Video Compression for Object Detection Algorithms

Leonardo Galteri, Marco Bertini, Lorenzo Seidenari, and Alberto Del Bimbo
Media Integration and Communication Center, Università degli Studi di Firenze
Viale Morgagni 65 - 50134 Firenze, Italy
{*name.surname*}@unifi.it

***Abstract*—Video compression algorithms have been designed aiming at pleasing human viewers, and are driven by video quality metrics that are designed to account for the capabilities of the human visual system. However, thanks to the advances in computer vision systems more and more videos are going to be watched by algorithms, e.g. implementing video surveillance systems or performing automatic video tagging.**

This paper describes an adaptive video coding approach for computer vision-based systems. We show how to control the quality of video compression so that automatic object detectors can still process the resulting video, improving their detection performance, by preserving the elements of the scene that are more likely to contain meaningful content.

Our approach is based on computation of saliency maps exploiting a fast objectness measure. The computational efficiency of this approach makes it usable in a real-time video coding pipeline. Experiments show that our technique outperforms standard H.265 in speed and coding efficiency, and can be applied to different types of video domains, from surveillance to web videos.

I. Introduction

Video streaming applications that require transmission of a high number of streams to some central server, have to deal with issues such as limited bandwidth channels or a bandwidth bottleneck on the server itself. Some examples are video surveillance networks, UHF video streaming networks connecting dash cams installed in police cars, or streams transmitted from wearable devices or smartphones. All these applications require to transmit videos with a reasonable high-quality for further processing by vision-based systems, e.g. to identify anomalous activities, detect and identify persons, and detect objects. A way to improve the video compression methods currently used is to reduce the amount of irrelevant information transmitted in the video stream, compressing more the parts that do not contain semantically interesting objects. Typically this is performed by computing visual saliency maps based on some model of the human visual system.

Differently from this approach our method is designed to compute a binary saliency map designed for computer vision-based systems, considering the case of a system that performs automatic object detection. To this end we exploit the objectness measure, that quantifies how likely an image window is containing an object of any class [1]. An objectness saliency map is computed from window proposals, and it is combined with H.265 quantization parameter map. We propose a method to learn a map that indicates to the codec which parts of the frame are relevant for a computer vision algorithm, combining a semantic and a low-level cue, maintaining also the perceptual quality for human viewers. Detector performance improves especially for low bitrates ($\sim$ 1k–2k Kb/s), while for similar bitrates our compression method preserves perceptual quality of relevant regions better than standard codecs. An additional result is that the proposed method is much faster than standard video coding.

II. Previous work

Saliency and objectness: Visual saliency deals with properties and qualities of items that stand out with respect to their neighbors, and typically arises from contrasts in terms of color, luminance, motion, etc. Works in this area typically aim at predicting salient points of human eye fixation [23] or modeling visual attention [6]. Because of this relation with the human visual system and its attentional model, saliency has been used to identify parts of images or video frames that are to be compressed lightly, to preserve their visual content.

However, a detector may need to handle objects that are not visually conspicuous or that do not draw human gaze, thus an object proposal method should be able to deal also with objects that are not salient. Hosang *et al.* [14] have recently presented a comparison of the current state-of-the-art in objectness proposal methods. Uijlings *et al.* [25] propose a method that requires no parameter learning, combining exhaustive search and segmentation in a data-driven selective search. The approach is based on hierarchical grouping of regions, using color, texture and region features. The method of Zitnick and Dollár [28], called Edge Boxes, computes a scoring function in a sliding window fashion. Scoring is performed measuring the number of edges that exist in the box minus those that are members of contours that overlap the box's boundary. The method of Cheng *et al.* [9] is the fastest approach, as reported in the comparison of [14], and uses a simple linear classifier over edge features, that is trained and applied in a sliding window manner. The efficiency of this approach is due to the use of approximated features, binarized normed gradients that give the name (BING) of the method.

