A Moving Object Detection Scheme in Codestream Domain for Motion JPEG Encoded Movies

Masaaki Fujiyoshi, Yuji Tachizaki*, and Hitoshi Kiya

Dept. Inform. & Commun. Systems Eng., Tokyo Metropolitan University,
6–6 Asahigaoka, Hino-shi, Tokyo 191–0065, Japan
mfujiyoshi@ieee.org, tachizaki-yuji@sd.tmu.ac.jp, kiya@sd.tmu.ac.jp

Abstract. This paper proposes a scheme for detecting moving objects (MOs) from a Motion JPEG (MJ) coded video recorded by a stationary camera. The proposed scheme detects MOs without decoding a compressed video, whereas the ordinary motion detecting schemes have to decompress the video. For MOs detection, the correlation based on the positive and negative sign of discrete cosine transformed (DCT) coefficients of video frames is used in the proposed scheme. A DCT sign is encoded separately from its corresponding magnitude, so the signs are directly extracted from a MJ compressed codestream, that is, no need to decompress the coded video. In the proposed scheme, MOs are detected in each 8×8-sized block of two adjacent video frames where the block is the MJ compression unit. Experimental results show that an usage of surrounding blocks in a detection decreases both false positive and false negative detections.

Keywords: Motion detection, Video surveillance, DCT-SPC, Bitstream.

1 Introduction

Video surveillance systems are widely used from security applications such as intrusion detection [1] to non-secure issues like human behavioral analysis [2]. The fundamental and important component in these systems is moving objects (MOs) extraction from a video, and many MOs detection schemes have been proposed [3–7].

Ordinary schemes uses the spatial information such as color or intensity [3,4], but these schemes have to fully decode the compressed codestream before MOs detection in the systems which videos are compressed by a video encoder [5]. To overcome this inconvenience, MOs detection schemes in compressed domain have been proposed [6,7], but coded videos have to be decompressed to the compressed domain.

This paper proposes a MOs detection scheme in the codestream domain for surveillance systems with a stationary camera, in particular, IP-based remote

* The author is currently with Chubu Electric Power Co., Inc., Japan.

G. Qiu et al. (Eds.): PCM 2010, Part II, LNCS 6298, pp. 595–606, 2010.

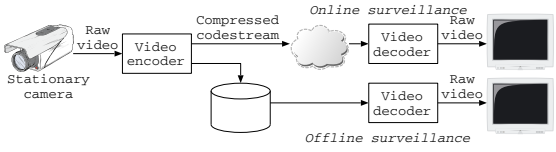


Fig. 1. The assumed video surveillance system (a cloud represents a network)

surveillance systems. The positive and negative sign of discrete cosine transformed (DCT) coefficients of MJ coded video frames are leveraged in the proposed scheme which DCT signs are obtained directly from a codestream without decoding it. The proposed scheme detects MOs by a similarity based on DCT sign phase correlation (DCT-SPC) between two adjacent frames.

2 Preliminary

The goal of the proposed MOs detection scheme is described in this section. MJ that is assumed as the video encoding technology in surveillance systems is then explained. DCT-SPC that the proposed scheme bases on is also given.

2.1 Goal of the Proposed MOs Detection

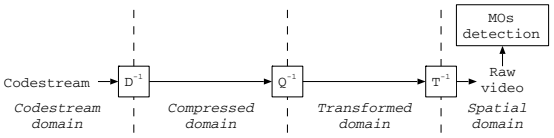
Figure 1 shows a block diagram of the assumed video surveillance system. In the system, a stationary video camera is connected to an encoder to compress video for efficient transmission and/or storage. For remote surveillance applications, the compressed video is transmitted to the observer side. The compressed video is stored for ex-post inspection.

Ordinary MOs detection schemes generally run in the spatial domain [3,4]. Thus, in the system mentioned above, these schemes have to fully decode the compressed video before MOs detection as shown in Fig. 2 (a). Either schemes in compressed domain using motion vectors, quantized coefficients, or macro block sizes [6,7] have to decode the codestream (Fig. 2 (b)). This decoding introduces extra computational costs and time consumption to MOs detection.

