Non-Blind Structure-Preserving Substitution Watermarking of H.264/CAVLC Inter-Frames

Thomas Stütz, Florent Autrusseau and Andreas Uhl

***Abstract*—In this work we propose a novel non-blind H.264/CAVLC structure-preserving substitution watermarking algorithm. The proposed watermarking algorithm enables extremely efficient watermark embedding by simple bit substitutions (substitution watermarking). The bit-substitutions change the motion vector differences of non-reference frames. Furthermore our watermarking algorithm can be applied in applications scenarios which require that watermarking preserves the length of the bitstream units (structure-preserving watermarking). The watermark detection works in the image domain and thus is robust to video format changes. The quality and robustness of the approach are in depth evaluated and analyzed, the quality evaluation is backed up by subjective evaluations. Comparison to the state-of-the-art indicates a superior performance of our watermarking algorithm.**

I. Introduction

H.264 watermarking has been researched intensely[1] and is of great interest due to its wide applicability in the context of DRM (digital rights management). This paper presents a novel H.264 CAVLC watermarking technique that allows to implement watermarking by simple and efficient bit substitutions of the compressed bitstream (substitution watermarking). Additionally our algorithm is structure preserving, i.e., precisely preserves the length of the bitstream and even of the bitstream's smaller units. In the case of H.264, structure preserving watermarking denotes watermarking algorithms in which the network-abstraction layer units (NAL units / NALUs are small units which form the entire H.264 bitstream) have exactly the same length in the watermarked content and the original content. The structure preservation for H.264 is required for the watermarking of Blu-Ray content. The length preservation is required as the video has to fit on a Blu-Ray disc. The internal structure has to be preserved as often byte-based addressing schemes are employed in production and presentation, e.g., the meta-data on Blu-Ray discs employs byte-based addressing schemes. Blu-Ray players can enhance the presentation with additional online content, which employs byte-based addressing (BD-J) as well.

Another important application scenario of our approach is the online distribution of video content, which benefits from efficient adaptation of the content to the user's device requirements and to the user's current bandwidth. H.264 enables the efficient adaptation of content based on simple bitstream operation, e.g., the aspect ratio can be changed by dropping parts of the bitstream. The scalable extension of H.264/CAVLC offers even more adaptation possibilities (bitrate, resolution, frame rate, quality), which are also implemented with simple bitstream operations. As meta-formats for adaptation employ byte-addressing, structure-preserving watermarking works well together with such adaptation operations, as the byte-addressing remains unchanged.

So there are important applications, which require structure-preserving H.264 watermarking. In our proposed watermarking algorithm the embedding stage is split into an analysis and a substitution stage; analysis must only be conducted once, afterwards the embedding of different marks requires only extremely light-weight bit substitutions. Thus the embedding of numerous marks in real-time with very low computational complexity is possible; of utmost importance for streaming individually marked content to numerous clients.

Our main contribution is the proposal of a new non-blind structure preserving H.264 CAVLC watermarking approach. A further contribution is the thorough analysis of the approach with respect to robustness and quality. The quality evaluation not only employs state-of-the-art quality assessment tools, but also presents an actual subjective quality evaluation. Our evaluations focus on 720p content (the leading mobile phone's resolution and also occasionally employed for Blu-Ray content).

A brief review and comparison to the state-of-the art of H.264 watermarking with a focus on structure-preserving watermarking is presented in section II. In section III an overview of H.264 is given, while section IV briefly summarizes quality evaluation of visual data. Our structure-preserving H.264 CAVLC watermarking approach is presented in section V. Experimental results with respect to quality and robustness are presented in section VI. Finally section VII concludes the paper.

II. Previous work

There is a considerable amount of scientific literature on structure-preserving watermarking [1][2][3] for the MPEG-2 format, which is a simple video compression system compared to H.264. For MPEG-2 structure-preserving watermarking is easier as compared to H.264, because MPEG-2 employs less prediction and content-adaptive coding. Therefore MPEG-2 offers more syntax elements which can be simply replaced in the

T. Stütz is with the Department of Multimedia Technology, University of Applied Sciences Salzburg, 5412 Puch/Salzburg, Austria, e-mail: thomas.stuetz@fh-salzburg.ac.at

A. Uhl is with the Department of Computer Sciences, University of Salzburg, 5020 Salzburg, Austria, e-mail: uhl@cosy.sbg.ac.at

F. Autrusseau is with the LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, rue Christian Pauc, BP 50609, 44306 Nantes, France

Special thanks to SONY DADC, in particular Reinhard Blaukovitsch, for the cooperation in the project and the insights into industry requirements. This work has been partially supported by the FFG bridge project 834165.

[1] ACM digital library reports 180 publications on H.264 watermarking. IEEE Xplore reports 72 publications on H.264 watermarking.

bitstream and employed for structure preserving watermarking. Probably due to the difficulty of implementing structure-preserving watermarking for H.264 the corresponding literature is far less extensive: only one approach can be employed for structure-preserving H.264 CAVLC watermarking [4] and only one for structure-preserving CABAC watermarking [5]. However, there is a considerable amount of literature for H.264 watermarking. Many schemes exploit the H.264 encoding process to embed the watermark during compression. The main advantage of such approaches [6] is that the error introduced by watermarking is not propagated further (at the cost of some bitrate increase). Other schemes work on the bitstream, mostly to reduce the computational burden of compression-integrated watermarking schemes. It has to be noted that H.264 bitstream watermarking actually performs entropy-decoding, such that the syntax elements can be accessed, watermarked (e.g., the quantized DCT coefficients), and again entropy-encoded. The approaches presented in [7] are examples for H.264 bitstream watermarking. Most related to our application requirements is the setup in the work of Zou and Bloom [8], [4], that discusses substitution watermarking for intra frames of H.264 CAVLC bitstreams, but is not capable to watermark inter coded frames (the vast majority of frames is commonly coded as inter frames, some encoders use intra frames only once at the start of a sequence). Thus methods for substitution watermarking of inter frames are needed. The approach of Zou and Bloom [4] modifies the intra-prediction modes which can be implemented by bit-substitutions of H.264 CAVLC bitstreams. Suitable substitutions have to be found in a complex analysis stage, which has to consider intra and inter drift, while our algorithm offers a lightweight analysis stage. Furthermore the marking space, i.e., the number of watermarkable blocks, of our approach is larger than the marking space of [8], [4], [5] (see section VI-D) and thus fewer frames are needed for watermark embedding. Alternatively the larger marking space can be employed to improve the robustness of our approach (a lower detection threshold can be selected for the same probability of alarm). A substitution watermarking algorithm for CABAC, based on motion vector data changes, was presented by the same authors in [5]. However, CABAC and CAVLC are entirely different, and thus the applicable changes are different. Motion vector data are encoded context-adaptively in CABAC, and thus a computationally complex analysis stage is required in the approach of [5], while our CAVLC approach is extremely lightweight in comparison. Additionally the number of candidate changes is smaller by an order of magnitude and thus the CAVLC algorithm performs better in terms of a larger marking space, which reduces the number of watermarked frames or leads to an improved robustness (see section VI-D).

TABLE I
CAVLC: CODING OF MVDS

Index	Codeword	MVD
0	1	0
1	010	1
2	011	-1
3	00100	2
4	00101	-2
5	00110	3
6	00111	-3
7	0001000	4
8	0001001	-4
9	0001010	5
10	0001011	-5
11	0001100	6
...	...	...

III. Overview of H.264

The design of H.264 follows the classic hybrid video coding approach [9]. The frames are processed in 16x16 macroblocks. Each macroblock can be predicted using previously processed macroblocks of the same frame (intra-prediction) or other frames (inter-prediction). The macroblocks can be further subdivided (sub-macroblock partitions), the smallest block size is 4x4. A coded video sequence always starts with the coded data of an intra-predicted frame (I frame). The distortion of I frames spreads on all subsequently decoded frames due to inter prediction. After an I frame inter-predicted frames that may use one reference frame (P frame) or two reference frames (B frame) follow. Frames (even P and B frames) may be used as reference, frames which are not used for inter-frame prediction are called non-reference frames. Inter-prediction is conducted by motion estimation and motion compensation, which are conducted with quarter pixel accuracy. The motion vectors (MVs) of a block are predicted by neighbouring blocks (a detailed description can be found in [10]) and the motion vector difference (MVD) is actually coded in the bitstream (which codes quarter pixel differences). There are two distinct coding modes in H.264, namely CAVLC and CABAC. CAVLC is computationally less expensive (at the cost of a lower compression performance) and thus is employed in cases where computational complexity constraints outweigh the compression performance. Typical applications are in the context of mobile devices (720p has become the resolution of the lead devices), where power and computational constraints outweigh compression. Furthermore 720p content on Blu-Ray discs can be coded with H.264/CAVLC: there is no need for higher compression as 720p typically fits on a Blu-Ray anyway. In H.264/CAVLC MVDs are not coded context-adaptively, but with variable-length signed exponential Golomb codes. Table I shows the coding of MVD values, each MVD (last column) is coded by an exponential Golomb code (in the column labelled "Codeword"). A separate MVD is coded for the x- and y-direction.