Video coding: Traditional adaptive video compression approaches do not consider the semantic content of video and instead adapt compression depending on the requirements of the network or device used to deliver video to the end user. Semantic video compression, instead, alters the video by taking into account objects [15], [18] or a combination of objects and events [4], using pattern recognition techniques. Bagdanov *et al.* [2] have proposed to use image features correlated with downstream detector features, like corners

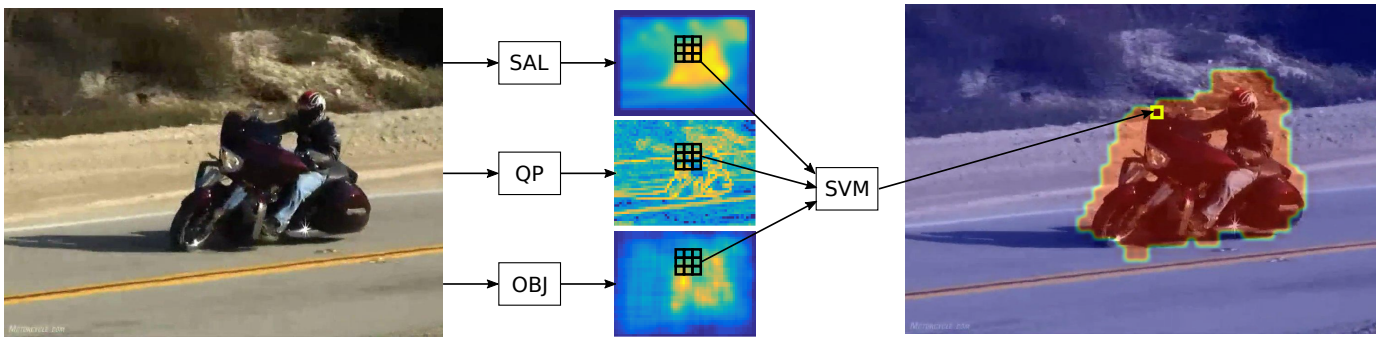

Fig. 1. Our system pipeline. Binary saliency maps are predicted using [9], [22], [24] fused with our learned model. The final binary map is shown rightmost.

and edges, to select frame areas to be maintained at higher visual quality and smoothing the rest. This approach has been applied, as pre-processing step to H.264 encoding, to surveillance videos. Videos compressed using [2] have a very low bitrate. Moreover the performance of pedestrian detection is better on videos compressed with this approach with respect to those compressed with H.264 at a similar bitrate.

Chen *et al.* [8] have addressed surveillance video coding, segmenting moving objects from background using low-rank and sparse decomposition, and coding the background with a few independent frames, based on their linear dependency, to further reduce their temporal redundancy. Guo *et al.* [11] follow a similar approach, but separating background from moving objects by learning a background dictionary, and coding motion together with the reconstruction coefficients of the background.

Use of saliency maps for video coding has been proposed by Gupta *et al.* [12], where low-level and high-level saliency features are combined and used to perform a non-uniform bit allocation over video frames. Since computation of these saliency maps is expensive the authors propose to use a shot-detection method to select a reduced number of frames for this computation. Hadizadeh and Bajić [13] have proposed to add a saliency distortion term in the rate distortion optimization (RDO) processing of H.264/AVC, to improve the coding quality of regions of interest (ROI). Saliency is computed with an extended version of the Itti-Koch-Niebur saliency model [16], with an improved temporal saliency that accounts for camera motion. The proposed RDO is $\sim$ 3 times slower than the standard approach. Li *et al.* [19] have proposed the use of saliency maps in H.265/HEVC coding to drive the quantization parameter of the coding units. An initial map is computed using graph-based visual saliency [7], then a Markovian algorithm concentrates saliency in a few locations, followed by a Gaussian filtering.

Differently from [8], [11] the proposed approach can be applied to standard video coding such as H.264 or H.265. Differently from [13], [17], [19], [29] the proposed approach does not compute a saliency that attempts to mimic the human visual system; instead, our saliency exploits objectness measures in to order to obtain a better performance with object recognition algorithms. Differently from [2] the proposed coding does not pre-process video frames to drive the allocation of bits of the encoder, but rather intervenes directly on the quantization parameter of the encoder, thus avoiding an extra processing step.

III. Coding with learned saliency

The goal of the proposed approach is to learn a saliency map that can drive compression of video frames in a way that is friendly for computer vision algorithms. The map is based on few features that are fast to compute, so to allow the application of the method to tasks that require real-time coding such as surveillance. Features are also related to the task of a computer vision-based system; in particular we have addressed the problem of object detection, that can be useful for both surveillance and automatic video tagging. This means that the map should indicate which part of the frame contains an object of interest for the algorithm. A secondary goal is to preserve visual quality, in terms of human visual system, for these objects.

