

Third, in order to evaluate the coherency between the above-defined saliency map and the ground-truth fixation map obtained over the eye tracking data, we considered the density fixation maps, obtained for 80 s of video, and averaged over 30 human observers [3]. For each map, two metrics of two types are considered: distribution-based metrics (the Kullback-Leibler Divergence – KL-D) and location-based metrics (area under the ROC curve - AUC).

2 State of the art

The saliency detection is a research field in which several sound studies have already been advanced for uncompressed images [4], [5], [6], [7], and videos [8], [9], [10], [11], [12], [13].

Since current day visual content is preponderantly stored and exchanged in compressed formats and as the full decompression process is not only time consuming but computation expensive as well, models computing the saliency directly from compressed images (namely JPEG [14]) and videos (namely MPEG-4 ASP [15] and MPEG-4 AVC [1], [2]) have also been proposed.

In order to extract the saliency maps, Fang et al [14] no longer considers pixel representation of the image but a transformed domain related to the JPEG compression. The features (intensity, color, texture) are directly extracted from the 8×8 JPEG discrete cosine transform (DCT). In order to extract the intensity and color maps, the JPEG native YCrCb transformed color space is translated into the RGB transformed color space, and then the intensity and color features are extracted according to the Itti's principles (the Y channel represents the luminance component, while Cr and Cb represent the chroma components). The texture feature is given by the AC coefficient in YCrCb color space. The global saliency map is obtained through a so-called coherent normalized-based fusion method, i.e. through a weighted addition of the elementary maps. The experimental results are obtained on 1000 images and correspond to the AUC (Area Under the ROC Curve) between the human fixation and the saliency maps;an average AUC value of 0.93 is obtained and shown to be larger than the values corresponding to six other state of the art studies.

To the best of our knowledge, [15], [1], [2] are the first studies devoted to video saliency detection in the compressed domain.

Fang et al proposes a saliency detection model in MPEG-4 ASP. This model uses DCT coefficients of unpredicted frames (*I* frames)

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

to get static features and predicted (*P* and *B* frames) to get motion information. *YCrCb* color space is used in MPEG-4 ASP video bitstream. The *AC* coefficients represent texture information for image blocks. The motion vectors are then extracted to get the motion feature. The combination of the static and the motion features is then applied based on a dynamic fusion. The experimental results are obtained on 50 video sequences and correspond to calculate the KLD and the AUC between the saliency map and the fixation map at saccade locations; it is shown that this model is validated by a $KLD = 1.828$ and $AUC = 0.93$.

In our previous work [1][2], we proposed a saliency detection model in MPEG-4 AVC stream (before the entropic coding). The first step consists in extracting the elementary saliency maps: the intensity and color maps are related to by energy of the luma and chroma residual coefficients, while the orientation is given by the gradient of the intra prediction modes. The motion is extracted as the amplitude of the inter motion vectors. The second steps ensured the statistical coherency for each individual map: the outliers are eliminated, the elementary maps are normalized to belong to a dynamic range of [0, 1] and an average filtering with fovea size kernel is applied to each map. The third step consists on pooled the obtained saliency maps. It requires both computing the static saliency map as a combination of the intensity, color and orientation maps, and then combining it to the motion saliency map so as to obtain the MPEG-4 AVC saliency map. In [2], the applicative performances are evaluated under a robust m-QIM watermarking framework [16]. The video corpus consists of 6 videos sequences of 20 minutes each. For prescribed data payload (of 30, 60, and 90 bits/second) and robustness (BER of 0.07, 0.03, and 0.01 against transcoding, resizing and Gaussian attacks respectively), the saliency information increases the transparency by an average value of 6 dB in PSNR, 0.003 in NCC and decrease the DVQ (Digital Visual Quality) by 1400. This experiment was not considering the color saliency map. In [1], the comparison between the above-defined map and the ground-truth is done by considering a corpus of 8 video sequences. The differences between the MPEG-4 AVC saliency map and the EyeTracker density fixation maps are evaluated by computing the Kullback-Leibler divergence and the AUC (area under the ROC curve). The KLD results in an average value of 0.83 and the AUC is an average value of 0.84.

3 Method presentation

3.1 HEVC overview

The nowadays emerging HEVC (High Efficiency Video Coding) standard brings improvements over MPEG-4 AVC, so as to increase the compression capabilities, especially for high resolution videos [17].

HEVC offers larger and more flexible prediction and transform block sizes, greater flexibility in prediction modes, more sophisticated signaling of modes and motion vectors and larger interpolation filter for motion compensation.

HEVC video sequences are structured, in the same way as MPEG4-AVC, into Groups of Pictures (GOP). A GOP is composed of an *I* (intra) frame and a number of successive *P* and *B* frames (unidirectional predicted and bidirectional predicted, respectively). The *I* frame describes a full image coded independently, containing only references to itself. The unidirectional predicted frames *P* use one or more previously encoded frames (of *I* and *P* types) as reference for picture encoding/decoding. The bidirectional predicted frames B consider in their computation both forward and backward reference frames, be they of *I*, *P* or *B* types.

