HYBRID VIDEO OBJECT TRACKING IN H.265/HEVC VIDEO STREAMS

Serhan Gül, Jan Timo Meyer*, Thomas Schierl, Cornelius Hellge, Wojciech Samek*

Fraunhofer Heinrich Hertz Institute
Video Coding & Analytics Department
Einsteinufer 37, 10587 Berlin, Germany

ABSTRACT

In this paper we propose a hybrid tracking method which detects moving objects in videos compressed according to H.265/HEVC standard. Our framework largely depends on motion vectors (MV) and block types obtained by partially decoding the video bitstream and occasionally uses pixel domain information to distinguish between two objects. The compressed domain method is based on a Markov Random Field (MRF) model which captures spatial and temporal coherence of the moving object and is updated on a frame-to-frame basis. The hybrid nature of our approach stems from the usage of a pixel domain method that extracts the color information from the fully-decoded I frames and is updated only after completion of each Group-of-Pictures (GOP). We test the tracking accuracy of our method using standard video sequences and show that our hybrid framework provides better tracking accuracy than a state-of-the-art MRF model. *

***Index Terms*—** object tracking, compressed domain analysis, video surveillance, H.265/HEVC, Markov Random Field (MRF)

1. INTRODUCTION

Video object tracking is an important component of many applications in computer vision such as motion-based recognition, human-computer interaction, automated surveillance, and traffic monitoring [1]. Tracking can be defined as the problem of estimating the trajectory of an object in the image plane as it moves around a scene. A tracking algorithm assigns consistent labels to the tracked objects at subsequent frames in a video [2].

Tracking algorithms can be classified in two broad categories according to their domain of operation: pixel domain and compressed domain. Pixel domain algorithms are characterized by their high accuracy but can also have high computational complexity. Such high computational complexity can limit their usage in scenarios requiring real-time processing of several video streams in parallel. One example scenario is the analysis of videos obtained from surveillance cameras.

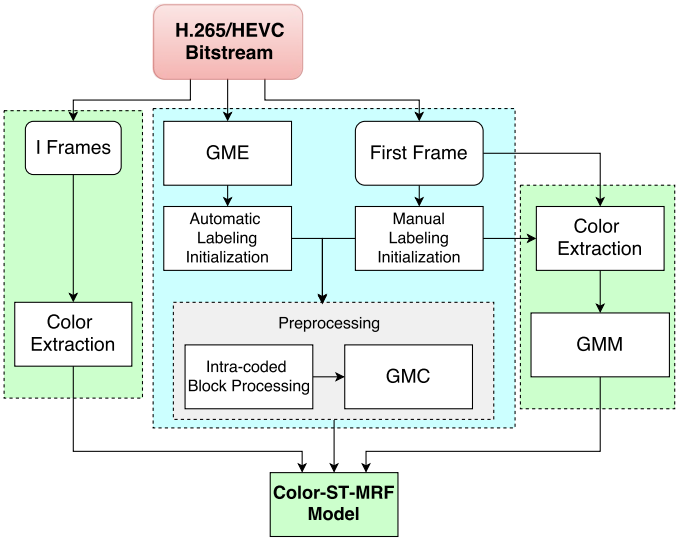

Fig. 1: Block diagram of the proposed method. Blue shaded area shows the components of the ST-MRF model from [3], green shaded areas show the added components that incorporate pixel domain information.

On the other hand, compressed domain algorithms operate with the data encoded in compressed video bitstream such as motion vectors, block coding modes or transform coefficients of the motion-compensated prediction residuals. Compressed domain approaches generally have lower computational cost compared to pixel domain approaches since they avoid a full decoding of the video, thereby reducing the amount of processing and storage requirements significantly. However, they usually perform worse in terms of tracking accuracy due to the lack of full pixel information. A good compromise between the respective advantages of pixel and compressed domain algorithms can be achieved by a *hybrid* framework, which resorts to pixel domain information sparingly, thereby improving the tracking accuracy while still maintaining a low computational complexity.