IV. Quality assessment

As briefly mentioned in the introduction, this work mainly focuses on evaluating the robustness and quality performances of an H.264 CAVLC watermarking operating within the compressed bitstream. In this section, we present both subjective and objective quality assessment of watermarked contents.

A. Subjective Quality Assessment

Ultimately, the quality of the marked content will be judged by human observers, and thus, running a subjective experiment is the best way to evaluate the impact of the watermark on the quality of the protected image/video. During subjective tests, quality (or annoyance) scores are collected from human observers within a controlled environment. Subjective experiments have been of high interest for many decades among the scientific community. Early experiments were conducted to determine an optimal viewing distance on television monitors [11], or the detection threshold of a simple spot on a CRT screen [12], and led the researchers to do an attempt to estimate the subjective quality [13]. Evidently, subjective experiments had to be standardized, in order for other researchers to be able to reproduce and/or compare the results. Thus, the International Telecommunication Union (ITU) has published various reports and recommendations for conducting subjective experiments. Both the Radiocommunication (ITU-R) and Telecommunication (ITU-T) sectors of the ITU are regularly issuing some recommendations for quality assessment (both objective and subjective) of digital images and videos. Two recommendations of particular interest are [14] and [15]. The recommendation [15] notably specifies the viewing conditions, monitor settings (resolution, contrast), the importance of anchoring is highlighted. It is for instance advised to use at least 15 non expert observers. Some advices are given on the duration of the experiment, and on the possible protocols to use. Among the most commonly used protocols, we can cite the ”Double-Stimulus Impairment Scale” (DSIS) and the ”Double-Stimulus Continuous Quality Scale” (DSCQs). In [14], alternative protocols are suggested, the ”Absolute Category Rating” (ACR) or the ”Pair Comparison” methods are amongst the most common methods. Commonly, the outcome of a subjective experiment is to collect the mean opinion scores (MOS) from the observers for the given input subjective dataset. The MOS are simply computed by averaging the collected scores of all observers for a given content. The alternative to subjective experiments is to utilize objective quality metrics (OQM), which are methods whose goal is to predict the perceived quality. In the upcoming sub-section, we will review the most common types of OQMs.

B. Objective Quality Assessment

Objective quality metrics are mainly of two types. On one hand, statistical quality metrics are very widely used, PSNR, RMSE, or SSIM belong to this category. On the other hand, advanced HVS-based OQMs (such as VIF[16], VSNR[17], CPA[18] or C4[19]) exploit some properties of the human visual system (such as contrast sensitivity, contrast masking, or luminance adaptation) to provide a prediction of the MOS (predicted mean opinion scores are commonly referred to as MOSp). Thus, once the subjective scores are collected (MOS are gathered), and the MOSp computed for a given set of metrics, a metric performance evaluation is performed. The Video Quality Experts Group (VQEG) issued a report in 2008 [20] providing an analysis of various assessment methods as well as several tools that can be used to evaluate the performances of objective quality metrics. Statistical or advanced HVS metrics could be either full reference (FR) or reduced reference (RR) or even no reference (NR). For a FR metric, the original content is needed as an input, along with the distorted content which needs to be assessed. A RR metric needs the content to be assessed along with a reduced set of features from the original content to compute the MOSp. Finally, NR metrics only need the distorted content as an input in order to provide a prediction. Usually, FR metrics exhibit better performances at predicting the MOS. In the following, only FR metrics are used.

V. A Novel Inter Frame Substitution Watermarking Algorithm

Our proposal for a novel H.264 CAVLC watermarking algorithms takes advantage of MVD modifications. Our watermarking algorithm modifies original MVDs such that the MVD of the watermarked content has equal length as the original MVD and thus the length of the NAL units is preserved as well.

The developed watermarking algorithm is robust, non-blind (additional information is required for detection), and zero-bit (only the presence of a watermark will be detected) [21]. The watermarking process is easily reversible, i.e., the watermark can be losslessly removed by a simple substitution of the original bit-sequences. In this paper we only present results for watermarking non-reference inter frames, which do not cause any inter frame drift and thus enable very efficient distortion assessment.

A. Embedding

The watermark embedding aims to alter a distinctive feature (in the current implementation average luminance) of applicable macroblocks. The macroblock’s average luminance is changed by altering its MVD. In the current implementation only macroblocks of type P16x16 are analyzed (the majority of the macorblocks of B frames are commonly of this type). Furthermore the MVD change modifies an entire 16x16 block of image data, which enables robust detection. The embedding consists of two stages: an analysis stage and a substitution stage (see fig. 1). Input to the embedding are the H.264 bitstream to be watermarked and the watermarking parameters, such as the key for random watermarking bit generation and a quality control parameter MbDist (macroblock distortion), which is used to control the embedding distortion. Only macroblock MVD changes are considered for watermarking which result in a distortion less than MbDist (in this work we use the mean squared error for distortion estimation). The distortion is computed between the original macroblock and the MVD changed and reconstructed macroblock. Output of the process are the watermarked bitstream and little additional information for detection (“Detection Info” in fig. 1). In the analysis stage each macroblock is checked for watermarking suitability, thereby several conditions have to be met such that a macroblock is employed for watermarking. Only inter predicted macroblocks are considered for watermarking, and each length-preserving and sign-preserving MVD change is

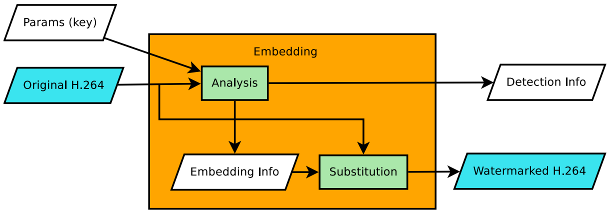


Fig. 1. Watermark embedding

evaluated. Only codes with the same sign (either codes contained in dashed red boxes or solid blue boxes in table I) and same length are evaluated. First the quality is checked (after application of the change) which is done by computing the MSE (mean squared error) between the original macroblock (with the original MVD) and the modified macroblock. Only if the obtained MSE is below the quality control parameter MbDist the change is considered valid. With the parameter MbDist the embedding strength can be adjusted. Second the impact on the feature (avg. luminance) is computed, which is later used in the detection process. If a macroblock has at least one MVD change that increases the feature and at least one MVD change that decreases the feature, the macroblock is suitable for watermarking. The change with the strongest increase is employed to encode a 1, while the change with the strongest decrease is employed to encode a -1. The macroblock position and frame number and the original feature are recorded in detection info. The watermark embedding algorithm is summarized briefly as follows.

For each inter-predicted macroblock:

- Evaluate original block's feature (avg. luminance)
- Apply length-and-sign-preserving MVD change and check
  - Embedding distortion is below MbDist (MSE).
  - Feature difference is sufficient (avg. luminance larger then, e.g., 0.25)
- If there are two groups of changes (increase feature, decrease feature) use the 'decrease feature' change to encode a minus one, and the 'increase feature' change to encode a one.

B. Detection

The detection can be performed in the image domain and does not require any re-encoding. As the presented approach is non-blind, we can assume perfect registration / alignment (temporal and spatial). The actual implementation of registration is well covered in computer vision literature and can for example be solved by storing SIFT interest points [22] of the watermarked frames as well as the detection information. These features can later be used to register the content. Specific solutions for watermarking have also been proposed [23], [24].