Our method is based on modern video coding algorithms, specifically we use H.265, and the open source implementation x265 as practical reference implementation, but it can be applied also to H.264. Compression quality can be controlled with different strategies, e.g. using variable bitrate (VBR), constant bitrate (CBR) or with a constant rate factor (CRF); these approaches allocate the bits to different granularity levels of video, then reach the preallocated budget while encoding by adjusting the quantization parameter (QP) [5]. In particular, x265 trades distortion for bitrate, following different strategies, that typically adopt psycho-visual options that improve the perceived visual quality (e.g. favoring wrong motion over blur, or preserving the energy of the reconstructed patch). This approach is geared towards preserving the image appearance as much as possible. If the coded video has to be principally consumed by a machine we only need to keep the video quality good enough for the task at hand to be completed successfully. The video created with our approach is completely H.265 compliant and requires no changes in the decoder.

We propose to generate a binary saliency map, created according to the quadtree partitioning of the coding tree unit (CTU), indicating whether keeping the originally estimated QP or setting QP to the maximum value 51 (highest compression)

[24]. We take into account *i)* an objectness saliency map, *ii)* motion map, *iii)* visual saliency map and *iv)* QP map, and propose to combine them to predict a binary map. The objectness map is directly related to the task of the computer vision system that will consume the video, i.e. object detection, while motion and saliency map are used to weight how much visually conspicuous are the objects in the scene; finally, the QP map account for visual features such as texture and also for the secondary goal of the proposed approach, that is to obtain visually pleasant objects for the possible human viewers. Fig. 1 shows a schema of the proposed approach.

More formally, let us consider a video frame t, a patch of $N \times N$ pixels centered at location (x, y), and $M_i, i = 1 \dots 4$ maps listed above. We define the operator $\mathcal{NN}_N(x, y, M)$ that extracts the vector $[M_i(x - N/2, y - N/2) \dots M_i(x + N/2, y + N/2)]$ concatenating the values of a neighborhood of size $N \times N$ centered on (x, y).

To learn how to fuse the maps with the best performing weights, we learn a function $f(\mathbf{x}) \rightarrow [0, 1]$ using a RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2)$ setting γ and C by five-fold cross validation.

Positive samples are patches whose center belong to detected objects, while negative patches are sampled from the background. Patch samples can be collected in an unsupervised manner by running an object detector with a high threshold on few frames; as an example on TownCenter we collect ~ 4000 samples using only 2 frames.

Each saliency map M_i is normalized to zero mean and unit variance. In the following we review the maps, extracted with known algorithms, that we combine to create the proposed objectness-based saliency map.

Objectness Saliency Map: Considering a set of object proposals $\mathcal{S}_k \subseteq \mathcal{S}$ we can define an objectness saliency map M_k by accumulating how many proposals of $\mathcal{S}_k$ are present in each frame at each position. Let us denote $\mathcal{S}_k^t$ the set of proposals $\mathcal{S}_k$ at frame t. For every pixel $p = (x, y)$ of frame t, we compute the corresponding objectness saliency map value $M_k^t(p)$ as the count of proposals enclosing this position:

$$M_k^t = \sum_{s \in \mathcal{S}_k^t} \Psi_s \quad (1)$$

where for each proposals $s \in \mathcal{S}_k^t$ we define the function of pixels p in an image:

$$\Psi_s(p) = \begin{cases} 1 & \text{if } p \in s \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The map $M(x, y, t)$ represents at any moment of the video, how much each pixel is relevant with respect to the set $\mathcal{S}_k$. The more proposals overlap in one position the more likely this pixel belongs to an object and is therefore relevant for video coding. The BING objectness proposal measure [9] has been used, because of its low computational cost that makes it amenable for inclusion in a real-time compression system.

Motion Map: It is simply the difference, pixel by pixel, of two consecutive frames.

Visual Saliency Map: To consider the visually salient elements of the scene, we have selected the Fast and Efficient Saliency (FES) map [22], based on estimating saliency of local feature contrast in a Bayesian framework. Again, its good computational performance make it usable in a real-time context.

QP Saliency Map: This map is the quadtree representation of the QP values used in the CTU of the HEVC encoder. QP values are computed over 16×16 pixels by the x265 encoder. This saliency is related to the energy of each patch considered.