A frame in HEVC is partitioned into coding tree units (CTUs), which each covers a rectangular area up to 64x64 pixels depending on the encoder configuration. Each CTU is divided into coding units (CUs) that are signaled as intra or inter predicted blocks. A CU is then divided into intra or inter prediction blocks according to its prediction mode. For residual coding, a CU can be recursively partitioned into transform blocks.

As in our previous work [1], saliency maps are obtained by first computing elementary saliency maps and then performing their post-processing and pooling. We extract the saliency map only from *I* and *P* frames.

3.2 HEVC elementary saliency maps

When defining our saliency map, we consider that the luma residual coefficients which represent the difference between the current block and its neighborhood would provide same intensity information corresponding to the one obtained by the center-surround difference at the still image intensity map. In our previous work [1], the intensity map in MPEG-4 AVC video stream is defined by computing the energy luminance for each 4x4 luma transform block.

Such a technique would not be appropriate in the context of a varying transform block sizes as in HEVC, where several transform block sizes are supported: 4x4, 8x8, 16x16 and 32x32. The basic transform coding process of the prediction residual in HEVC is very similar to that of MPEG4-AVC. It is based on integer DCT basis functions, except for 4x4 luma transform blocks, in which case a DST-based transform is performed.

To compute the intensity saliency map from HEVC video stream, two steps are required. We first compute the luminance energy of the transform block (TB) and then we calculate the luminance energy of each 4×4 region inside the TB.

We first extract the transformed and quantified luma coefficients for each TB directly from the compressed stream. By applying the energy conservation property between DCT or DST transformed and spatial domain, the luminance energy of a TB is computed according to:

$$M_{TB} = \sum_{i}^{s} \sum_{j}^{s} {Y_{ij}}^2 \quad (1)$$

Where *s* is the size of TB, *i* and *j* are coefficient coordinates and *Y* is the luma residual coefficient.

We calculate the luminance energy of a 4×4 region inside TB as following:

$$M_i(k) = M_{TB}/N \quad (2)$$

Where *k* is the 4x4 region index in the frame and *N* is the total of 4x4 regions in TB. The intensity conspicuity map will be obtained by displaying M_i where the highest values represent the salient blocks.

3.2.1 Color map

Thorough analogy to the way in which the intensity saliency was defined, color saliency will be based on color energy.

In our previous work [1], we first extract from the compressed MPEG-4 AVC video stream chroma residual coefficients. The color information (Cr,Cb) is then used to calculate the two opponent color pairs RG (Red/Green) and BY(Blue/Yellow). Finally, we compute the color saliency map as the sum of the energy in the double color-opponent RG and BY space. For the

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

same reason as for intensity map, this technique is not appropriate with HEVC stream.

The chroma TB size of HEVC is half the luma TB size in each dimension, except when the luma TB size is 4x4, in which case a single 4x4 chroma TB is used for the region covered by four 4x4 luma TBs.

To compute color saliency map from HEVC video stream, only chroma DC coefficients, which represent the average color of the chroma transform block TB, are extracted. First, we calculate, for each 4×4 region inside TB, a color average for each of the chroma color components Cr and Cb.

$$DC^c(k) = \sqrt{DC^c_{TB} \times DC^c_{TB}/N} \quad (3)$$

Where k is the 4x4 region index in the frame, c is the color component, DC_{TB} is the DC coefficient in TB and N is the total of the 4x4 regions in TB.

Then, based on the average color, we calculate the average opponent-color pairs RG_k and BY_k for the associated 4x4 region k. Finally the color map is computed according to:

$$M_c(k) = {RG_k}^2 + {BY_k}^2 \quad (4)$$

The color conspicuity map will be obtained by displaying M_c where the highest values represent the salient blocks.

3.2.2 *Orientation map*

Compared to MPEG-4 AVC, changes in the intra prediction process has been introduced in HEVC in both prediction block sizes and prediction modes. HEVC supports variable intra prediction block sizes from 64x64 down to 4x4. As MPEG-4 AVC, DC and planar mode are defined, while intra angular prediction directions are augmented from 8 to 33.

According to intra HEVC paradigm, the prediction modes reflect the orientation of the corresponding block with respect to its neighboring blocks. The orientation map will be computed by analyzing the discontinuities among the intra prediction modes of intra frame blocks: blocks which feature the same direction as their neighborhood are considered as non salient while blocks with different orientation modes are considered as salient.

The building of the orientation map starts by analyzing the intra prediction block sizes. Large intra prediction blocks are considered as non salient regions. In the remaining cases, values of the prediction modes are extracted; then, the obtained orientation for each 4×4 block will be compared to those obtained for a set of neighboring blocks.

The M_o orientation map is computed according to:

$$M_o(k) = \begin{cases} Card(\{O_l = O_k; \forall l \in V\}) \; if \;\; PB\; size \geq 8\times 8 \\ 0 \quad else \end{cases} \quad (5)$$

where k is the block index in the frame, V is the n block neighborhood and l is the block index belonging to V.

3.2.3 *Motion map*

In addition to the advanced motion vector prediction presented in prior standards, HEVC defines a new inter prediction mode: the merge mode, which derives the motion information from spatially and temporally neighboring blocks. Compared to MPEG-4 AVC, HEVC includes asymmetric motion partitioning and share the accuracy of motion compensation, which is in units of one quarter of the distance between luma samples.