In this paper we present a hybrid object tracking method which extends the compressed domain method [3] proposed for tracking H.264/AVC videos using color information obtained by fully decoding the I frames and apply our method to H.265/HEVC video bitstreams. The results show that our

*Serhan Gül und Jan Timo Meyer contributed equally.

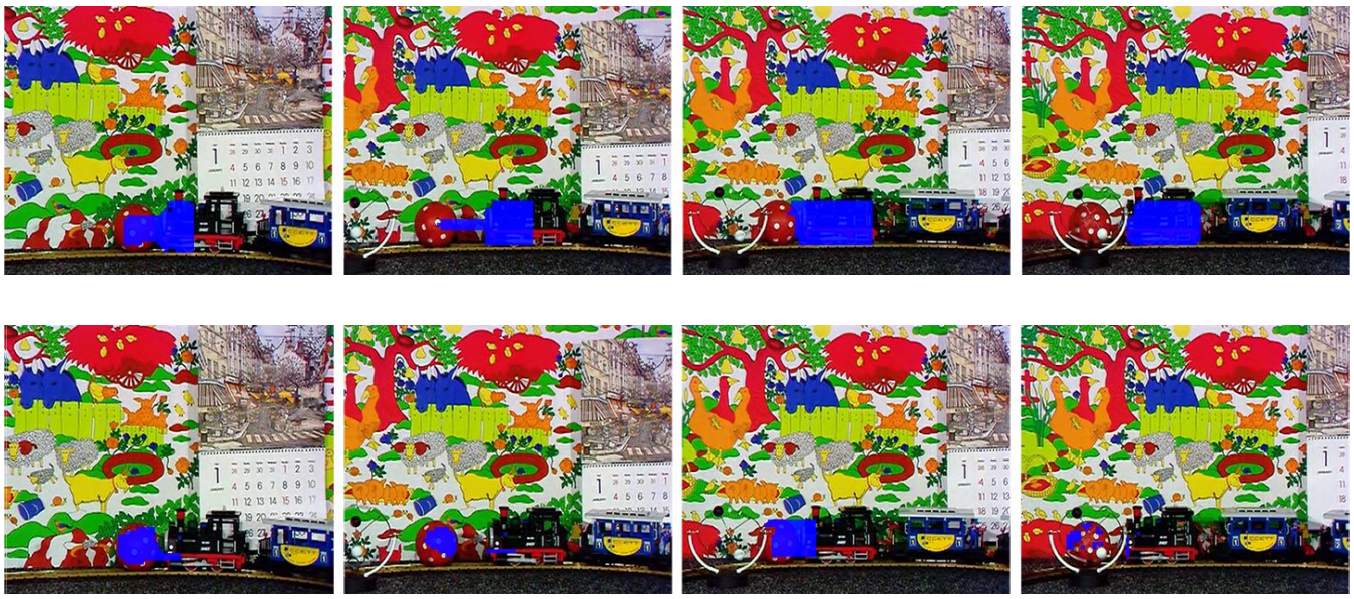

Fig. 2: Comparison of the tracking performance of [3] (top) and our proposed method (bottom) for frames 130, 190, 240, and 275 (left to right) of the *Mobile Calendar* sequence. Blue denotes true positives detected by the respective methods. Our method manages to track the ball fairly accurately along the sequence whereas [3] fails to distinguish between the ball and the train during some parts of the sequence.

method helps to overcome limitations of the original algorithm in [3] and increases the tracking accuracy.

The remainder of this paper is organized as follows. Section 2 presents some of the state-of-the-art compressed domain and hybrid video object tracking algorithms. Section 3 describes the proposed hybrid video object tracking algorithm. Experimental results of our approach are reported in Section 4, and Section 5 concludes the paper.