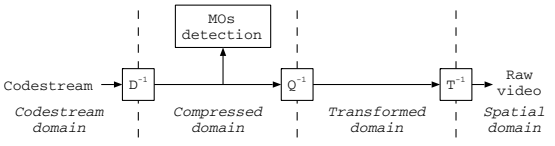
The proposed scheme detects MOs from the compressed video without decompressing the video, i.e., in the codestream domain (Fig. 2 (c)). The assumptions and requirements of the proposed scheme are summarized below.

1. System
   (a) A stationary camera takes videos.
   (b) A video is encoded by MJ.
2. MOs detection
   (a) MOs are directly detected from a codestream.
   (b) The background frame is never used.
   (c) No learning process is required.

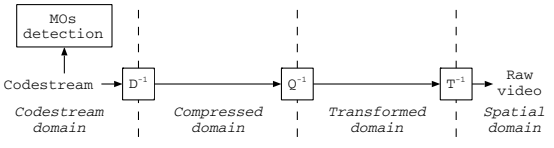
The proposed scheme utilizes the positive and negative sign of DCT coefficients to satisfy the requirements, and have to neither memorize nor update the background frame.



(a) Conventional schemes in the spatial domain [3,4].



(b) Conventional schemes in the compressed domain [6,7].



(c) Proposed scheme.

Fig. 2. MOs detection in different domains (D^{-1}: entropy decoding, Q^{-1}: inverse quantization, and T^{-1}: inverse transformation)

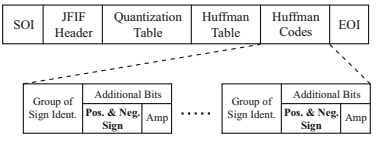


Fig. 3. An example of JPEG codestream

2.2 Motion JPEG

In this paper, the video encoding technology used in surveillance systems is assumed to be MJ as it is still widely used, in particular, in IP-based remote surveillance systems. Since MJ encodes each frame of a video by JPEG [11], JPEG encoding algorithm is briefly described here.

In JPEG, pixel values in an original image are shifted and the image is divided into 8×8-sized non-overlapping blocks, referred to as DCT blocks in this paper. The two dimensional DCT, then, is applied to each DCT block to produce one DC and 63 AC coefficients. All the coefficients are quantized according to a table scaled by Q-factor. Finally, the quantized coefficients are coded by an entropy coder with a Huffman table. The AC coefficients are directly coded, whereas the difference between two consecutive DCT blocks is coded for DC coefficients.

Figure 3 shows the structure of a JPEG codestream that is generated from a grayscale image and with Huffman encoder. The start of image (SOI) marker is the head of a JPEG codestream. The JPEG File Interchange Format (JFIF) header contains information such as the image size. The next two entities are the quantization table for scalar quantization and the Huffman table for entropy encoding. Then, the entropy-coded DCT coefficients are put. The end of image (EOI) marker follows the last byte of a codestream.

An entropy code consists of a Huffman code and an additional bits. The Huffman code represents the amplitude, whereas the positive and negative sign is represented by the most significant bit of the additional bits. Since this positive and negative sign is independent of other bits in a codestream, the signs are directly acquired from JPEG and MJ codestreams.

2.3 DCT-SPC

DCT-SPC is a correlation that is used to estimate the shift values and the similarity between two signals [8,9]. DCT-SPC uses only the sign of DCT coefficients of signals, similar to that phase correlation or phase only correlation uses only the phase of discrete Fourier transformed coefficients of signals [10], in which DCT-SPC and phase correlation have a theoretically strong relation [8,9].

Let $G(k)$ be the N-point DCT coefficients of N-point signal $g(n)$. The DCT sign is defined in terms of $G(k)$ and its corresponding absolute value, $|G(k)|$, as

$$s_G(k) = \begin{cases} \dfrac{G(k)}{|G(k)|}, & |G(k)| \neq 0 \\ 0, & |G(k)| = 0 \end{cases}. \tag{1}$$

DCT sign product, $C_s(k)$, between $g_1(n)$ and $g_2(n)$ is given as

$$C_s(k) = s_{G_1}(k) s_{G_2}(k). \tag{2}$$

Then, DCT-SPC, $c_s(n)$, is defined by using $C_s(k)$ as

$$c_s(n) = \frac{1}{N} \sum_{k=0}^{N-1} K_k C_s(k) \cos\left(\frac{\pi n k}{N}\right), \tag{3}$$

where

$$K_k = \begin{cases} \dfrac{1}{2}, & k = 0 \\ 1, & k \neq 0 \end{cases}. \tag{4}$$

DCT-SPC $c_s(n)$ has a spike peak for similar signals as well as phase correlation has.