For each GOP, we define the motion saliency map from HEVC stream as the global motion amplitude, computed by summing the motion amplitude over all the P frames in the GOP, at the same corresponding block position:

$$M_m(k) = \sum_{P \in GOP} \sqrt{{MVDx_k}^2 + {MVDy_k}^2} \quad (6)$$

where $(MDVx_k, MVDy_k)$ denote horizontal and vertical components of motion vectors difference in the P frame block k, and M_m represents the global motion amplitude among the P frames GOP; the larger this M_m value, the more salient the k block position.

3.3 Elementary saliency maps post-processing

The obtained saliency maps for each feature are now to be normalized to the same dynamic rang. This is achieved on each individual map, by three steps approach, Fig. 1.

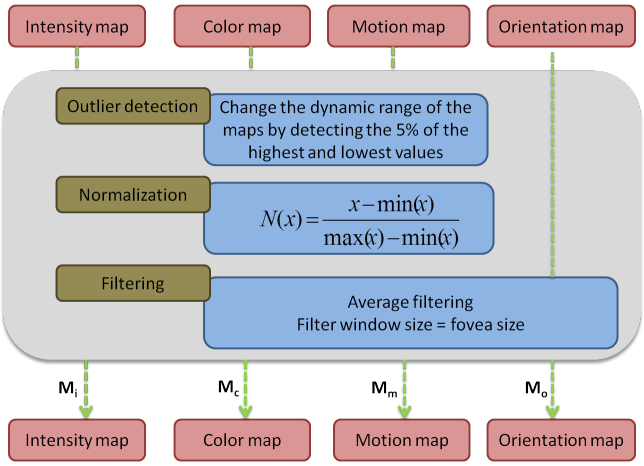

Figure 1 Elementary saliency maps post-processing

First, outlier detection is performed: the 5% largest and the 5% lowest values are eliminated.

Then the remaining values are mapped to the [0 1] interval through an affine transform.

Finally, an average filtering, with the window size equal to the fovea area is applied.

In the case of the orientation map where its values belong to [0 1], the first two post-processing operations are skipped.

3.4 Saliency maps pooling

Individual features (intensity, color, orientation, motion) are processed separately to produce individual feature maps, which are then fused to obtain the saliency map that globally represents the most salient regions.

In our work, we first combined among them the static features maps into a static saliency map, then we merged it with the motion saliency map to finally have a global saliency map.

In the following subsections, we describe the pooling techniques to create the static saliency map and then the global saliency map.

3.4.1 *Static saliency map fusion formulas*

The static saliency map is computed as a linear combination of the intensity, color, and orientation normalized maps as follows:

$$M_s = \beta_1 M_i + \beta_2 M_C + \beta_3 M_o) \quad (7)$$

where β_1, β_2, and β_3 are the parameters determining respectively the weight for the intensity, color and orientation map.

Color advantage fusion: here, we consider the equation (7) with $\beta_1 = 0.2, \beta_2 = 0.6,$ et $\beta_3 = 0.2$.

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

Orientation advantage fusion: here, we consider the same equation (7) with $\beta_1 = 0.2, \beta_2 = 0.2, \text{et } \beta_3 = 0.6$

Intensity advantage fusion: here we take $\beta_1 = 0.6, \beta_2 = 0.2, \text{et } \beta_3 = 0.2$

Mean fusion: the same as the previous ones with equal weights $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$.

Max fusion: This is a winner takes all strategy in which the maximum value between the three maps is taken for each block.

$$M_S = max(M_i, M_c, M_o) \quad (8)$$

Multiplication fusion: a block by block multiplication is done, corresponding to a logical AND.

$$M_S = M_i \times M_c \times M_o \quad (9)$$

3.4.2 Spatio-temporal saliency map fusion formulas

The sptio-temporal saliency map is obtained according to a fusion function *f* applied on spatial saliency map and motion saliency map.

In our study, we consider 8 fusion functions on each static saliency map (calculated according to the 6 previously presented formulas) and the motion map: we shall thus obtain 48 saliency maps. Each of these 48 saliency map will be compared to the density saliency map in order to validate its performances and to know the better combination that gives us the best HEVC saliency map.

Mean fusion[4]: this fusion technique takes the pixel average of both static and dynamic saliency map.

$$M_F = (M_S + M_D)/2 \quad (10)$$

Max fusion[18]: this is a winner takes all strategy in which the maximum value between the two saliency maps is taken for each pixel.

$$M_F = max(M_S, M_D) \quad (11)$$

Multiplication fusion[18]: a pixel by pixel multiplication is done, corresponding to a logical AND.

$$M_F = M_S \times M_D \quad (12)$$

Maximum skewness fusion[18]: The static saliency map is modulated by its maximum value. The dynamic saliency map is modulated by its skewness value. The reinforcement term gives more importance to the areas that are salient both in static and dynamic way.

$$M_F = \alpha M_S \times \beta M_D + \gamma(M_S + M_D) \quad (13)$$

with $\alpha = \max(M_S)$, $\beta = skewness(M_D)$ and $\gamma = \alpha\beta$

Binary threshold fusion[19]: A binary mask is used to exclude spatiotemporal inconsistent areas and to enhance the robustness of the final saliency map when the global motion parameters are not estimated properly.

$$M_F = \max(M_S, M_D \cap M_B) \quad (14)$$

Motion priority fusion[20]: this fusion technique states that a viewer might pay more attention to the motion caused by a moving object even when the static background is more attractive.

$$M_F = (1-\alpha)M_S + \alpha M_D, \quad (15)$$

With $\alpha = \lambda e^{1-\lambda}$ and λ=max (M_S)-mean (M_D).