2. RELATED WORK

Several moving object detection and tracking approaches in compressed domain were proposed over the years with recent work focusing on H.264/AVC video bitstreams [4]. Poppe et al. [5] introduced a syntax level algorithm based on the size of a macroblock (MB) in bits. In an initial training phase, their algorithm creates a background model, and in the following steps new images are compared with this model to determine MBs that correspond to moving objects. Laumer et al. [6] proposed an algorithm that is solely based on MB types in H.264/AVC video bitstream to detect moving objects. They exploit the inherent motion information contained in MB types based on encoder's motion estimation process, and define MB type weights for all possible MB types. Assigned weights indicate the probability that a MB belongs to a moving object in the scene. In a following work [7], Laumer et al. extended their approach by creating spatio-temporal weight maps using MB type information, and additionally used quantization parameters (QP) of macroblocks to apply individual thresholds to the block weights in order to segment the video frames.

Approaches that combine pixel and compressed domain were also presented in the literature. Wojaczek et al. [8] used MB type information as a pre-processing step for fast evaluation of video streams. Upon detection of an object, a pixel domain person detector is employed to obtain a finer segmentation. By fusing information from the two domains, full video decoding is only performed when necessary. You et al. [9] proposed a tracking algorithm using probabilistic spatio-temporal MB filtering to segment and track multiple objects in real-time. First, they use clustering to all non-skip MBs in P frames. Then they apply spatial filtering to discard isolated MBs followed by temporal filtering to remove erroneous MBs. To refine the object trajectory, they further employ background subtraction in I frames and motion interpolation in P frames.

One of the earliest works that employed a Markov Random Field (MRF) model for tracking was published by Zeng et al [10]. They merge similar motion vectors (MV) into moving objects through minimization of the MRF energy and define different MV types. Then they treat the tracking problem as a Markovian labeling procedure on the classified MV field. Their MRF model considers the spatial continuity and temporal consistency of MVs to track moving objects. However, their method does not take into account the potential camera motion. Hence, it is only suitable for tracking objects captured by fixed cameras [4]. One of the prominent works using an MRF model was presented in [3] by Khatoonabadi & Bajić. Instead of first classifying MVs into multiple types as in [10], their method uses MV observations di-

rectly to compute a motion coherence metric. Furthermore, they employ global motion compensation to deal with camera movements and propose a new method to assign MVs to intra-coded blocks based on the MVs of their neighbouring blocks.

3. PROPOSED METHOD

Khatoonabadi & Bajić [3] treat the tracking problem in a Bayesian framework. MV information in video bitstream is used to infer the block labels $\omega^t \in \{0,1\}$ in frame t given the labels ω^{t-1} in frame $t-1$. The maximum-a-posteriori solution (MAP) for ω^t is found by maximizing

$$\omega^t = \underset{\omega}{\operatorname{argmax}} \underbrace{P_{\mathrm{mc}}(\omega^{t-1} \,|\, \omega, \kappa^t)}_{\substack{\text{motion} \\ \text{coherence}}} \underbrace{P_{\mathrm{sc}}(\kappa^t \,|\, \omega)}_{\substack{\text{spatial} \\ \text{compactness}}} \underbrace{P_{\mathrm{tc}}(\omega)}_{\substack{\text{temporal} \\ \text{continuity}}} \quad (1)$$

where κ^t is the observed motion information. The three terms in Eq. (1) correspond to three fundamental characteristics of rigid objects: motion coherence, spatial compactness, and temporal continuity. Motion coherence is defined as the relative consistency of the MVs belonging to an object. The ST-MRF model of [3] characterizes the object's motion by a single representative MV and computes the deviation of a MV of a block from this representative object MV. Spatial compactness measure is based on the observation that most rigid objects have a compact shape. Therefore, chance of a block belonging to an object is high if its spatial neighbours are known to belong to that object. Accordingly, ST-MRF model quantifies compactness by computing a weighted sum of labels assigned to the neighbourhood of the current block. Temporal continuity measures the overlap between the labeling of the previous frame.