The next section proposes an MOs detection scheme based on DCT-SPC for MJ encoded surveillance videos. The proposed scheme utilizes DCT signs of frames for MOs detection in codestream domain.

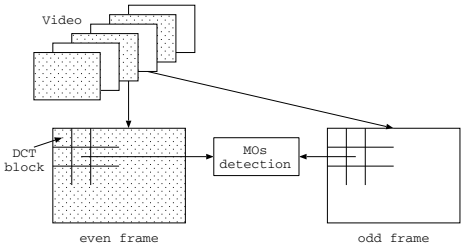


Fig. 4. The proposed scheme detects MOs between two adjacent frames per DCT block

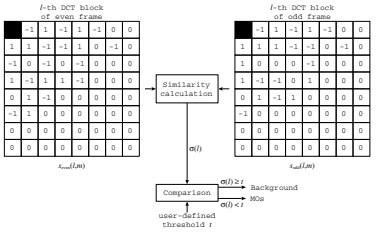


Fig. 5. The proposed scheme detects MOs using similarity $\sigma(l)$ based on DCT-SPC [8, 9]. Boxes with 1, 0, and −1 represent DCT sign for AC coefficients, and filled boxes represent DC coefficients that are not used to compute $\sigma(l)$.

3 Proposed Scheme

This section proposes a scheme to detect MOs from MJ compressed videos in codestream domain for video surveillance systems with a stationary camera. The algorithm of the proposed scheme and the essences enabling MOs detection in codestream domain are described in subsequent sections. At the last of this section, detection using multiple DCT blocks for more robust detection is described.

3.1 Algorithm

The proposed scheme detects MOs in each DCT block based on DCT-SPC between two adjacent frames. Here, MOs detection between a set of even and odd frames is considered, as shown in Fig. 4. The following steps are applied to each set of even and odd frames for MOs detection (Fig. 5).

1. $l := 0$.
2. From the part of the codestream corresponding to the even frame, DCT sign bits for all 63 AC coefficients in the l-th DCT block of luminance component are extracted. DCT signs are represented by $s_{\text{even}}(l, m)$, where $m = 1, 2, \ldots, 63$ and $m = 0$ represents the DC coefficient.
3. DCT signs for the subsequent odd frames, $s_{\text{odd}}(l, m)$ of luminance component, are obtained from the codestream.

4. For the l-th DCT block in the even and odd frames, it computes the following measure,

$$\sigma(l) = \frac{\sum_{m=1}^{63} s_{\mathrm{even}}(l,m)s_{\mathrm{odd}}(l,m)}{\sum_{m=1}^{63} |s_{\mathrm{even}}(l,m)s_{\mathrm{odd}}(l,m)|}. \tag{5}$$

5. If $\sigma(l) < t$, where t represents the user-defined threshold, it is determined that motion exists between the l-th block of the even and odd frames.
6. $l := l + 1$. Continue to Step 2 unless $l = L$, where L represents the number of DCT blocks in a frame.

By applying this algorithm to a video sequence, MOs are detected two by two frames.

3.2 MOs Detection in Codestream Domain

The proposed scheme focuses attention on the positive and negative sign of AC coefficients so that it detects MOs directly from the compressed video. DCT signs of AC coefficients make a DCT-SPC-based MOs detection available and are directly extractable from the MJ compressed codestream.

MOs Detection Based on DCT-SPC. In surveillance systems with a stationary camera, the background of the scene without MOs is not changed through frames of a video sequence. Therefore, a MOs detection using frames of a video sequence in the systems results in finding the difference between a target and the reference frames, where an even and the successive odd frames are the target and reference frames in the proposed scheme.