Dynamic weight fusion[21]: in this fusion method, the weights of the static and dynamic saliency maps are determined by the ratio between the means of both maps for each frame.

$$M_F = (1-\alpha)M_D + \alpha M_S, \quad (16)$$

Where α = mean (M_D)/ (mean (M_S) + mean (M_D))

Scale invariant fusion[22]: in this fusion technique, the input images are analyzed at three different scales from 32×32 to 128×128 to original image size. Three fused maps are obtained which are finally combined linearly into the final spatio-temporal saliency map.

$$M_F = \sum_{l=1}^{3} w_l M_F^l \quad (17)$$

Where $M_F^l = (1-\alpha)M_D + \alpha M_S \; with \; \alpha = 0.5$ is the fused map at scale l and the coefficients of the linear combination are $w_1 = 0.1, w_2 = 0.3 \; and \; w_3 = 0.6$.

4 Experimental results

In this section, we consider a public database [3]. It contains 8 video sequences of 10 seconds each one. For each video, the eye-tracker data are extracted for 30 observers. The distance between observers and the display was set to 3m. The resolution of the display was 1920×1080 with 50Hz frame rate. Based on those results, a density fixation maps are calculated and given with each video.

Before calculating our HEVC saliency maps, those video sequences were encoded in the HEVC standard. The GOP is composed only of *I* and *P* frames, its size is set to 5 and the frame size is set to (576×720). The HEVC reference software (JCT-VC HEVC) is completed with software tools allowing the parsing of these elements and their subsequent usage, under syntax preserving constraints.

Our experiment consists of comparing the obtained saliency maps according to different fusioning formulas by calculating the distance between the saliency map and the density fixation map using two measures: KL-D which is a distribution-based metric and the AUC which is a location-based metric. We use here the Borji's implementation for both KL-D and AUC calculation and we compare each obtained map to existing saliency detection models. According to the used code the KL-D is the distance between the saliency map and the density fixation map and the AUC is the area under the ROC curve of the saliency map and the binarised density fixation map. To binarise the density fixation map, we used the threshold as the half of maximum value of the entire map.

Figures 2-9 represent the result of the comparison of the obtained saliency maps with four methods of the state of the art, namely: Ming Cheng et. al. [7] (referred to as the Ming method), Hae Seo et.al. [12] (referred to as Hae), Stas Goferman [13] (referred to as Gof) and our previous work in MPEG-4 AVC video stream [1] (referred to as AVC). In the case of the AVC method, the best result in each spatio-temporal fusion technique computed in [1] is used.

As a general tendency, Figures 2-9 bring to light that saliency extraction from the HEVC stream outperforms (in both KL-D and AUC sense) the three investigated uncompressed domain state-of-the-art methods. However, no sharp conclusion can be drawn when comparing the HEVC domain to AVC domain: the performances depend on both the static and spatio-temporal saliency pooling technique.

In order to quantify these behaviors we define and compute two coefficients ϱ_{Mij} and η_{Mij}, defined as follows:

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

$$\varsigma_{Mij} = \frac{KLD_{Mj} - KLD_{Mi}}{KLD_{Mj}} \quad (1)$$

where KLD_{Mi} represents the minimal KL-D value over all the six static pooling formulas for a given spatio-temporal saliency pooling formula i, i=1, 2,…8 (the compressed domain saliency maps). KLD_{Mj} is the KL-D value of the maps Mj, j = 1,2,3,4 (the four state of the art maps, Ming, Hae, Gof, AVC).

$$\eta_{Mij} = \frac{AUC_{Mi} - AUC_{Mj}}{AUC_{Mj}} \quad (2)$$

where AUC_{Mi} represents the maximal AUC value over all the six static pooling formulas for a given spatio-temporal saliency pooling formula i, i=1, 2,…8, (the compressed domain saliency maps) and AUC_{Mj} is the AUC value of the maps Mj, j = 1,2,3, 4 (the four state of the art maps, Ming, Hae, Gof, AVC).

According to these definitions, a gain with respect to the state of the art is reflected by positive ϛ and η values.

The ϛ and η coefficients are reported in Tables 1 and 2, respectively.