The detection process can be divided into three distinct tasks, bit extraction, correlation and decision (see figure 2). The overall watermarking system including detection is illustrated in figure 3. The bit extraction takes advantage of the detection info, it computes the feature (avg. luminance) of a possibly watermarked macroblock and compares it to the recorded original feature. If the computed feature is larger or equal, a 1 is extracted; if it is smaller a -1 is extracted. In the correlation step, the extracted bit sequence $\vec{e}$ is compared to the possibly embedded watermark bit sequence $\vec{w}$. The number of embedded bits is n. More precisely, the detector response z is computed by:

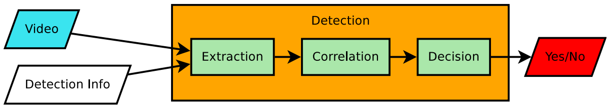


Fig. 2. Watermark detection

$$z = 1/n \sum_{i=1}^{n} e_i \times w_i$$

Finally depending on the detector response and a user-defined false positive probability a decision is made. Thereby the detector response is compared to a detection threshold $\mathcal{T}(p_{fp})$, and if the detector response is larger than $\mathcal{T}$ the watermark is decided to be present. The user-defined false positive probability determines how likely it is to detect the watermark in content that does not contain the watermark. Overall the algorithm for detection can be briefly summarized as follows:

For each possibly watermarked macroblock in the image domain:

- Compute the block feature (avg. luminance) of the possibly watermarked macroblock and compare the block feature to the original feature and return -1 for a decrease and 1 for an increase.
- Compute detection statistic (a measure of correlation between embedded and extracted sequence).
- Decide whether watermark is present or not.

This kind of watermark detection is referred to as detection based on hard decision decoding [25, sect. 2.4.2.1] and is well known and has already been analysed.

The analysis of the distribution of the detector response can be divided into two cases, the watermark has not been embedded ($\mathcal{H}_0$) and the watermark has been embedded ($\mathcal{H}_1$).

After embedding and without any distortions (e.g. recompression) the detector response will always be 1, as we do not have any inaccuracies / randomness in the embedding and detection processes and thus each bit will be correctly extracted. Distortions from recompression and other signal processing operations will introduce errors and shift the detector response for watermarked content slightly towards zero.

In case of $\mathcal{H}_0$ the distribution of the detector response follows a Binomial distribution, which can be approximated by a Gaussian distribution.

Figure 4 gives an overview how the number of embedded bits and the false positive probability determine the appropriate threshold for detection. The more bits are embedded the lower the threshold for a given false positive probability can be chosen.

For our application scenarios, the exact determination of the false positive probability is of ultimate importance. In case of of a false positive in the online distribution scenario, i.e.,

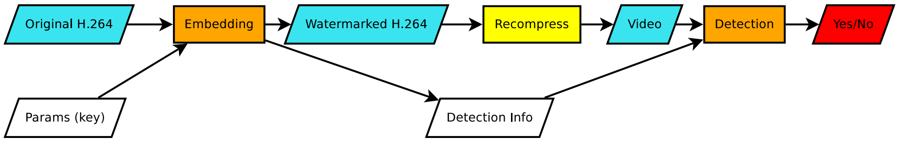


Fig. 3. Watermarking overview

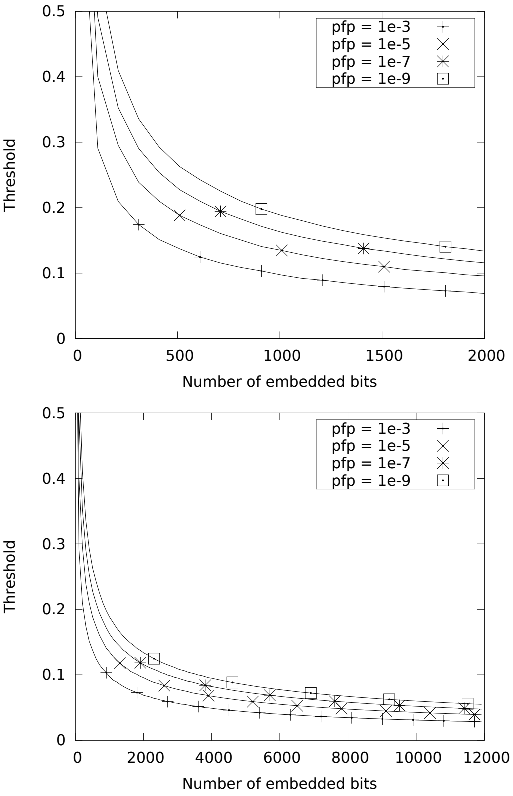


Fig. 4. The threshold for various probabilities of false alarm as a function of the number of embedded bits

a watermark is detected in pirated content, a user is found guilty of distributing the content and certain actions (legal or technical) are taken, which should prevent the user from further piracy. Given that the user is innocent the rage and the following (social) media coverage can severely harm the business of the content distributor. Such an event is greatly feared and its risk must be known to be very low.

VI. Experiments

In the following we present results on the basis of 4 different 720p sequences (Canal, Depart, Ebu, Elephant) with 250 frames. The sequences reflect different natural content, as well as one artificially generated sequence (Elephant). The sequences have been encoded with H.264 CAVLC with a Qp of 13 (very high quality). The lower the quality (higher Qp) the more macroblocks are coded in P16x16 (as the frames become smoother and larger partitions more efficient). Thus our approach works even better for lower quality video. An I(BP)* prediction structure with non-reference B-frames is employed, which is a common and reasonable configuration. The Blu-ray specification even requires B-frames to be non-reference frames. Only P16x16 macroblocks (having no sub-partitions) of B-slices have been employed for watermarking. The search range for candidate MVD changes has +/- 16 in each direction and the MVD change with the maximum feature difference below the distortion threshold was selected. Furthermore we rejected MVD changes that modified the average luminance feature by less than 0.25.

TABLE II
Average number of embedded bits per seconds for each sequence and embedding distortion

Sequence / MbDist	100	25	4
Canal	154.0	81.4	9.8
Depart	294.2	138.4	2.4
Ebu	407.2	277.3	11.3
Elephant	214.6	210.9	111.3

Our watermarking algorithm can be employed with different values of the embedding distortion parameter MbDist (distortion measured in MSE). Both the MbDist and the source material (video) have an impact on the number of bits that can be embedded (see table II for results on our test sources). The encoder decides on the basis of the source video which macroblocks are encoded as P16x16 blocks. The analysis stage of watermark embedding only chooses MVD changes which result in a macroblock distortion (in MSE) that is below MbDist. Therefore the reduction of MbDist severely reduces the set of candidate MVD changes for highly textured sequences (natural video content, especially for the Depart sequence), while the reduction of MVD changes is far less severe for computer generated content, such as the Elephant sequence (because of the smoothness of the computer-generated content). For highly textured sequences slight spatial shifts (the result of a MVD change) lead to higher MSE distortions.

In the next section we present evidence that even the highest embedding strength offers very good / excellent quality.

A. *Quality Evaluation*

A subjective experiment was conducted, 42 observers were enrolled, their acuity was checked as well as normal color vision. The ACR (absolute content rating) protocol was used. For this protocol, a single video sequence is displayed in the center of the screen, and the observer is asked for a quality score after every displayed sequence. The resolution of the tested video sequences was 1280 × 720. The quality score was (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad), MOS were computed across the 42 observers, as well as standard deviations for every content across observers, the maximum standard deviation was below 1 (0.93). During the experiment, 56 videos were evaluated, the database was built as follows. Among the four input sequences three were collected from the VQEG datasets (Canal, Depart and Ebu) and one sequence was an artificial (cartoon) sequence (Elephant). Every sequence was either watermarked and re-encoded, or only re-encoded. The watermarking technique was presented in section V. Six quantization parameters were used (Qp = 24, 28, 32, 36, 40, 44) for both watermarking and re-encoding scenarios. The original input sequences and watermarked sequences were also considered in the experiment (using a Qp of 13). Overall four original sequences have been subjected to two distortions (watermarking and coding or coding only) and each resulting sequence has been re-encoded with 7 quantization parameters. In summary this results in a dataset of 56 sequences. Each sequence was 10 seconds long (250 frames). For each observer, the experiment duration was in between 15 and 20 minutes.

The main objective of splitting the dataset into two parts: watermarking and coding was to analyse any perceptual quality loss due to the watermark embedding. To this end, figure 5 shows a histogram representing the difference between watermarked and coded sequences.Positive x-axis values means that the watermarked sequences presented a higher quality score than the coded version (and negative values along the x-axis means the coded sequence had a higher quality score). The y-axis simply counts the number of occurrences for all observers and for all sequences. As we can notice on this figure, the histogram bins are symmetrically distributed around zero, The symmetry of the histogram indicates that the differences in quality perception between watermarked and non-watermarked videos is random, i.e., there is no difference in the perceived quality of watermarked or non-watermarked videos.

Figure 6 shows the Mean Opinion Score as a function of the quantization parameter for all tested sequences. The gray lines represent the differences between marked and coded sequences. It is interesting to notice that these differences are centered around zero, which means that depending on the sequences and the Qp, the allocated quality score could either be higher for the watermarked sequence or for the coded sequence. On this figure, the symbols (arbitrarily positioned at Qp=22) represents the original marked (squares) and coded (diamonds) sequences.