Although [3] displays a state-of-the-art tracking accuracy for many standard test sequences, it fails to distinguish between two nearby objects with very similar MVs. One example case is illustrated in Figure 2 for the *Mobile Calendar* sequence. At around frame 100 the ball touches the train, and they start to move in the same direction. Thus, they are both assigned very similar MVs by the encoder. As a result, a part of the train is also detected as the target object, and the ST-MRF model starts to track that part of the train as well. As the ball detaches from the train, the method stops tracking the ball completely and starts to track only a part of the train. Obviously, MVs are not informative enough to resolve such tracking ambiguities.

We propose a hybrid approach to video object tracking which relies on MV information, but also takes into account the color information (Figure 1). Our approach assumes that the color of the tracked object does not change too much over the video. Parallel to the partial decoding of P-frames and processing of MVs extracted from them, I frames in the bitstream are fully decoded. As the user selects the tracking object in the first frame to initialize the labels for the ST-MRF model, we utilize the first frame to create a Gaussian Mixture Model (GMM) based on the color distributions of the object p_{obj} and the background p_{back}. For this, we use the median intensity values of 4×4 pixel blocks in YCbCr color space. We found out that four Gaussians are sufficient to model the intensity distributions in each color channel. We assume that the estimated color probabilities stay (more or less) constant throughout the sequence and calculate the posterior probability that a block color $\chi(\mathbf{k})$ belongs to the object in terms of inter-frame likelihood with respect to the frame 0.

$$p_{\mathrm{col}}(\chi(\mathbf{k})) = \frac{p_{\mathrm{obj}}(\chi(\mathbf{k}))}{p_{\mathrm{back}}(\chi(\mathbf{k}))} \quad (2)$$

Note that for efficiency reasons we compute $\chi(\mathbf{k})$ only for I frame blocks $\mathbf{k}$. For every I frame we regularize the MAP solution by adding the following color term to Eq. (1)

$$P_{\mathrm{col}}(\omega) = \prod_{\mathbf{n}:\Psi(\mathbf{n})=1} p_{\mathrm{col}}(\chi(\mathbf{n})) \quad (3)$$

where Ψ represent the candidate labeling. For efficiency reasons we do not compute color information from P-frames, i.e., $P_{\mathrm{col}}(\omega)$ is set to a constant. We optimize for ω^t using the Iterated Conditional Modes (ICM) algorithm [11].

Fig. 3: Tracking accuracies of [3] (left) and the proposed approach (right) in Frame 51 of the *Hall Monitor* sequence. Green, blue, and red colors denote true positives, false positives, and false negatives, respectively.

4. EXPERIMENTS

4.1. Experimental setup

We consider the standard test video sequences listed in Table 1. The sequences are provided in raw (YUV420) format. We carry out our accuracy evaluations on parts of the sequences for which the ground truth is available (up to 100 frames). As a benchmark for our experiments, we used the H.264/AVC (JM 18.0) encoded videos provided by the authors of [3]. They are encoded in H.264/AVC high profile with the GOP structure IPPP, i.e. with only one I frame at the beginning and subsequent P frames. In order to evaluate the performance of the ST-MRF and our proposed method

Table 1: Averages of precision, recall, and F-measure values attained by [3] and our proposed method, respectively, over various test sequences encoded with QP=28.

Method	Encoder	Measure	Mobile Calendar	Coast guard	Stefan (SIF)	Hall Monitor	Flower Garden	Table Tennis	City	Foreman	Avg.
ST-MRF [3]	H.264/AVC JM 18.0	***Precision***	75.2	50.6	66.7	65.9	65.4	89.6	86.1	91.1	**73.8**
		Recall	84.2	91.6	69.3	83.6	86.1	88.7	97.5	90.2	**86.4**
		F-Measure	78.5	63.9	67.0	73.4	73.4	89.0	91.4	90.4	**78.4**
ST-MRF [3]	H.265/HEVC x265	***Precision***	75.3	55.9	63.2	69.6	45.8	92.9	82.4	91.8	**72.1**
		Recall	85.3	90.9	63.5	79.4	92.8	85.5	98.0	89.1	**85.6**
		F-Measure	79.5	68.6	62.1	74.0	58.3	88.8	89.5	90.4	**76.4**
Proposed	H.265/HEVC x265	***Precision***	84.6	63.2	81.9	77.9	61.0	92.9	92.0	93.9	**80.9**
		Recall	83.7	89.6	46.9	72.6	88.0	85.5	87.8	85.6	**80.0**
		F-Measure	83.5	73.3	56.9	74.9	70.5	88.8	89.2	89.4	**78.3**