	Ming[7]	Hae[12]	Gof[13]	AVC[1]
Mean (stat_max)	0.41	0.39	0.31	-0.03
Max (stat_max)	0.39	0.37	0.28	-0.07
Multiplication (stat_mean)	0.12	0.08	-0.03	-0.58
Maximum skewness (stat_mean)	0.39	0.36	0.28	-0.07
Binary threshold (stat_max)	0.34	0.31	0.22	-0.19
Motion priority (stat_max)	0.16	0.13	0.01	0.27
Dynamic weight (stat_max)	0.41	0.39	0.31	-0.05
Scale invariant (stat_max)	0.41	0.39	0.31	-0.02

TABLE 1 KLD GAINS BETWEEN HEVC SPATIO-TEMPORAL SALIENCY MAP FUSION TECHNIQUES AND THE STATE OF THE ART METHODS [7] [12] [13] [1]

Table 1 shows that when comparing the HEVC saliency map extracted in the HEVC domain to the three uncompressed-domain methods based on the KL-D, with singular exceptions, the ϛ coefficient is larger than 0.1 (its maximal value reaching 0.41). The worst performances are provided by the (Multiplication, static_mean) pooling combination, when the Gof method outperforms by 3% the HEVC saliency detection. When compared to the AVC saliency extraction, the pooling technique has a bigger impact in the overall performances:

- the (Mean, stat-max), (Dynamic weight, stat_max) and (Scale invariant, stat_max) combinations result in quite equal good performances, the ϛ being lower than 5%;
- the (Max, stat_max), (Multiplication, stat_mean), (Maximum skewness, stat_mean) and (Binary threshold, stat_max) combinations result in better performances for the AVC saliency map extraction;
- the (Motion priority, stat_max) combination ensures better performances for the HEVC saliency extraction.

A similar analysis can be performed based on the η coefficient reported in Table 2. This time, all the figures show that HEVC outperforms the three state-of-the-art uncompressed domain methods with gains ranging from 6% to 23%. Moreover, HEVC and AVC saliency extraction feature equally good performances: the absolute value of the η coefficient is always lower than 3%.

	Ming[7]	Hae[12]	Gof[13]	AVC[1]
Mean (stat_max)	0.23	0.19	0.18	0.00
Max (stat_max)	0.22	0.19	0.18	0.00
Multiplication (stat_mean)	0.10	0.08	0.06	-0.03
Maximum skewness (stat_mean)	0.22	0.19	0.18	0.00
Binary threshold (stat_max)	0.21	0.18	0.17	0.03
Motion priority (stat_max)	0.18	0.15	0.13	-0.02
Dynamic weight (stat_max)	0.23	0.19	0.18	0.01
Scale invariant (stat_max)	0.23	0.19	0.18	0.00

TABLE 2 AUC GAINS BETWEEN HEVC SPATIO-TEMPORAL SALIENCY MAP FUSION TECHNIQUES AND THE STATE OF THE ART METHODS [7] [12] [13] [1]

5 Conclusion

This paper investigates whether the information related to the human visual saliency is still preserved at the level of the HEVC compressed stream syntax elements.

In this respect, we define elementary intensity, color, orientation and motion saliency maps. They are related to luma residual coefficients, chroma residual coefficients, intra prediction modes and motion vectors difference respectively.

These individual maps are pooled according to 6 static formulas: color advantage, orientation advantage, intensity advantage, mean, max and multiplication. Over each of these 6 static formulas, we consider 8 spatio-temporal formulas, namely: Mean, max, multiplication, maximum skewness, binary threshold, motion priority, dynamic weight and scale invariant.

The experiments consider 8 video sequences [3] and compare the 6x8 HEVC saliency maps to three state of the art uncompressed domain methods references as well as to results obtained in our previous study devoted to the MPEG-4 AVC [1].

By computing both Kl-D and HEVC, it is brought to light that the saliency can be extracted directly from the HEVC stream syntax (after the entropic decoding), without any overall loss in performances. These results prove that, as its predecessor MPEG-4 AVC, although not designed by exploiting saliency principles, the HEVC standard preserves the visual saliency at the stream syntax elements level.

From the practical point of view, this result opens the door towards a large variety of applications, like video retargeting, object segmentation and discovery or video surveillance.