Moreover, five Objective Quality Metrics were tested on this subjective dataset (PSNR, SSIM[26], CPA1[27], CPA2[18] and VIF[16]). The performances of the metrics were assessed in terms of wRMSE, RMSE, Rank correlation, Outlier Ratio, Kappa coefficient, and Linear correlation. Table III provides the metrics performances for all five metrics. It is obvious from table III that the VIF metric outperforms all others for the six tested performances tools, and PSNR exhibits the worst overall performances. Thus, in the following our analysis will focus on these two metrics.

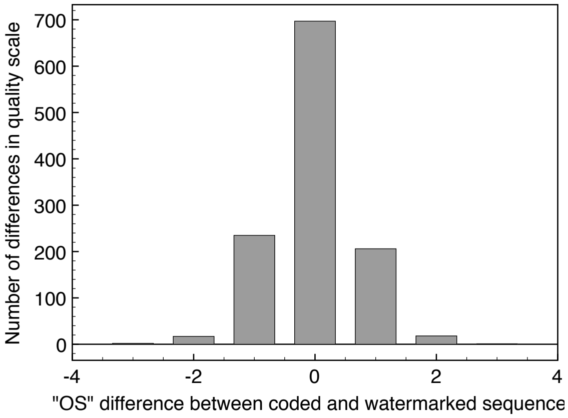


Fig. 5. Watermarked versus coded: differences in opinion scores

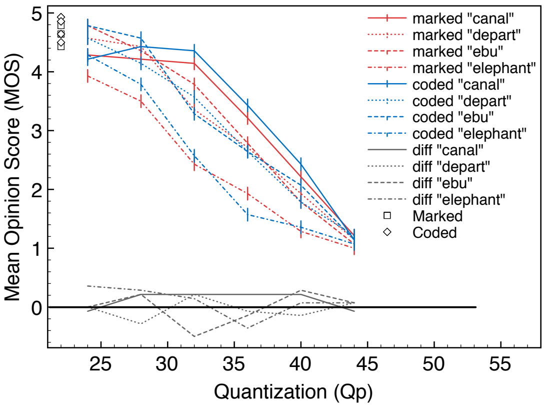


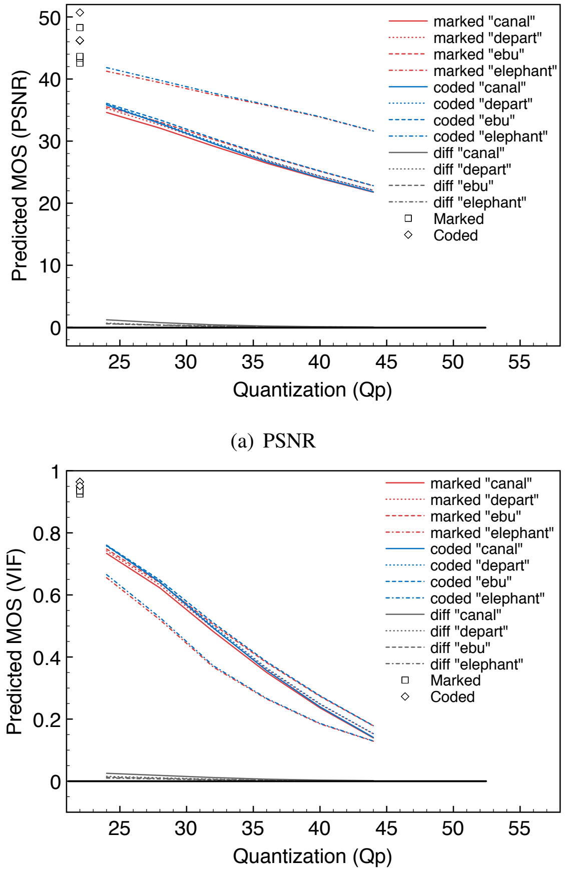
Fig. 6. Subjective mean opinion scores for three coding parameters (different Qp)

Figures 7(a) and 7(b) respectively show the MOSp for PSNR and VIF as a function of Qp. We can observe that neither PSNR, nor VIF can notably differentiate coded sequences and watermarked ones. It is interesting to notice that both metrics disagree concerning the assessment of the Elephant sequence (computer generated sequence). A further analysis showed that for all tested metrics, except VIF, the predicted scores for the Elephant sequence were seen as presenting a significantly higher perceptual quality than the other 3 sequences. This explains the overall low metrics performances shown in table III. This particular behavior is clearly visible on figure 8(a) representing the PSNR plotted as a function of the MOS. For low to good quality, the PSNR values for the Elephant sequence are about 9dB higher than the remaining sequences, whereas the MOS for this sequence was slightly below others (see figure 6). Such a behavior is not apparent for the VIF metric (see figure 8), which presents a linear distribution of MOS versus MOSp, and as we have seen above, is capable of discriminating the Elephant sequence as having an overall lower quality.

Given our subjective results we conclude that watermarked

TABLE III
OBJECTIVE QUALITY METRICS PERFORMANCES

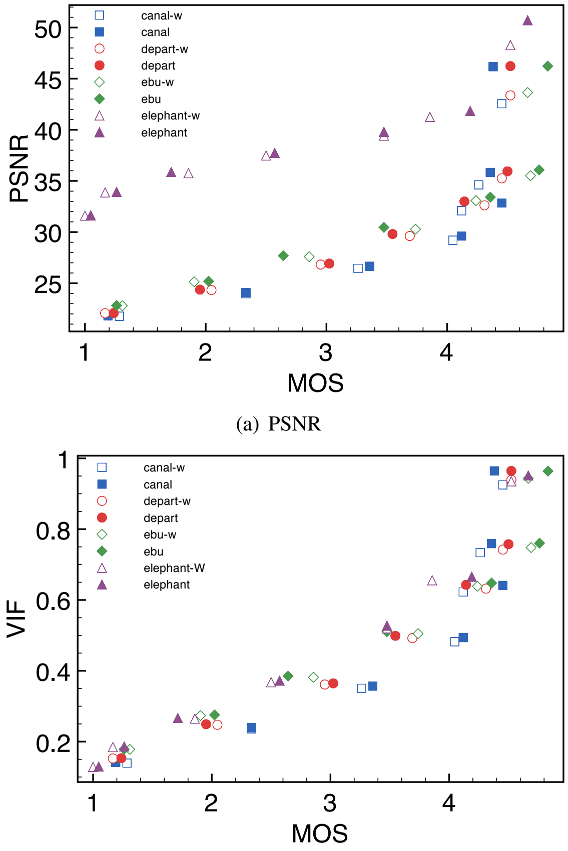
	wRMSE	RMSE	RankCorr	OR	Kappa	LinCorr
PSNR	13.6665	0.9419	0.6792	0.7321	0.1754	0.6679
SSIM	12.2094	0.8328	0.7695	0.6071	0.3345	0.7544
CPA1	7.6107	0.7189	0.8090	0.4643	0.4393	0.8229
CPA2	7.7724	0.6481	0.8666	0.4464	0.5186	0.8589
VIF	1.1941	0.2422	0.9621	0.1607	0.8146	0.9815



(a) PSNR

(b) VIF

Fig. 7. OQM predictions on coded or watermarked sequences for various Qp



(a) PSNR

(b) VIF

Fig. 8. OQMs plotted as a function of the mean opinion score

content can not be visually distinguished from not watermarked content (when played as video). This may come surprising as the allowed embedding distortion is high with a MbDist of 100 in terms of MSE. However, the watermarking approach implicitly takes advantage of temporal masking effects, as due to the algorithm design the watermark is always embedded in high motion areas, in which distortion is perceived less pronounced by human observers.

B. Robustness Evaluation to H.264 and H.263

The robustness of our watermarking algorithm highly depends on the embedding strength as defined by the parameter MbDist as the number of embedded bits is primarily determined by this parameter. Recompression is the main focus of our robustness evaluation, most importantly recompression with H.264 and H.263. The employed software for recompression was x264[2] (ultrafast, varying quality parameter Qp) for H.264 compression and ffmpeg[3] (vcodec=mpeg4, varying quality scale Qs) for H.263 compression. The chosen quality ranges correspond to qualities from excellent to bad for both H.264 and H.263.

The false positive probabilities in the figures 10, 12, 15 and 16 have been derived from the Binomial distribution [25]. For watermarked content the false negative probabilities have been derived from a Gaussian distribution which has been fitted to the data generated by 50 different watermarking keys [25], [21] (in figures 15 and 16).