To detect the difference arising from MOs' existence, the proposed scheme utilizes a similarity between frames where the similarity given by Eq. (5) is a special form of DCT-SPC [8]. Similarity $\sigma(l)$ between consecutive two frames gives a large value for DCT blocks in which no MO exists, whereas $\sigma(l)$ is small in a DCT block where MOs are recorded. The proposed scheme detects MOs simply by comparing $\sigma(l)$ with user-defined threshold t as described in Step 5.

It is noted that the proposed scheme does not use DCT signs of DC coefficients, because neglecting DC coefficients does not affect $\sigma(l)$ and DC coefficients sometime fluctuate according to the lighting condition of the scene that introduces wrong detection. Moreover, $\sigma(l)$ is adaptively weighted by the number of AC coefficients which are non-zero in both frames. This adaptive weighting prevents the proposed scheme from detecting the background as MOs.

DCT Signs Extractable from MJ Codestreams. In Steps 2 and 3 of the algorithm in the proposed scheme, DCT signs for AC coefficients are extracted from the MJ encoded codestream without decompressing the video sequence. As mentioned in Sect. 2.2, the DCT sign of an AC coefficient is represented as the

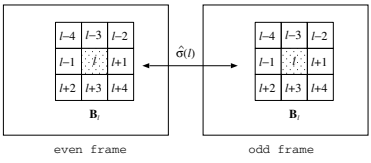


Fig. 6. An example of multiple DCT blocks for robust MOs detection. $\mathbf{B}_l$ consists of 3×3 DCT blocks

most significant bit of additional bits in an entropy code. Thus, the proposed scheme directly extracts DCT signs from the MJ compressed codestream without decoding it.

As mentioned in Sect. 2.2, for DC coefficients, the difference between two consecutive DCT blocks is encoded, so DCT signs of DC coefficients cannot be obtained directly from the codestream. The proposed scheme, however, does not use the DCT sign of DC coefficients for computing similarity $\sigma(l)$, as described in the previous section. So similarity given by Eq. (5) makes contribution to not only MOs detection but also DCT signs extraction from codestreams without decompressing video sequences.

3.3 More Robust Detection

The algorithm proposed in Sect. 3.1 decides whether a DCT block contains MOs based on the DCT-SPC based similarity that is calculated by using the focused DCT block itself. This section proposes another detection algorithm using multiple DCT blocks for more robust MOs detection, because a MO often cover several DCT blocks.

The extended similarity between the l-th block of the even and odd frames is defined as

$$\hat{\sigma}(l) = \frac{\displaystyle\sum_{b_l \in \mathbf{B}_l} \sum_{m=1}^{63} s_{\mathrm{even}}(b_l, m) s_{\mathrm{odd}}(b_l, m)}{\displaystyle\sum_{b_l \in \mathbf{B}_l} \sum_{m=1}^{63} |s_{\mathrm{even}}(b_l, m) s_{\mathrm{odd}}(b_l, m)|}, \tag{6}$$

where $\mathbf{B}_l$ is a set of DCT blocks centering on the l-th DCT block, and b_l represents the location of a DCT block in $\mathbf{B}_l$. For 3×3 blocks shown in Fig. 6, $\mathbf{b}_l$ is represented as

$$\mathbf{B}_l = \{l-4, l-3, l-2, l-1, l, l+1, l+2, l+3, l+4\}. \tag{7}$$

If $\mathbf{B}_l = \{l\}$, Eq. (6) comes down to Eq. (5). Then, if $\hat{\sigma}(l) < t$ as in Step 5 of the algorithm proposed in Sect. 3.1, it is decided that MOs exist in the l-th DCT block.

(a) ViSOR (45th frame). (b) Lab (167th frame). (c) HallMonitor (30th frame). (d) PETS2001 (360th frame).

(e) Ground truth for (a). (f) Ground truth for (b). (g) Ground truth for (c). (h) Ground truth for (d).