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

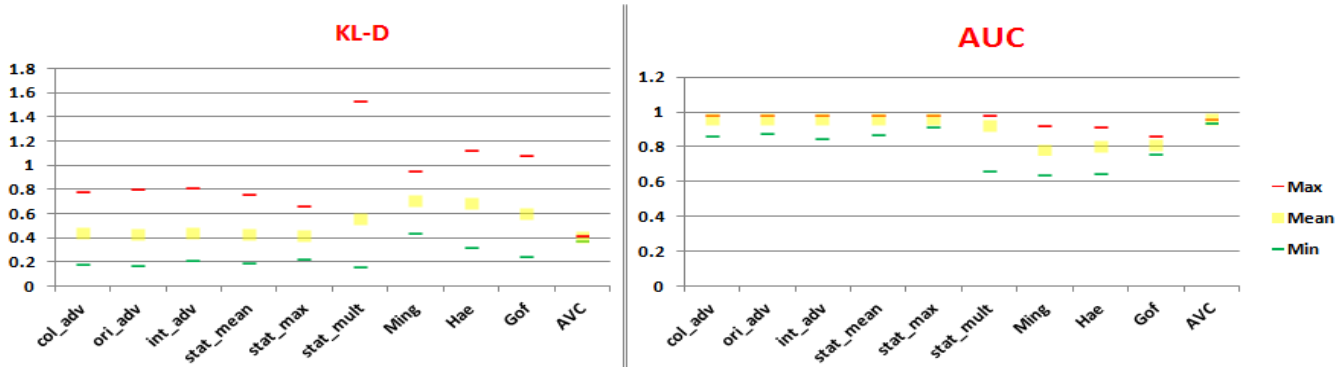

FIGURE 2 MEAN FUSION

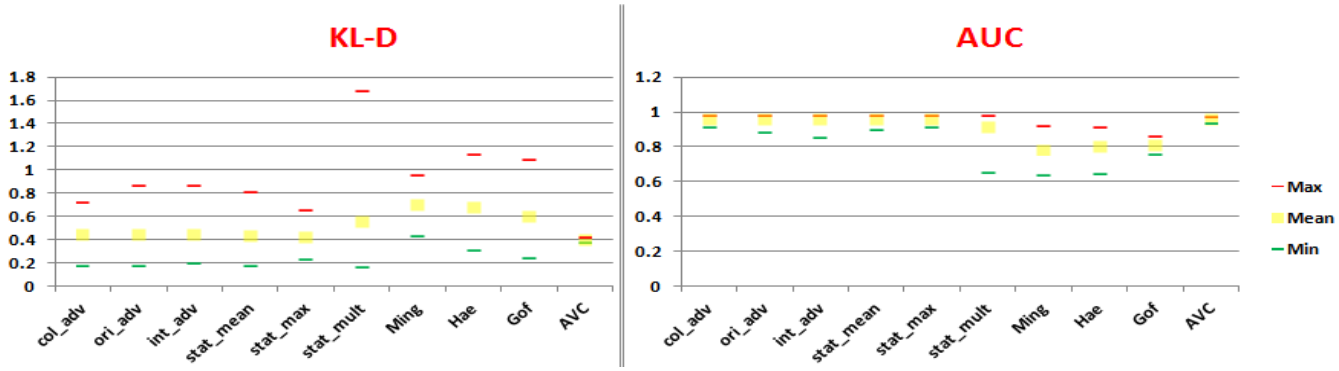

FIGURE 3 MAXIMUM FUSION

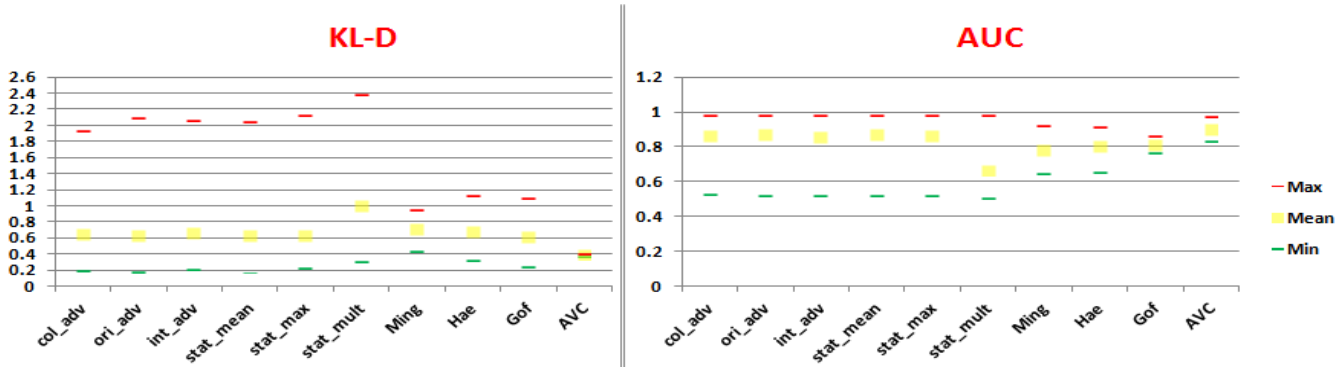

FIGURE 4 MULTIPLICATION FUSION

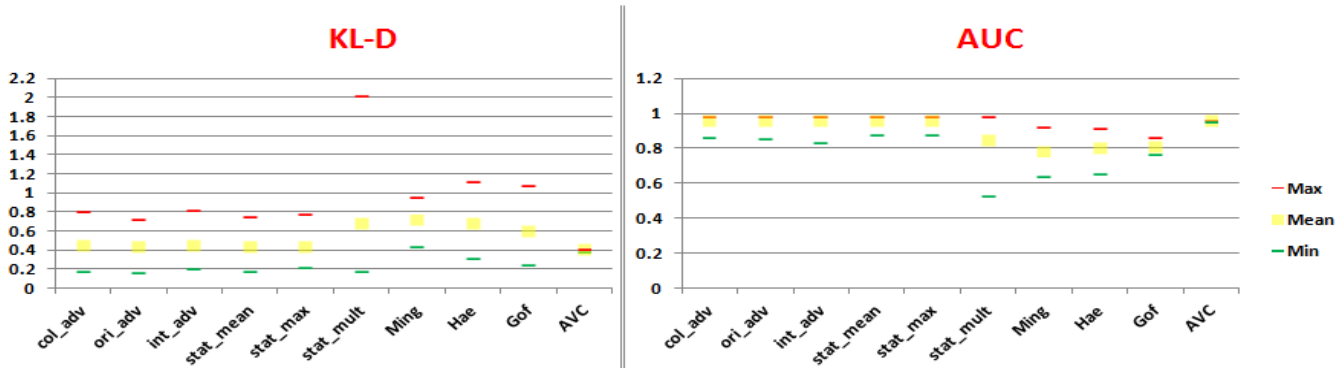

FIGURE 5 MAXIMUM SKEWNESS FUSION

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

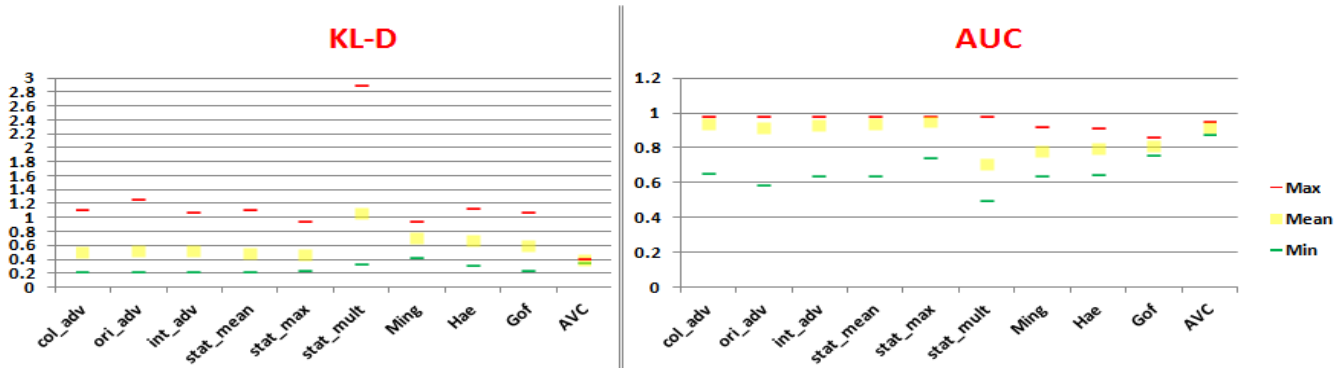

FIGURE 6 BINARY THRESHOLD FUSION

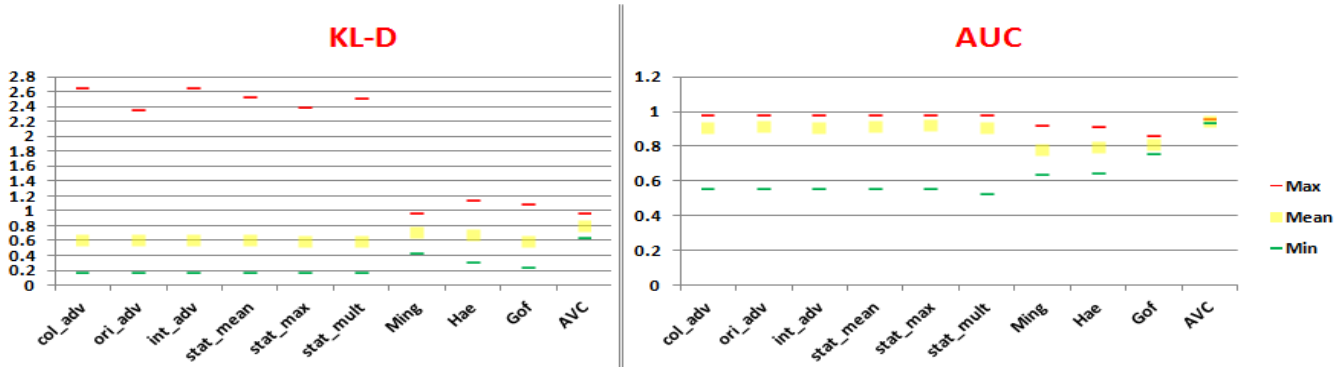

FIGURE 7 MOTION PRIORITY FUSION

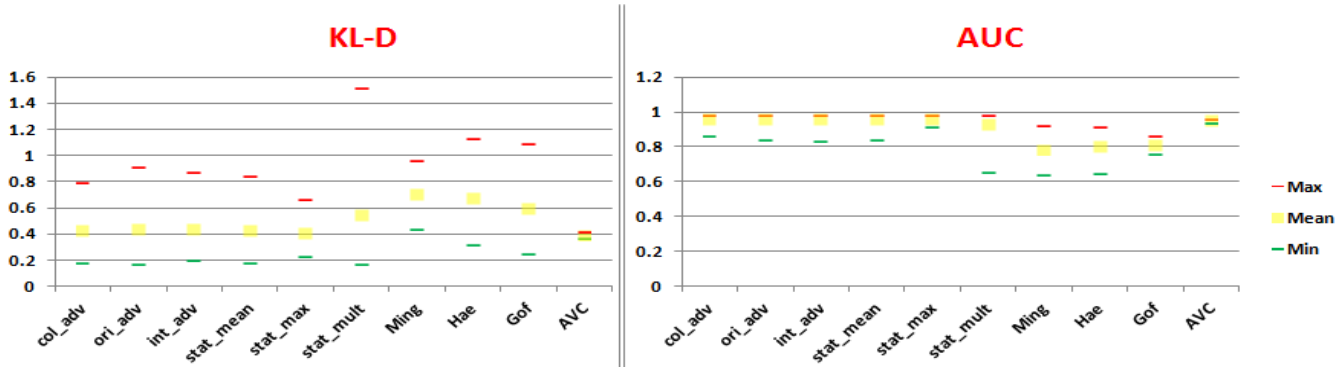

FIGURE 8 DYNAMIC WEIGHT FUSION

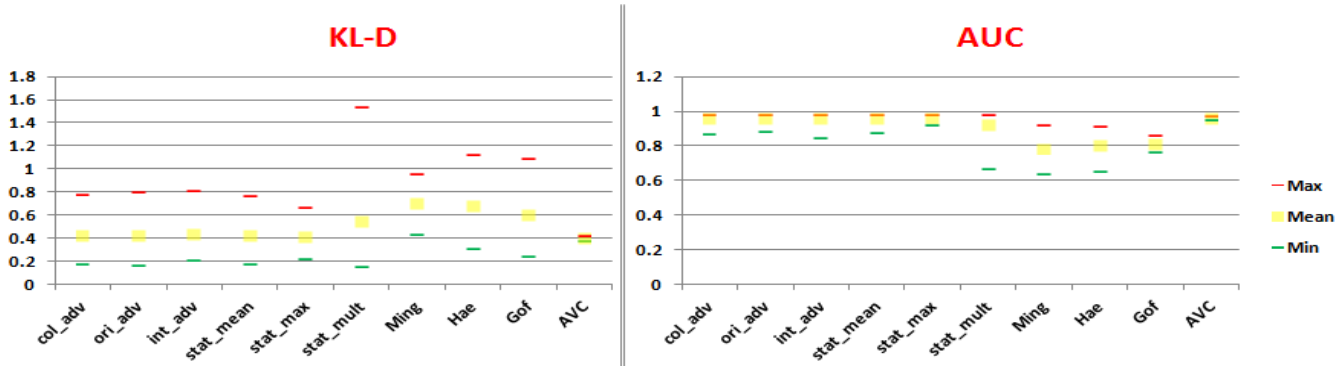

FIGURE 9 SCALE INVARIANT FUSION

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-107

References

[1] M. Ammar, M. Mitrea, M. Hasnaoui, and P. Le Callet, "Visual saliency in MPEG-4 AVC video stream," in *IS&T/SPIE Electronic Imaging International Society for Optics and Photonics*, 2015, p. pp. 93940X–93940X.

[2] M. Ammar, M. Mitrea, and M. Hasnaoui, "M. MPEG-4 AVC saliency map computation," in *IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics*, 14AD, p. 90141A–90141A.