Figure 9 plots the detector response against the Qp employed in x264 compression. Results are given for 4 different sequences and different embedding strengths (MbDist). A detection threshold has to be chosen such that it separates the detector responses of un-watermarked content (dashed blue lines in fig. 9) and watermarked content (solid red lines in fig. 9). For MbDist 100 and 25 the selection of detection threshold that separates un-watermarked from watermarked content is obviously possible, and even for a MbDist of 4 the detector response for an un-watermarked content is always below the detector response of the same watermarked content. While this figure can only give a first impression, a better interpretation of the detector values is obtained if we consider the associated false positive probability for the obtained detector response. Figure 10 plots the exponent of $1/\texttt{pfp}$ in basis 10, i.e., a value of 8 corresponds to a false positive probability of $1/10^8$, against the Qp. Note that the negative exponents of the $\texttt{pfp}$ have been clipped at 8. We notice high robustness to H.264 compression even for very bad quality and even for low embedding strengths, only at

[2] x264 0.85.1448 Ubuntu_2:0.85.1448+git1a6d32-4

[3] FFmpeg version SVN-r0.5.1-4:0.5.1-1ubuntu1.2, Copyright (c) 2000-2009 Fabrice Bellard, et al

MbDist=4 the detection performance decreases significantly (although even this limited robustness may be sufficient for the protection of high quality content). We further have to point out that the ultrafast settings of the x264 encoder can be considered the worst case, as these settings introduce heavy distortions (but are fast). Thus the presented results correspond to a worst case scenario for robustness, other encoders (or other settings for x264) preserve a better quality and thus the watermark is more reliably detected.

Even better is the robustness to H.263 compression as summarized in similar figures (see fig. 11 and 12).

However, single detector responses only represent a single sample from a random experiment; a more thorough analysis has to draw several samples from the random experiment in order to enable a statistical analysis of the underlying distributions. Given that the distortions are computationally expensive (repeated H.264 and H.263 encoding / decoding) the more extensive analysis focuses on medium quality recompression which can be considered the default case in many application scenarios (e.g., illegal file sharing). In the following experiments 50 watermarking keys have been employed both for detection in content watermarked with the same key ($\mathcal{H}_1$) and content that has not been watermarked with the same key ($\mathcal{H}_0$). The experiments have been conducted for different embedding strengths (MbDist). The figures 13 and 14 contain histograms of the detector response under $\mathcal{H}_0$ (on the left, detector responses distributed around 0) and $\mathcal{H}_1$ (on the right). The resulting distribution under $\mathcal{H}_0$ follows approximately a Gaussian distribution, the parameters can be either estimated on the basis of the obtained detector responses (the dashed line in the histograms) or exactly by the Binomial distribution (the solid line with upward triangles). We notice that the prediction on the basis of the Binomial distribution is very close to the fitted Gaussian distribution. The solid line with downward triangles is the fitted Gaussian distribution for the detector responses of watermarked content. If the embedding strength is reduced the detection performance slowly decreases, i.e., the distributions of Z under $\mathcal{H}_0$ and Z under $\mathcal{H}_1$ are less and less separated. However, even for an embedding strength of 25 the distributions are clearly separated, and for a embedding strength of 4 many sequences still present well separable distributions. The same information (as contained in histograms) can be plotted in ROC (receiver operation curves). The results are only shown for the Depart and the Elephant sequence as the results of these two are the most different, the performances of the other sequences are in between. The different behaviour of these two sequences is due to the different video characteristics, on the one hand the relatively smooth computer-generated Elephant sequence on the other hand the highly textured Depart sequence, with high, but local and independent motion (it contains a sequence of a cross country running race, each runner moves independently of the others). For the Depart sequence there are simply too few MVD changes that result in a distortion below an MSE of 4. On the other hand the MSE penalty of an MVD change in a smooth sequence, such as Elephant, is far less pronounced, resulting in many watermarkable blocks and thus a higher robustness.

Figures 15 and figures 16 show ROC plots for common recompression attacks with H.264 and H.263. In the ROC plots the exponent (of basis 10) of the false positive probability (x-axis) is plotted against the false negative probability (y-axis). Thus an x-axis value of -8 corresponds to a probability of 10^{-8}. The closer an ROC curve is to the axes the better is the performance of the associated watermarking scheme (the proposed algorithm with different parameters of MbDist). We plotted results starting from a very high false positive probability of 10^{-1} to a a low false positive probability of 10^{-9}, which should contain results for most practical systems. The schemes with MbDist=100 and MbDist=25 exhibit excellent performance for all sequences and both distortions (see figures 15 and 16). Only for a MbDist=4 problems are encountered as too few candidate blocks are found. The number of watermarkable blocks depends on the source characteristics, MVD changes in highly textured content result in an higher MSE. In a practical system, one would simply need to embed the watermark into more frames, i.e., a longer sequence.

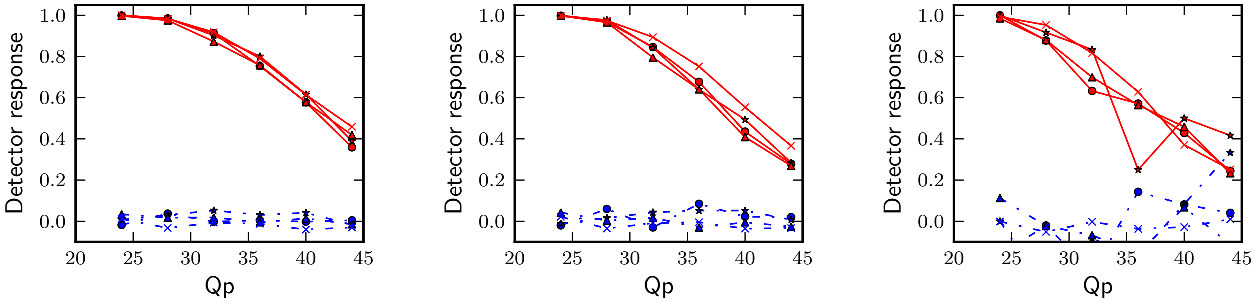
In conclusion, at a MbDist of 25 and above we can extremely reliably detect the presence of the watermark even for highly compressed sequence (both H.264 and H.263) and even at lower embedding strength many sequences show a very good detection performance.

It is also notable, that the detector responses do not change significantly for the different embedding distortions, only the number of embedded bits increases significantly with higher embedding distortions. The higher the number of embedded bits the lower the detection threshold can be chosen for a given false positive probability. Thus lower embedding distortions solely require to watermark more frames to achieve higher robustness.

C. *Watermarking Attacks*

In the following we present results for a relevant subset of the Checkmark watermarking evaluation framework [28] and additionally we present first results for a targeted attack on our watermarking system. As the embedding distortion mainly influences the number of embedded bits, only results for an embedding distortion of 100 (MbDist) are given in tables IV and V. Although the attacks significantly reduce the detector responses, the decrease is not a problem as long as the number of embedded bits is sufficient. The number of embedded bits can be simply increased by watermarking more frames (our results are only for 10 seconds video clips).

While the Checkmark attacks are generic attacks against watermarking systems, an attacker might exploit the specifics of a watermarking system to design a targeted attack. For our approach an attacker can analyze the H.264 bitstream and identify candidate macroblocks, i.e., macroblocks which have equal-length MVD options. As an attacker does not have access to the original, she can not determine whether the macroblock has been considered watermarkable. The MVDs of candidate macroblocks then can be set to a canonical value, e.g., the minimum MVD value. This attack removes the watermark, but also introduces heavy distortions, i.e., the attacked video suffer from flicker and many frames show a

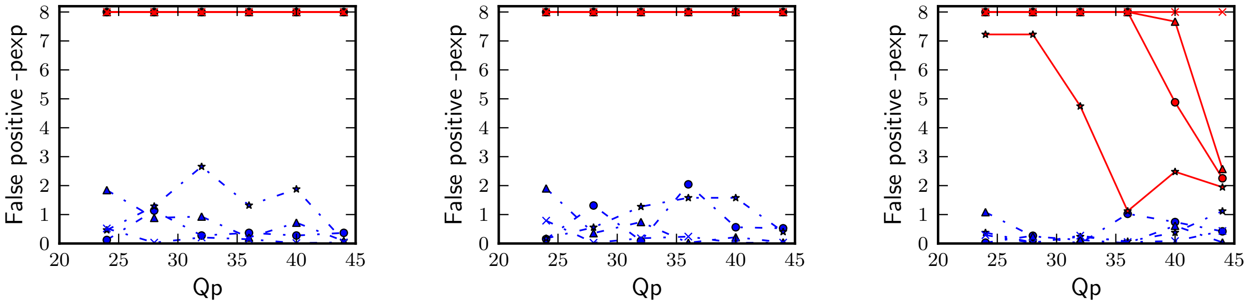


(a) MbDist=100

(b) MbDist=25

(c) MbDist=4

Fig. 9. Detector response for varying Qp of x264 (720p, 250 frames) for watermarked content (red) and not watermarked content (blue)