on H.265/HEVC encoded videos, we made slight modifications to the ST-MRF to accommodate the block structure of H.265/HEVC. We encoded all test sequences to H.265/HEVC using x265 software with a GOP structure of IPP...IPP. We adopted this GOP structure since our method needs I frames to periodically regularize the MAP solution. We chose the GOP size as 25 to compromise between the computational cost of fully decoding the I frames and improved tracking accuracy (due to frequent update of the color information). We retained the same cost coefficients from ST-MRF related to motion coherence, spatial compactness, and temporal continuity. Furthermore, we found out in a pre-study that assigning MVs to intra-coded blocks (based on MVs of their neighbours) as in [3] does not noticeably improve tracking. Thus, we did not include this technique in our analysis.

4.2. Results

In our evaluation, we analyse two different factors that might affect the performance of the considered methods: encoder choice and pixel domain information introduced by our proposed method. For the first analysis, we investigate the tracking performance of [3] over standard test sequences encoded with H.264/AVC and H.265/HEVC, respectively. For the second analysis, we compare the tracking accuracy of our proposed method on H.265/HEVC vach ideos to the tracking accuracy of [3] on H.265/HEVC and H.264/AVC videos, respectively. Table 1 shows the obtained precision, recall, and F-measure values at QP=28. Results show that the overall performance of ST-MRF with H.264/AVC and H.265/HEVC are comparable. Hence, it is shown that the ST-MRF method of [3] (with slight modifications) can track objects accurately with H.265/HEVC encoded videos as well.

Secondly, our proposed color-ST-MRF method improves precision and F-measure values significantly for most of the test sequences. Figure 3 illustrates the reduction of the false positives through our method in the *Hall Monitor* sequence. Considering the fact that our method updates the ST-MRF based labelling once at every I frame (which happens 2-3 times for most of the sequences), boosting effect of the color information on tracking accuracy is remarkable. Figure 2 displays how our proposed method outperforms [3] in the *Mobile Calendar* sequence. As discussed in Section 3, ST-MRF method of [3] fails to track the ball during a certain part of the sequence. Our proposed method exploits color information to mitigate this issue and tracks the ball fairly accurately. However, in some sequences, the addition of color information comes with the drawback of a decreased number of true positives in cases where parts of the tracking object and background have similar colors. Hence, we observe a slight degradation in recall values compared to ST-MRF in such cases.

As a further analysis, we investigate how different QP values affect the tracking accuracy of our proposed method and compare the results to [3]. Figure 4 shows that our proposed method consistently outperforms [3] in precision and F-measure for QPs $\in \{16, 20, \cdots 40\}$. We observe a decrease in precision for both methods at higher QPs. Since high quantization levels decrease the visual quality, perfor-