IV. Experiments

We evaluate structural similarity index (SSIM) [27], a visual quality metric that models the perception of compression artifacts, and the average log miss rate for a pedestrian detection scenario typical of surveillance videos. Furthermore to evaluate how the method generalizes we evaluated mean average precision (mAP) of generic object detectors on videos downloaded from YouTube. We compare these metrics with many baselines comprised the standard x265 codec for different bitrates. The performance of the proposed method has been evaluated using three measures, as the video bitrate is varied with the CRF: structural similarity index (SSIM) [27], a visual quality metric that models the perception of compression artifacts, and two measures used in object detection: the average log miss rate and mean average precision of object detectors.

Surveillance Videos: We used Town Center [3] composed by a 5 minutes HD video, recorded at 1920×1080 @ 25 fps from a fixed camera, showing people walking in the street of a town, with 71,500 ground truth annotations of persons;

In the first experiment, we have compared the proposed method with a number of baselines. The original video has been compressed with the proposed method and with the baselines. We processed the resulting videos with the ACF pedestrian detector [10] and its detection results have been compared with the ground truth annotations. The baselines are: *i)* the x265 implementation of H.265 ; *ii)* a combination of QP map and motion map; *iii)* a combination of QP map with visual saliency computed with the method proposed by Walter and Koch [26]; *iv)* a combination of QP map, motion map and visual saliency by Walter and Koch; *v)* a combination of QP map, motion map and FES visual saliency.

It has to be noted that using the visual saliency of [26] results in a computational cost that is higher than 150 ms; thus the combinations that use [26] are not suitable for applications that have constraints on processing time.

Fig. 2 reports the average log miss rate of the person detector; the lower the value the better the performance of the detector. The improvement with respect to the x265 baseline is due to the reduction of false detections, that are eliminated by the increased compression of non relevant parts of the frame. Comparison with the other baselines shows that adding the objectness-based map improves over other visual saliencies.

Fig. 3 evaluates the visual quality of the areas of interest for the detector, using the SSIM measure. The improvement obtained by the proposed approach is due to the fact that the encoder is able to better allocate the bits to the areas of interest, rather than distributing them also to the background. The comparison with the other saliency maps, that also beat the performance of standard x265, shows that the objectness measure better selects the elements of interest.

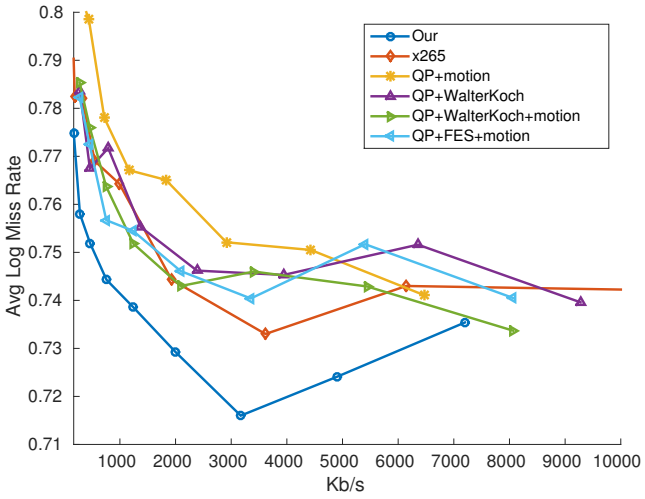

Fig. 2. Detection Average Log Miss rate varying bitrate on Town Centre (lower is better).

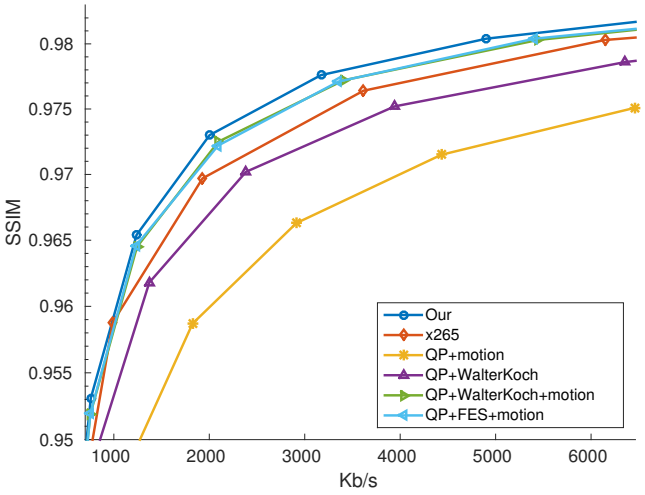

Fig. 3. SSIM varying bitrate on Town Centre (higher is better).