Fig. 7. Frames from video sequences. White regions are ground truth given in pixels

Table 1. Video sequences for evaluation

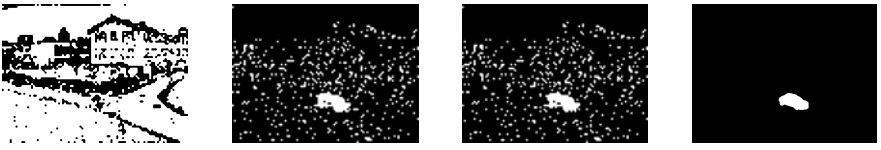
ViSOR[1]	320×240 pixels, 15 fps, 1–120 frames are used
Lab[2]	320×240 pixels, 15 fps, 108–227 frames are used
HallMonitor [13]	352×288 pixels, 30 fps, 1–120 frames are used
PETS2001[3]	768×576 pixels, 15 fps, 301–420 frames are used

4 Experimental Results

Three standard and one recorded video sequences are used for evaluation. A frame from videos and the corresponding ground truth images are shown in Fig. 7, and the detail of sequences are summarized in Table 1. Excepting for standard video distributed with the ground truth, the ground truth images are given by the authors. All video sequences are compressed by MJ with the condition that Q-factor is set to 50.

Detection accuracy is evaluated by receiver operator characteristics (ROC) curve which represents the relation between the false negative rate (FNR) and the false positive rate (FPR) for various threshold of the system. The system whose ROC curve is the closest to the origin serves the best performance. Equal error rate (EER), i.e., FNR = FPR, is also used for quantitative evaluation. Threshold t's are between 0.1 and 1 in 0.01 steps. For severe evaluation, the FNR and FPR are computed in pixels, though MOs are detection in DCT blocks.

[1] "Cam4" of Indoor Domotic Unimore D.I.I. setup in ViSor [12].
[2] Video recorded at the authors' lab. A person enters the room from left. He walks into the back, and then he returns to the front. Finally, he goes out to the right side.
[3] "Camera 2" of Testing frames in Dataset 2 for PETS2001 [14].



(a) Fixed weight. (b) DC and AC. (c) Proposed. (d) Ground truth.

Fig. 8. MOs detection examples. White regions represent MOs detected in DCT blocks. The similarities are derived with the only focused DCT block (PETS2001).

Table 2. Detection accuracy for the proposed scheme with Eq. (5). The distance to the origin is not represented in percentage.

Sequence	Threshold t	Distance
ViSOR	0.82	0.36
Lab	0.91	0.20
HallMonitor	0.92	0.46
PETS2001	0.88	0.13

4.1 MOs Detection Using the DCT-SPC-Based Similarity

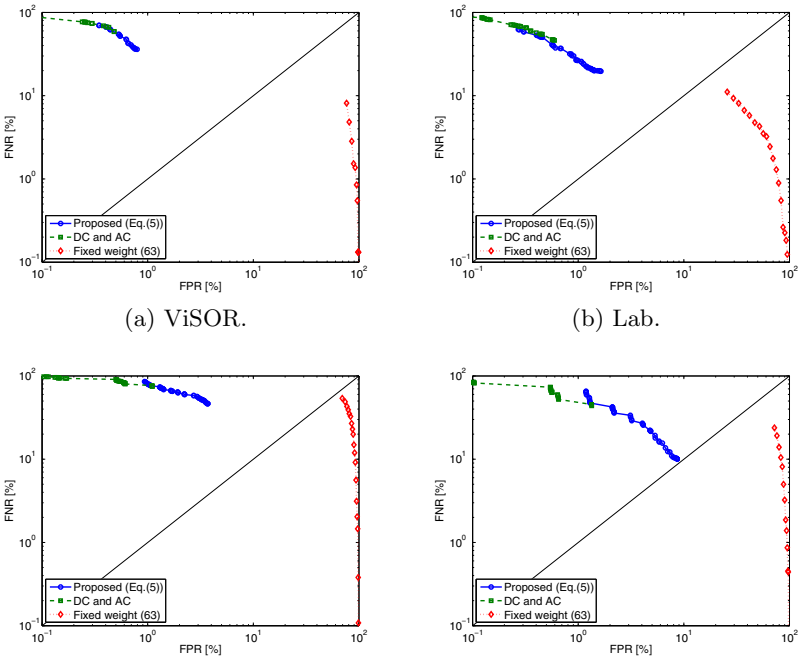
The detection accuracy using Eq. (5) was evaluated. Results with other similarities, i.e., the similarity weighted by the stationary value (the number of AC coefficients, 63) and that using DCT signs of DC and AC coefficients are compared with the proposed scheme. It is noted that similarities are calculated with the focused DCT block only in this section.