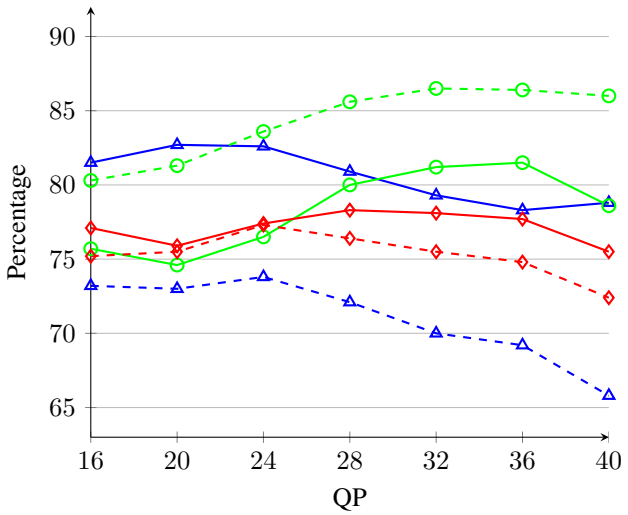

Fig. 4: Comparison of average Precision (—△—), Recall (—○—), and F-Measure (—◇—) values of the proposed algorithm and ST-MRF method of [3] (dashed) over various quantization parameters (QP).

mance of the block-based motion estimation is degraded at higher QPs. Thus, less accurate MVs are allocated to individual prediction blocks which leads to reductions in precision. Interestingly, we also observe that recall values improve at higher QPs. We believe that this is caused by the increased coarseness of MVs which causes the ST-MRF model to label a much larger area as part of the object. Hence, the number of false negatives decreases drastically increasing recall, at the expense of reduced precision.

5. CONCLUSION

In this paper, we proposed a hybrid object tracking method by combining a compressed domain spatio-temporal MRF model with a pixel-domain color-consistency measure based on the color information obtained from I frames of H.265/HEVC encoded videos. Our experimental results show that for the considered standard video sequences, our method has a better average tracking accuracy in terms of precision than the pure compressed domain ST-MRF model of [3]. Our approach is especially useful in video sequences in which the motion vectors of multiple objects are very similar to each other, such as *Mobile Calendar*. The promising results of our approach on I frames can be further improved by incorporating textual features on regions in which color is not discriminative with respect to the background. Such future enhancements are expected to increase the tracking accuracy also in terms of recall.

6. REFERENCES

[1] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in *Computer vision and pattern recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2411–2418.

[2] Alper Yilmaz, Omar Javed, and Mubarak Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, pp. 13, 2006.

[3] Sayed Hossein Khatoonabadi and Ivan V Bajić, "Video object tracking in the compressed domain using spatio-temporal markov random fields," *Image Processing, IEEE Transactions on*, vol. 22, no. 1, pp. 300–313, 2013.

[4] R Venkatesh Babu, Manu Tom, and Paras Wadekar, "A survey on compressed domain video analysis techniques," *Multimedia Tools and Applications*, pp. 1–36, 2014.

[5] Chris Poppe, Sarah De Bruyne, Tom Paridaens, Peter Lambert, and Rik Van de Walle, "Moving object detection in the h. 264/avc compressed domain for video surveillance applications," *Journal of Visual Communication and Image Representation*, vol. 20, no. 6, pp. 428–437, 2009.

[6] Marcus Laumer, Peter Amon, and Andreas Hutter, "Compressed Domain Moving Object Detection based on H.264/AVC Macroblock Types," *Proceedings of the International Conference on Computer Vision Theory and Applications*, pp. 219–228, 2013.

[7] Marcus Laumer, Peter Amon, Andreas Hutter, and Andre Kaup, "Compressed domain moving object detection by spatio-temporal analysis of H.264/AVC syntax elements," *2015 Picture Coding Symposium (PCS)*, pp. 282–286, 2015.

[8] Philipp Wojaczek, Marcus Laumer, Peter Amon, Andreas Hutter, and André Kaup, "Hybrid Person Detection and Tracking in H.264/AVC Video Streams," *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, pp. 478–485, 2015.

[9] Wonsang You, MS Houari Sabirin, and Munchurl Kim, "Real-time detection and tracking of multiple objects with partial decoding in h. 264/avc bitstream domain," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009, pp. 72440D–72440D.

[10] Wei Zeng, Jun Du, Wen Gao, and Qingming Huang, "Robust moving object segmentation on h. 264/avc compressed video using the block-based mrf model," *Real-Time Imaging*, vol. 11, no. 4, pp. 290–299, 2005.

[11] Julian Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302, 1986.