Fig. 4 shows an example of the various saliency maps computed. Fig. 5 shows an example of the compression of TownCenter frames, with the proposed method (lower bitrate, better person recognition) and using the standard x265 codec.

YouTube Videos: For this experiment we used YouTube Objects [20], commonly used to test video object detection methods for different types of objects, evaluating the proposed approach on 10 classes of objects, from cats to trains, to asses the capability to generalize. Since the YouTube Objects dataset videos have typically been captured with non fixed cameras,

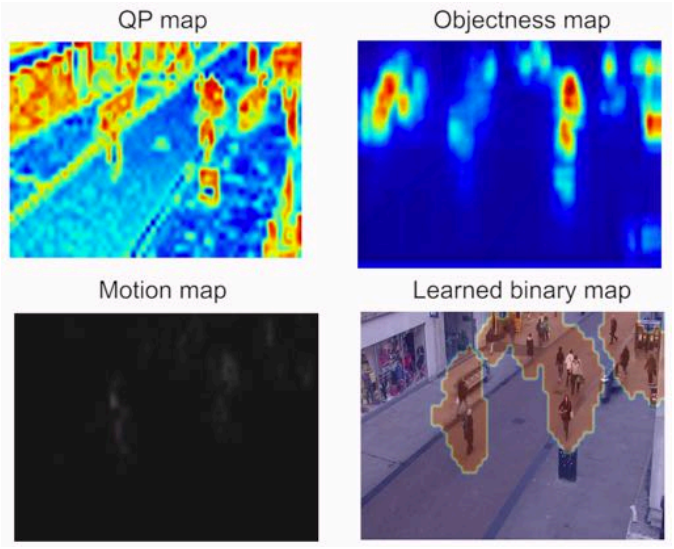

Fig. 4. Examples of saliency maps computed on Town Centre.

Fig. 5. Examples of video compression of Town Centre. The proposed method obtains a much lower bitrate with an improved person detection.

the motion map is not useful, and it has not been used in this experiment.

Given the sparsity of ground truth annotations of YouTube Objects (usually just 1-2 frames per video), we have annotated 1500 objects on a subset of the dataset. The method has been compared to a standard H.265 compression.

Object detections have been computed using Faster R-CNN [21], on the videos compressed with the proposed approach and with the baseline. We used the VGG-16 model, pre-trained on Imagenet and fine-tuned on PASCAL VOC2007 [21]. No tuning of the detector has been performed on the YouTube Objects dataset.

Fig. 6 reports the mean average precision for the 10 object classes of YouTube object. Again, similarly to the previous case, the improvements of the detection can be attributed to

the reduction of false positives.

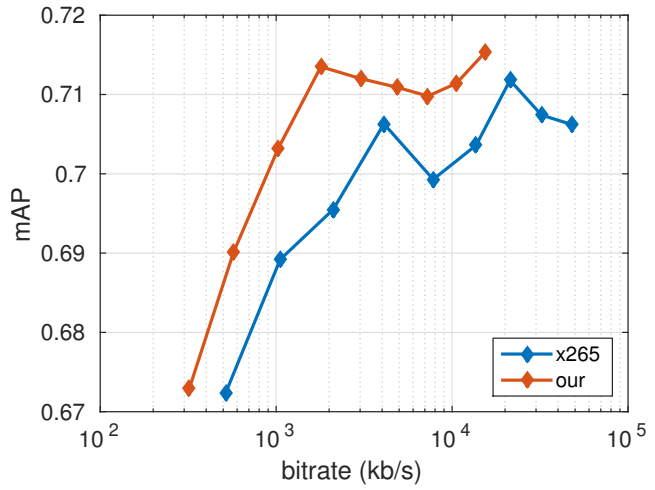

Fig. 6. Mean average precision on YouTube Objects varying bitrate (higher is better).

Fig. 7 shows an example of the various saliency maps computed. Fig. 8 shows an example of the compression of TownCenter frames, with the proposed method (lower bitrate, better person recognition) and using the standard x265 codec. It is clearly shown that "uninteresting" parts of the frame like the road are more compressed, while persons are kept in high quality.