[3] "ftp://.ivc.polytech.univnantes.fr/IRCCyN_IVC_Eyetracker_SD_2009_12/." .

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Adv. Neural Inf. Process. Syst.*, pp. 545–552, 2006.

[6] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, 2006.

[7] M. M. Cheng, J. Warrell, W. Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1529–1536.

[8] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression.," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, 2010.

[9] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues Categories and Subject Descriptors," in *Proceedings of the 14th annual ACM international conference on Multimedia*, 2006, vol. 32816, pp. 815–824.

[10] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Adv. Neural Inf. Process. Syst.*, vol. 21, no. 800, pp. 681–688, 2008.

[11] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Res.*, vol. 47, no. 19, pp. 2483–2498, 2007.

[12] H. J. Seo and P. Milanfar, "Nonparametric bottom-Up saliency detection by self-resemblance," in *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, 2009, pp. 45–52.

[13] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, 2012.

[14] Y. Fang, Z. Chen, W. Lin, and C. W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, 2012.

[15] Y. Fang, W. Lin, S. Member, Z. Chen, C. Tsai, and C. Lin, "A Video Saliency Detection Model in Compressed Domain," vol. 24, no. 1, pp. 27–38, 2014.

[16] M. Hasnaoui and M. Mitrea, "Multi-symbol QIM video watermarking," *Signal Process. Image Commun.*, vol. 29, no. 1, pp. 107–127, 2014.

[17] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[18] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, 2009.

[19] T. Lu, Z. Yuan, Y. Huang, D. Wu, and H. Yu, "Video retargeting with nonlinear spatial-temporal saliency fusion," *Proc. - Int. Conf. Image Process. ICIP*, pp. 1801–1804, 2010.

[20] J. Peng and Q. Xiao-Lin, "Keyframe-based video summary using visual attention clues," *IEEE Multimed.*, vol. 17, pp. 64–73, 2010.

[21] X. Xiao, C. Xu, and Y. Rui, "Video based 3D reconstruction using spatio-temporal attention analysis," in *2010 IEEE International Conference on Multimedia and Expo*, 2010, pp. 1091–1096.

[22] W. Kim, S. Member, C. Jung, C. Kim, and S. Member, "Spatiotemporal Saliency Detection and Its Applications in Static and Dynamic Scenes," *IEEE Trans. Circuits Syst.*, vol. 21, no. 4, pp. 446–456, 2011.