Figure 8 shows detection examples for "PETS2001," where results are generated using the best threshold from the perspective of the ROC characteristics. It is confirmed that MOs are detected by using the DCT-SPC-based similarity. The similarity with the fixed weight does not distinguish MOs from the stationary background. Using the DCT sign of DC coefficients in the similarity calculation gives almost no difference to detection accuracy.

Figure 9 shows the ROC curves for the three similarities. It is confirmed that the similarity with the fixed weight gives high FPR, because it takes the stationary background for MOs as shown in Fig. 8. The detection using DC and AC coefficients gives higher FNR than the similarity defined by Eq. (5) does. Consequently, the similarity given by Eq. (5) is superior to other two similarities. Table 2 summarizes the performance of the proposed scheme with Eq. (5).

4.2 Robust Detection Using Multiple DCT Blocks

Following the results given in the previous section that the similarity given by Eq. (5) is the best, detection using the extended similarity given as Eq. (6) is investigated here.



(a) ViSOR. (b) Lab.

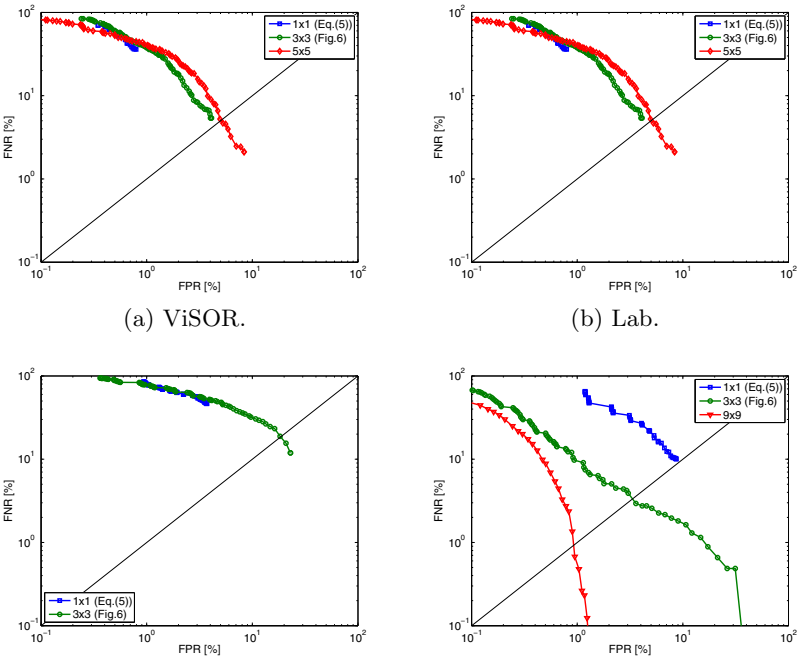
(c) HallMonitor. (d) PETS2001.

Fig. 9. ROC curves for the similarities using only the focused DCT block

Figure 10 shows ROC curves using multiple DCT blocks for MOs detection. Seven different sized square $\mathbf{B}_l$ are employed for evaluation; 1×1-sized (Eq. (5)), 3×3 (Fig. 6), 5×5, 7×7, 9×9, 11×11, and 13×13. For its simplicity, Fig. 10 only shows results for 1×1, 3×3, and the size serving the best performance in terms of ROC curve characteristics. Table 3 summarizes the detection accuracy in terms of the EER. Detection examples are shown in Fig. 11.

From Figs. 10 and 11, it is confirmed that the similarity using multiple DCT blocks improves detection accuracy. It is noted that the best size of the region used to derive $\hat{\sigma}(l)$ varies dependently on not only the frame resolution but also the ratio between the frame resolution and the size of MOs in a frame. For "PETS2001" in which each frame has 768×576 pixels, even a small car in the frame covers many DCT blocks. This results in that $\mathbf{B}_l$ consisting of 9×9 of DCT blocks serves the best detection accuracy. Though other sequences have the similar frame resolution, the MO's size in "HallMonitor" is smaller than those of other videos. Thus, 3×3-sized $\mathbf{B}_l$ gives the best performance.