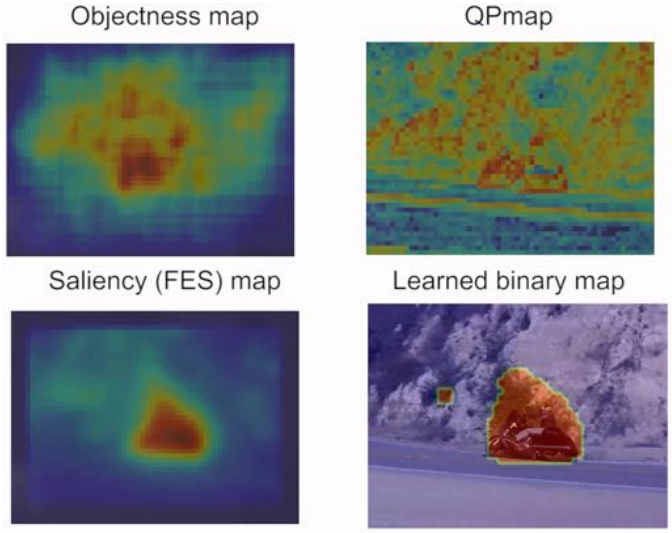

Fig. 7. Examples of saliency maps computed on Youtube objects dataset.

Computational costs: Another advantage of our method is in the reduction of computational costs in video encoding. In Figure 9 we show per-frame encoding time of our approach compared to H.265. We show the timing for three presets of increasing quality and varying the CRF. As expected for both approaches a lower CRF implies more encoding time, setting on average higher QPs, thus leading to less sparse quantized coefficients for each CU.

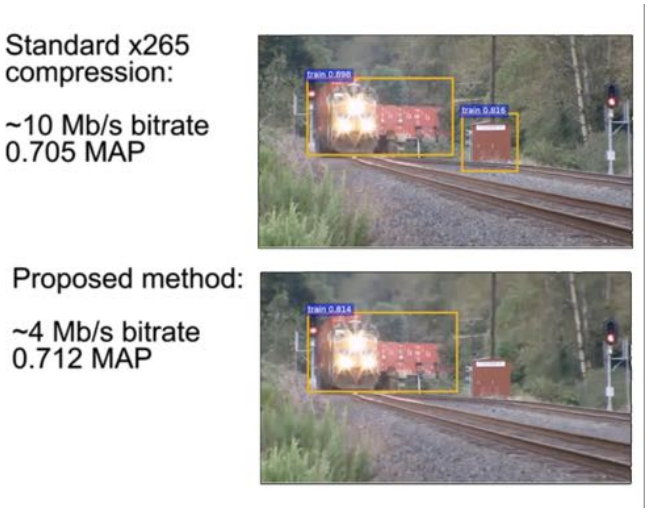

Fig. 8. Examples of video compression of Youtube objects. The proposed method obtains a much lower bitrate with an improved object detection.

This behavior is afflicting H.265 more than our adaptive coding algorithm. Frame blocks, predicted to be irrelevant, by our learned binary map, are quantized with the lowest quality. This mainly affects two aspects. Firstly the DCT coefficients will be mostly zero, leading to a less expensive coding later in the pipeline. Secondly, residues after block matching will be even more sparse and thus coded more efficiently.

Our method is more than two times faster than standard H.265 in coding video frames. This improvement is consistent for all H.265 presets.

V. Conclusions

We proposed a method for adaptive video coding based on a learned saliency. Our novel saliency drives video compression in order to preserve the appearance without damaging the performance of object detectors. Detector performance improves especially for low bitrates ($\sim$ 1k–2k Kb/s). We also show that for similar bitrates our compression method preserves perceptual quality of relevant regions better than standard codecs. Finally we also found that our codec is more than two times faster than standard H.265.

Acknowledgments: This work is partially supported by the "Social Museum and Smart Tourism" project (CTN01_00034_231545).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010.

[2] A. D. Bagdanov, M. Bertini, A. Del Bimbo, and L. Seidenari, "Adaptive video compression for video surveillance applications," in *Proc. of International Symposium on Multimedia*, 2011.

[3] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2011.

[4] M. Bertini, A. Del Bimbo, A. Prati, and R. Cucchiara, "Semantic adaptation of sport videos with user-centred performance analysis," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 433–443, 2006.