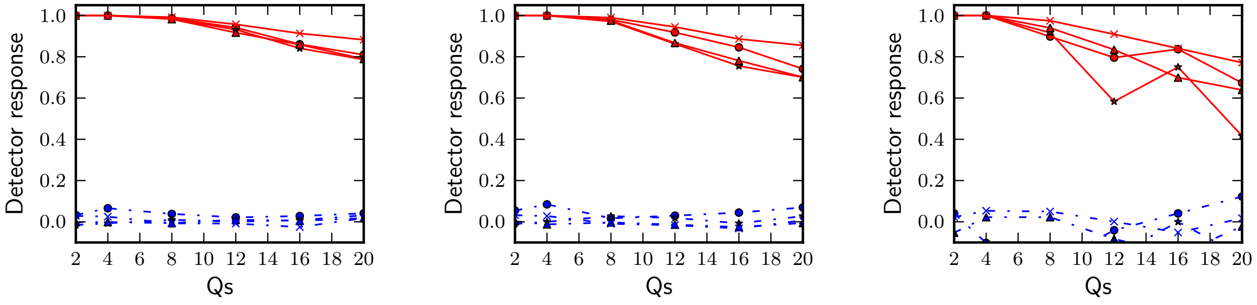


(a) MbDist=100

(b) MbDist=25

(c) MbDist=4

Fig. 10. Probability of a false positive for varying Qp of x264 (720p, 250 frames) for watermarked content (red) and not watermarked content (blue)

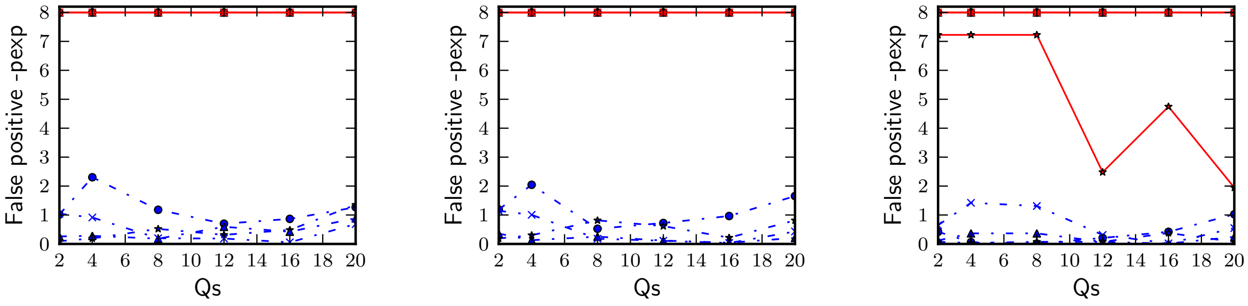


(a) MbDist=100

(b) MbDist=25

(c) MbDist=4

Fig. 11. Detector response for varying Qs for H.263 (ffmpeg, mpeg4, 720p, 250 frames) for watermarked content (red) and not watermarked content (blue)

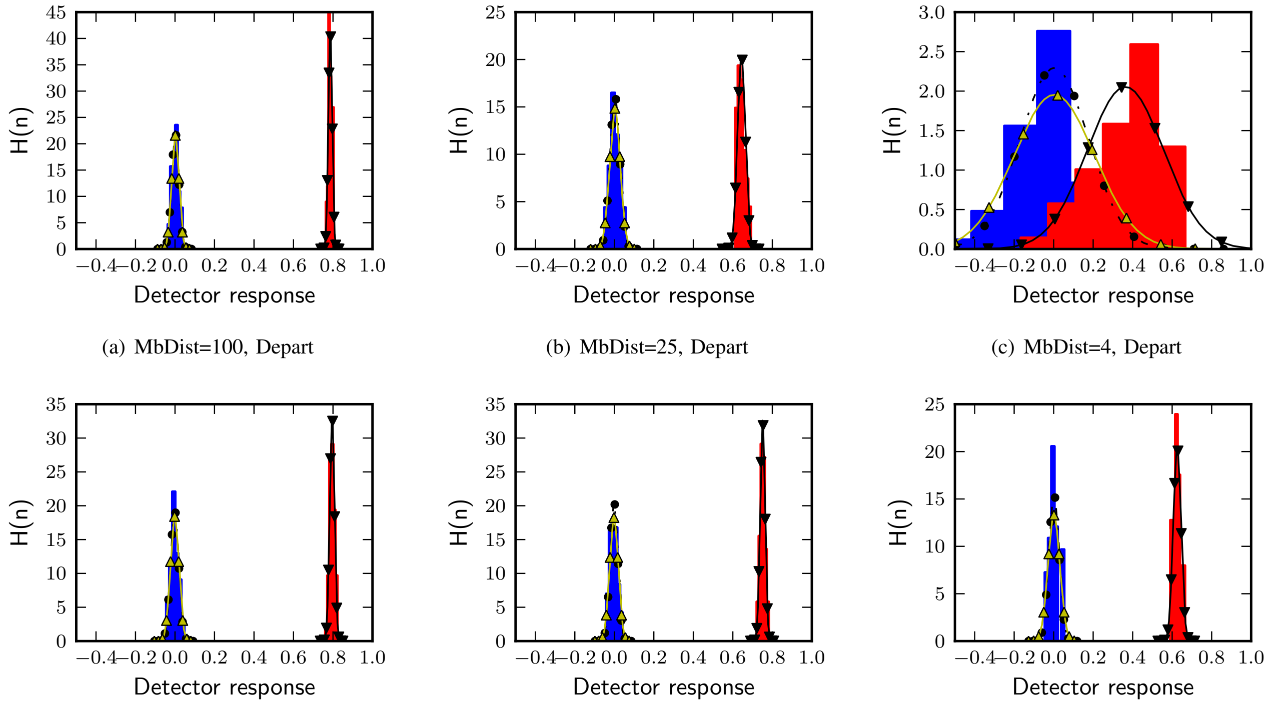


(a) MbDist=100

(b) MbDist=25

(c) MbDist=4

Fig. 12. Probability of a false positive for varying Qs of H.263 (ffmpeg, mpeg4, 720p, 250 frames) for watermarked content (red) and not watermarked content (blue)