Through the experimental results in this section, it is confirmed the proposed scheme detects MOs based on DCT-SPC and improves the detection accuracy by using surrounding DCT blocks as well as the focused DCT block in the similarity derivation.



(a) ViSOR. (b) Lab.

(c) HallMonitor. (d) PETS2001.

Fig. 10. ROC curves for the similarities using multiple DCT block, Eq. (6). '1x1' is the result using Eq. (5) and '3x3' uses blocks shown in Fig. 6. Other legends indicate the blocks forming $\mathbf{B}_l$, the region for deriving $\hat{\sigma}(l)$. $\mathbf{B}_l$ consisting of '5×5' DCT blocks serves the best performance for "ViSOR" and "Lab," and '3×3' and '9×9' are the best for "HallMonitor" and "PETS2001," respectively.

Table 3. The best EER served by the proposed scheme with Eq. (6)

Sequence	$\mathbf{B}_l$	Threshold t	EER [%]
ViSOR	5×5	0.92	5.0
Lab	5×5	0.92	4.5
HallMonitor	3×3	0.97	18.5
PETS2001	9×9	0.90	0.7

(a) One DCT block (Eq. (5)). (b) 9×9 DCT blocks. (c) Ground truth.

Fig. 11. The similarity computation using multiple DCT blocks serves more accurate MOs detection than that using the focused DCT block does (PETS2001)

5 Conclusions

This paper has proposed a MOs detection scheme in codestream domain based on DCT-SPC for video surveillance systems with a stationary camera. The proposed scheme detects MOs from a MJ compressed codestream without decompressing it, because the scheme focuses DCT signs that are directly extracted from the codestream. The DCT-SPC-based similarity that uses not only the focused DCT block but also surrounding DCT blocks improves detection accuracy.

Further works include an adaptation to videos with cluttered background and the analysis among the frame resolution, MOs' size, and the number of DCT blocks used in the similarity calculation.

References

1. Makarov, A.: Comparison of background extraction based intrusion detection algorithms. In: Proc. IEEE Int. Conf. Image Process., pp. 521–524 (1996)
2. Leykin, A., Tuceryan, M.: A vision system for automated customer tracking for marketing analysis: low level feature extraction. In: Proc. Human Activity Recognition and Modelling Workshop (2005)
3. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: principle and practice of background maintenance. In: Proc. IEEE Int. Conf. Comput. Vision, pp. 255–261 (1999)
4. Piccardi, M.: Background subtraction techniques: a review. In: Proc. IEEE Int. Conf. Syst., Man, Cybern., pp. 3099–3104 (2004)
5. Zen, H., Hasegawa, T., Ozawa, S.: Moving object detection from MPEG coded picture. In: Proc. IEEE Int. Conf. Image Process., vol. 4, pp. 25–29 (1999)
6. Nakajima, Y., Yoneyama, A., Yanagihara, H., Sugano, M.: Moving object detection from MPEG coded data. In: Proc. SPIE, vol. 3309, pp. 988–996 (1998)
7. Poppe, C., De Bruyne, S., Paridaens, T., Lambert, P., Van de Walle, R.: Moving object detection in the H.264/AVC compressed domain for video surveillance applications. J. Vis. Commun. Image R 20, 428–437 (2009)
8. Ito, I., Kiya, H.: DCT sign-only correlation with application to image matching and the relationship with phase-only correlation. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process., vol. I, pp. 1237–1240 (2007)
9. Ito, I., Kiya, H.: Multiple-peak model fitting function for DCT sign phase correlation with non-integer shift precision. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process., pp. 449–452 (2009)
10. Kuglin, C.D., Hines, D.C.: The phase correlation image alignment method. In: Proc. IEEE Int. Conf. Cybern. Soc., pp. 163–165 (1975)
11. Information technology — Digital compression and coding of continuous-tone still images — Requirements and guidelines, ISO/IEC Int. Std. 10918-1 (1994)
12. Video Surveillance Online Repository (ViSOR), `http://www.openvisor.org/`
13. CIPR Sequences, Center for Image Processing Research (CIPR) at Rensselaer Polytechnic Institute, `http://www.cipr.rpi.edu/resource/sequences/`
14. Datasets for 2001 IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance (PETS) (2001), `http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001-dataset.html`