[5] T. Biatek, M. Raulet, J.-F. Travers, and O. Deforges, "Efficient quantization parameter estimation in HEVC based on ρ-domain," in *Proc. of IEEE European Signal Processing Conference*, 2014.

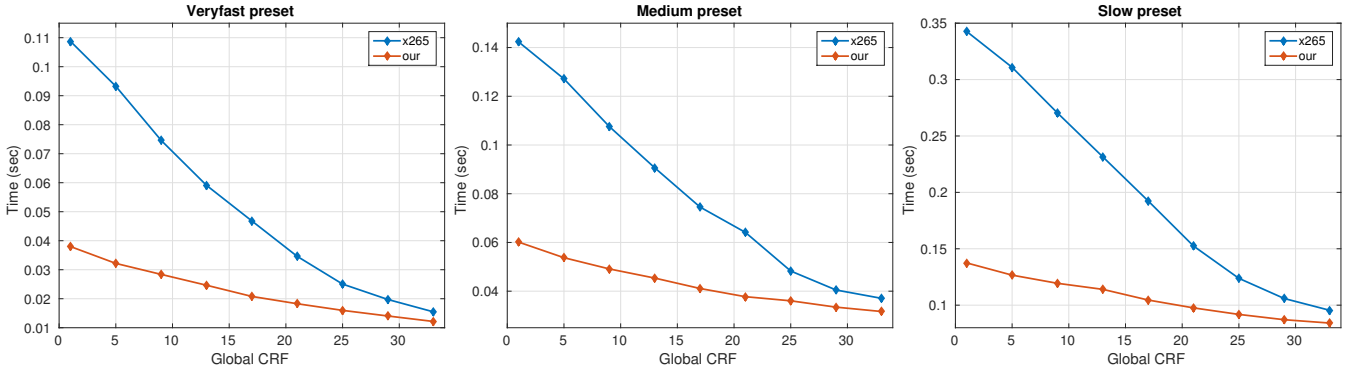

Fig. 9. Coding time, per frame, of our codec compared to reference H.265. Our method is faster and depends less on the CRF (lower is better).

[6] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, 2013.

[7] B. Bross, W. Han, G. Sullivan, J. Ohm, and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 9," JCTVC-K1003, ITU-T/ISO/IEC Joint Collaborative Team on Video Cod- ing (JCT-VC), Tech. Rep., Oct. 2012.

[8] C. Chen, J. Cai, W. Lin, and G. Shi, "Surveillance video coding via low-rank and sparse decomposition," in *Proc. of ACM Multimedia*, 2012. [Online]. Available: http://doi.acm.org/10.1145/2393347.2396294

[9] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

[10] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, 2014.

[11] X. Guo, S. Li, and X. Cao, "Motion matters: A novel framework for compressing surveillance videos," in *Proc. of ACM Multimedia*, 2013. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502145

[12] R. Gupta, M. T. Khanna, and S. Chaudhury, "Visual saliency guided video compression algorithm," *Signal Processing: Image Communication*, vol. 28, no. 9, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596513000982

[13] H. Hadizadeh and I. Bajic, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, 2014.

[14] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" in *Proc. of British Machine Vision Conference*, 2014.

[15] H.-J. Huang, X.-M. Zhang, and Z.-W. Xu, "Semantic video adaptation using a preprocessing method for mobile environment," in *Proc. of IEEE Computer and Information Technology*, 2010.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, 1998.

[17] H. Ke, H. Sun, L. Gao, and H. Wang, "A video image compression method based on visually salient features." *Journal of Digital Information Management*, vol. 12, no. 5, 2014.

[18] C. Kim and J.-N. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 122–129, 2002.

[19] Y. Li, W. Liao, J. Huang, D. He, and Z. Chen, "Saliency based perceptual HEVC," in *Proc. of IEEE International Conference on Multimedia and Expo Workshops*, 2014.

[20] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2012.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99.

[22] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Proc. of Scandinavian Conference on Image Analysis*, 2011. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-21227-7_62

[23] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. of European Conference on Computer Vision*, 2012.

[24] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, 2012.

[25] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11263-013-0620-5

[26] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608006002152

[27] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004.

[28] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. of European Conference on Computer Vision*, 2014.

[29] F. Zund, Y. Pritch, A. Sorkine-Hornung, S. Mangold, and T. Gross, "Content-aware compression using saliency-driven image retargeting," in *Proc. of IEEE International Conference on Image Processing*, 2013.