(a) MbDist=100, Depart

(b) MbDist=25, Depart

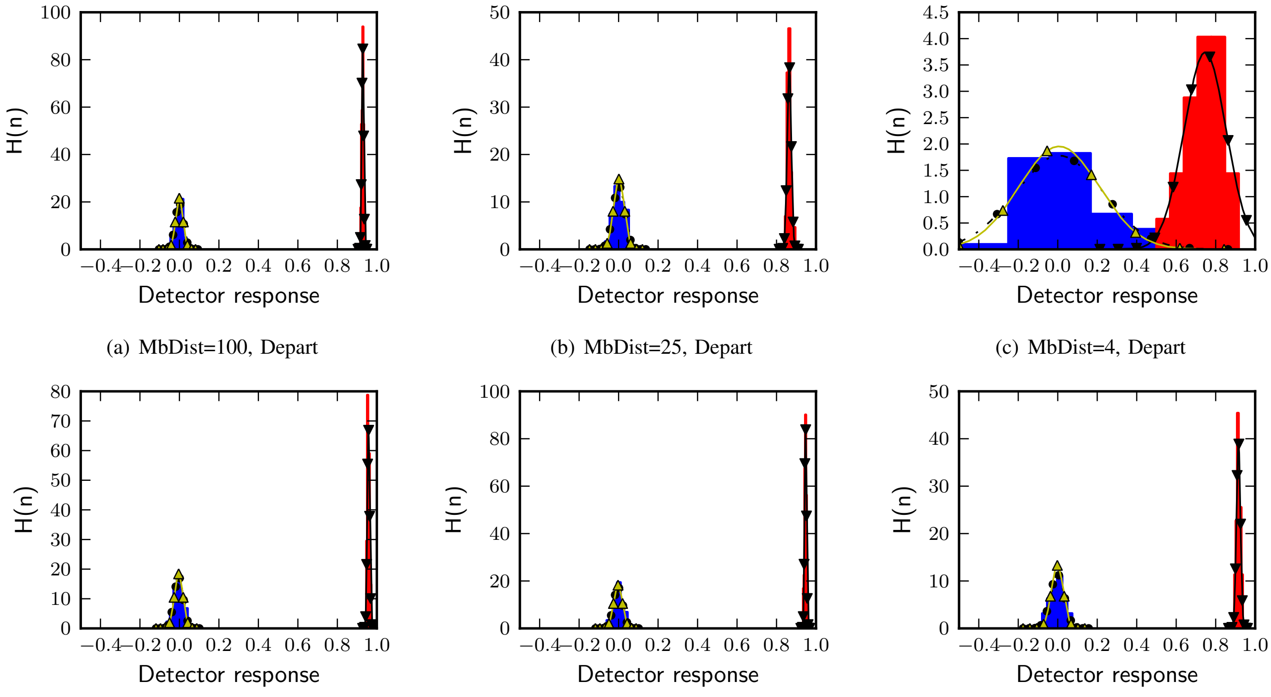
(c) MbDist=4, Depart

(d) MbDist=100, Elephant

(e) MbDist=25, Elephant

(f) MbDist=4, Elephant

Fig. 13. H.264 (x264, ultrafast, Qp 36): Histogram of 100 detector responses



(a) MbDist=100, Depart

(b) MbDist=25, Depart

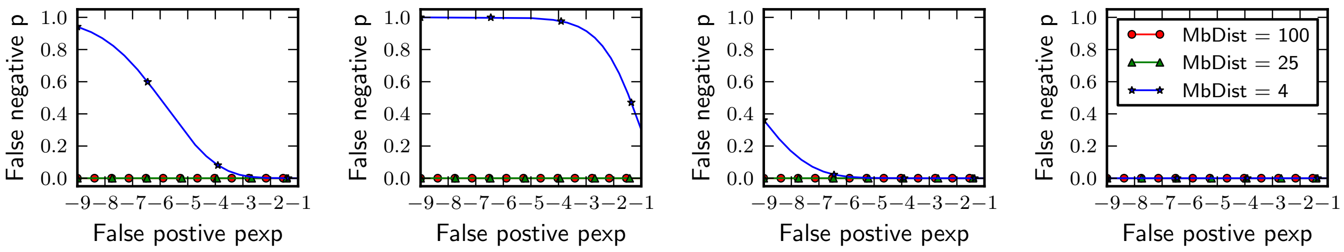
(c) MbDist=4, Depart

(d) MbDist=100, Elephant

(e) MbDist=25, Elephant

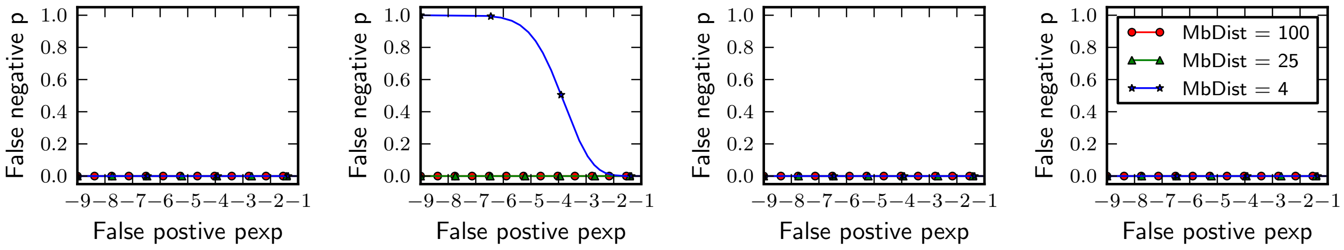
(f) MbDist=4, Elephant

Fig. 14. H.263 (ffmpeg, mpeg4, Qs=12): Histogram of 100 detector responses



(a) Canal (b) Depart (c) Ebu (d) Elephant

Fig. 15. ROC at H.264 compression (x264, ultrafast, Qp 36)



(a) Canal (b) Depart (c) Ebu (d) Elephant

Fig. 16. ROC at MPEG-4 compression (ffmpeg, mpeg4, qs 12)

TABLE IV
Checkmark attacks: detector responses

Dist. / Seq.	Canal	Depart	Ebu	Elephant
gaussian1	0.0766	0.095	0.160	0.125
gaussian2	0.0766	0.095	0.159	0.124
medfilt1	0.0961	0.093	0.150	0.117
medfilt2	0.0792	0.101	0.161	0.124
medfilt3	0.1064	0.100	0.155	0.112
sharpen1	0.0870	0.072	0.159	0.144
wiener1	0.0792	0.093	0.158	0.125
wiener2	0.0688	0.092	0.148	0.123
dpr1	0.0805	0.092	0.152	0.119
dpr2	0.0844	0.104	0.146	0.109

TABLE V
Checkmark attacks: false positive probability

Dist. / Seq.	Canal	Depart	Ebu	Elephant
gaussian1	0.00120	1.08e-7	5.61e-25	2.34e-9
gaussian2	0.00120	1.08e-7	1.07e-24	3.04e-9
medfilt1	0.00007	1.72e-8	3.15e-22	2.91e-8
medfilt2	0.00101	2.12e-8	2.91e-25	3.04e-9
medfilt3	0.00001	2.12e-8	1.01e-23	9.56e-8
sharpen1	0.00028	4.27e-5	1.07e-24	8.77e-12
wiener1	0.00101	1.95e-7	1.48e-24	2.34e-9
wiener2	0.00371	2.36e-7	1.06e-21	6.57e-9
dpr1	0.00071	2.36e-7	1.25e-22	1.39e-8
dpr2	0.00050	5.92e-9	2.63e-21	1.51e-7

heavily reduced quality due to blocking artifacts. Figure 17 shows the result of the attack on the Elephant sequence.

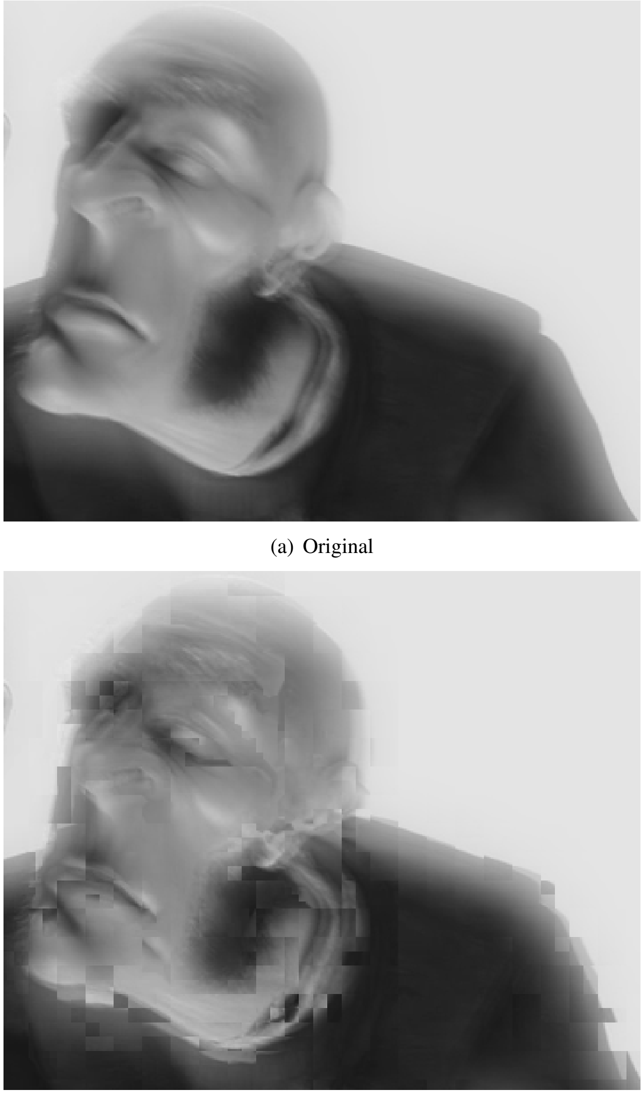
D. Comparison to Previous Work

The comparison to previous work focuses on the algorithms by Zou and Bloom [8], [4], [5], which are the only proposals that satisfy similar requirements, i.e., offer structure-preservation and substitution embedding for H.264. Neither the implementation nor the test data (a clip from the action movie ”Independence Day”) of these approaches could be made accessible by the authors. Therefore we are not able to perform a rigorous experimental comparison with exactly the same settings for all algorithms. However, we provide experimental results for our approach for four very different video sequences, which helps to put the results into perspective and enable a fair comparison. All the test data of our evaluation are publicly accessible (ftp://ftp.ivc.polytech.univ-nantes.fr/IRCCyN_IVC_H264_Watermarking_Structure_Preserving/) and thus future proposals can rigorously compare their approaches to our proposal.

Compared to Zou and Bloom’s approaches [8], [4], [5] our approach is superior in terms of the number of watermarkable blocks (see table VI). Thus fewer frames are needed to embed a watermark. Alternatively the robustness of our approach can be improved (a lower threshold can be selected for the same probability of alarm).

The comparison of the robustness on bit embedding level reveals that the performance is rather similar, while our robustness criterion (the average luminance feature difference must be larger than 0.25) leads to higher correlations for downsizing, the effect of downsizing and compression are almost the same (see table VII). A special case again is the computer generated Elephant sequence, which can be compressed very efficiently and thus the highest correlations are observed with this sequence.

The number of embedded bits has a tremendous effect on the threshold selection (for a fixed false positive probability) or the false positive probability (for a fixed threshold). Figure 18 plots the thresholds for different probabilities of false alarms and for the different approaches against the number of frames. As distortions affect the different approaches almost



(a) Original

(b) Attack

Fig. 17. Visual examples of an attack based on a bitstream analysis

TABLE VI
WATERMARKABLE BLOCKS PER FRAME OF 1080P VIDEO, THE RESULTS OF ZOU'S CAVLC [4] AND ZOU'S CABAC [5] COMPARED TO OUR APPROACH FOR DIFFERENT SEQUENCES

Zou & Bloom		Our approach			
CALVC	CABAC	Canal	Depart	Ebu	Elephant
0.625	1.319	17.672	41.196	57.420	18.820

similarly this allows a fair comparison of the performance of the approaches. The lower the threshold the more robust is the watermarking algorithm against distortions. Our algorithm requires significantly lower detection thresholds compared to the algorithms of Zou [4], [5].

VII. CONCLUSION

The proposed H.264/CAVLC watermarking algorithm enables structure-preserving H.264 watermarking. Due to the separation of embedding in an analysis and and a substitution stage, it is able to efficiently generate numerous video files with different watermarks embedded. Compared to previous work it offers a significantly increased marking space and improved robustness. While the analysis stage of the algorithm is lightweight compared to previous proposals, it still satisfies the invisibility constraints, which has been shown by subjective experiments. The algorithm offers high robustness to re-compression and sufficient robustness against standard watermarking attacks.

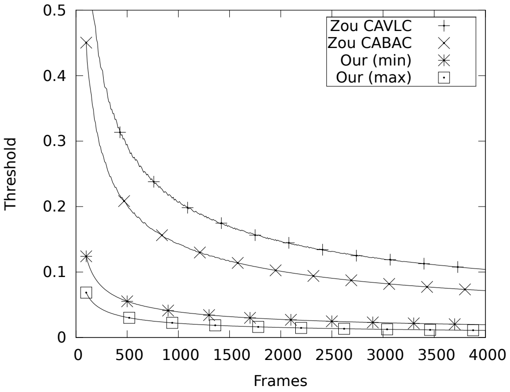


Fig. 18. The thresholds for the approaches of Zou and to our algorithm plotted as a function of frames for a false positive probability of 10^{-7}

REFERENCES

[1] T. Y. Chung, M. S. Hong, Y. N. Oh, D. H. Shin, and S. H. Park, "Digital watermarking for copyright protection of MPEG2 compressed video," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 3, pp. 895–901, 1998.

[2] Gerrit C. Langelaar and Reginald L. Lagendijk, "Optimal differential energy watermarking of dct encoded images and video," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 148–158, Jan. 2001.

[3] Bijan G. Mobasseri, "Watermarking of MPEG-2 video in compressed domain using VLC mapping," in *International Multimedia Conference, Proceedings of the 7th Workshop on Multimedia and Security, MM-SEC '05*, New York, NY, USA, Aug. 2005, pp. 91–94, ACM.

[4] Dekun Zou and Jeffrey Bloom, "H.264/AVC substitution watermarking: a CAVLC example," in *Proceedings of the SPIE, Media Forensics and Security*, Jan Jose, CA, USA, Jan. 2009, vol. 7254, SPIE.

[5] D. Zou and J. Bloom, "H.264 stream replacement watermarking with CABAC encoding," in *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME '10*, Singapore, July 2010.

[6] M. Noorkami and R. M. Mersereau, "A framework for robust watermarking of H.264 encoded video with controllable detection performance," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 1, pp. 14–23, Mar. 2007.

[7] M. Noorkami and R. M. Mersereau, "Compressed-domain video watermarking for H.264," in *Proceedings of the IEEE International Conference on Image Processing, ICIP '05*, Genova, Italy, Sept. 2005, pp. 890–893, IEEE.

[8] Dekun Zou and Jeffrey A. Bloom, "H.264/AVC stream replacement technique for video watermarking," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '08*, Las Vegas, NV, USA, Mar. 2008, pp. 1749–1752, IEEE.

[9] I. E. G. Richardson, *H.264 and MPEG-4 video compression: video coding for next generation multimedia*, Wiley & Sons, 2003.

[10] ITU-T H.264, "Advanced video coding for generic audiovisual services," Nov. 2007.

[11] L. C. Jesty, "The relation between picture size, viewing distance and picture quality," *Proc. Inst. Electr. Eng. Part B*, vol. 105, pp. 425–439, 1958.

[12] S. H. Baker and M. E. Carpenter, "Correlation of spot characteristics with perceived image quality," *IEEE Trans. Commun. Electron.*, vol. 35, pp. 319–324, 1989.

TABLE VII
AVERAGE CORRELATION FOR ZOU'S APPROACH [4] AND OUR ALGORITHM

Attack	Zou CAVLC (Full HD)	Our (Canal Full HD)	Our (Depart)	Our (Ebu)	Our (Elephant)
No Attack	0.9685	1.0000	1.0000	1.0000	1.0000
Downsize to 960x540	0.8133	0.9995	0.9998	0.9988	1.0000
Downsize to 480x270	0.2778	0.9488	0.9586	0.9669	0.9974
Downsize to CIF	0.1804	0.8497	0.8572	0.8942	0.9795
Downsize to CIF, 1M	0.1148	0.2643	0.2403	0.2108	0.8826
Downsize to CIF, 780K	0.1181	0.2489	0.2088	0.1710	0.8826
Downsize to CIF, 300K	0.1026	0.0878	0.1043	0.0882	0.6272

[13] Peter G. J. Barten, "Evaluation of subjective image quality with the square-root integral method," *J. Opt. Soc. Am. A*, vol. 7, pp. 2024–2031, 1990.

[14] ITU P.910, "Subjective video quality assessment methods for multimedia applications," Tech. Rep., Intl Telecom. Union, April 2008, SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS Audiovisual quality in multimedia services.

[15] ITU-R-BT.500-11, "Methodology for the subjective assessment of the quality of television pictures question itu-r 211/11, g," Tech. Rep., Intl Telecom. Union, 2004.

[16] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, May 2006.

[17] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, 2007.

[18] V. Pankajakshan and F. Autrusseau, "A multi-purpose objective quality metric for image watermarking," in *IEEE International Conference on Image Processing, ICIP'2010*, 2010, pp. 2589–2592.

[19] M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication*, vol. 23(4), pp. 239–256, 2008.

[20] VQEG MM, "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I," Tech. Rep., VQEG, 2008.

[21] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, Morgan Kaufmann, 2007.

[22] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[23] Hui Cheng and Michael A Isnardi, "Spatial temporal and histogram video registration for digital watermark detection," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'03)*, Sept. 2003, vol. II, pp. 735–738.

[24] Bertrand Chupeau, Lionel Oisel, and Pierrick Jouet, "Temporal video registration for watermark detection," in *Proceedings of the 2006 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, Apr. 2006, vol. II, p. 4.

[25] J. J. Eggers and B. Girod, *Informed Watermarking*, Kluwer Academic Publishers, 2002.

[26] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[27] Maurizio Carosi, Vinod Pankajakshan, and Florent Autrusseau, "Toward a simplified perceptual quality metric for watermarking applications," in *Proceedings of the SPIE conference on Electronic Imaging*, 2010, vol. 7542.

[28] Shelby Pereira, Sviatoslav Voloshynovskiy, M. Madueno, and Thierry Pun, "Second generation benchmarking and application oriented evaluation," in *Proceedings of the 4th Information Hiding Workshop '01*, Portland, OR, USA, Apr. 2001, vol. 2137 of *Lecture Notes in Computer Science*, pp. 340–353, Springer.