2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2009) December 7-9, 2009

MA3-B-5

A DCT Coefficient Sign-Based Background Model for Moving Objects Detection from Motion JPEG Coded Movies

Yuji TACHIZAKI, Masaaki FUJIYOSHI, and Hitoshi KIYA
Dept. of Information and Communication Systems Eng., Tokyo Metropolitan University, Hino, Tokyo 191–0065, Japan
Email: tachizaki-yuji@sd.tmu.ac.jp, mfujiyoshi@ieee.org, kiya@sd.tmu.ac.jp

***Abstract*—This paper proposes a robust moving objects detection method for Motion JPEG coded movies. The proposed method utilizes the similarity based on of the positive and negative sign of discrete cosine transformed (DCT) coefficients. The proposed method describes a time-varying background by an adaptive model consisting of multiple sets of DCT signs with respect to each 8×8 DCT block to adapt to non-stationary scenes. In addition, the multi block-based processing reduces false detection. Moreover, since DCT sign is encoded separately from its corresponding magnitude in a Motion JPEG codestream, the sign can be obtained without decoding a codestream. Experimental results comparing with a conventional method using Gaussian mixture model show the effectiveness of the proposed method.**

I. Introduction

Video surveillance systems seek to automatically identify events of interest in a variety of situations for intrusion detection [1], human behavioral analysis [2], and so on. The extraction of moving objects (MOs) from a movie is a fundamental and important factor in these systems.

The numerous MOs detection methods based on spatial information, such as color or intensity [3], [4], have been proposed. They, however, require decoding a codestream beforehand in systems which compress movies for storing and/or transmitting. Meanwhile, the MOs detection method for Motion JPEG movies has been proposed [5]. This method detects MOs by comparing two adjacent frames based on discrete cosine transformed (DCT) coefficients' sign phase correlation (DCT-SPC) [6], and it doesn't require decoding the compressed movies.

Though the background variation between the adjacent frames is relatively small, the method [5] incorrectly detects MOs in the cluttered background. To overcome such problem, a number of methods use the adaptive background modeling for non-stationary scenes such as waving trees, lighting changes, and shadows [3]. A mixture of Gaussians model (GMM) that describes the temporal behavior of pixels [7] and its successors are currently attractive well for their performance.

This paper proposes a MOs detection method which uses a block-wise multi-modal background modeling for Motion JPEG movies. The proposed method constructs the multi-modal background model of the sign of DCT coefficients with respect to each DCT block. By the modeling and multi block-based processing, the method can robustly detect MOs. In addition, the proposed method does not require decoding of Motion JPEG movie as well as the conventional method [5].

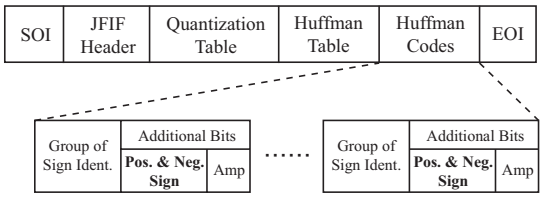

Fig. 1. JPEG Codestream.

II. Preliminary

This section mentions the algorithm and the codestream structure of Motion JPEG, and principles of MOs detection using the positive and negative sign of DCT coefficients [5].

A. Motion JPEG

Motion JPEG encodes each frame of a movie by JPEG [8]. In JPEG encoding, an original image is divided into *DCT block*s that respectively consist of 8×8 pixels. Then, two dimensional DCT is applied to each block in which a transformed block has one DC and 63 AC coefficients. DCT coefficients in each block are quantized according to the quantization table scaled by a Q-factor. Finally, AC coefficients are coded independently, and differences of DC coefficients between two consecutive DCT blocks are Huffman coded.

Fig. 1 shows a structure of JPEG codestream that is generated from a grayscale image. The start of image (SOI) marker is the head of a JPEG codestream and the JPEG File Interchange Format (JFIF) header contains information such as the image size. Then, the quantization table, the Huffman table, and Huffman codes are stored, where Huffman codes represent quantized DCT coefficients. Finally, the end of image (EOI) marker is put in the last of the codestream.

A Huffman code consists of an indicator of codeword group and additional bits. The latter part is further divided into the positive and negative sign bit and the coded amplitude of the corresponding DCT coefficient. Since this positive and negative sign is independent of other bits in the codestream, the sign can be directly obtained from the codestream without full decoding.

978-1-4244-5016-9/09/$25.00 ©2009 IEEE

B. MOs Detection Based on DCT Block Similarity

A similarity between a DCT block of two frames has been defined [5], as a special form of DCT-SPC [6]. In this section, the principle of MOs detection using this similarity is described.

1) Notations and Terminologies:

- s represents the sign of a DCT coefficient, i.e., take the value of 1 for the positive coefficient, -1 for the negative coefficient, 0 for the zero coefficient.
- $\mathbf{S}$ represents a vector consisting of the sign of 63 AC coefficients in a DCT block, i.e., $\mathbf{S} = [s_1, s_2, \ldots, s_{63}]$. In this paper, this vector is referred to as *DCT sign set*.
- $|\mathbf{S}|$ denotes the vector consisting of the absolute value of each element of vector $\mathbf{S}$, i.e., $|\mathbf{S}| = [|s_1|, |s_2|, \ldots, |s_{63}|]$.
- B represents the number of DCT blocks in a frame.
- i and j represent frame numbers, where $i \neq j$ and the first frame is represented as 0, i.e, $i, j \geq 0$.

2) MOs Detection by Comparison of Two Frames: Similarity $\sigma_{i,j,b}$ between the b-th DCT block ($b = 0, 1, \ldots, B-1$) of i-th frame and that of the j-frame is defined [5] as,

$$\sigma_{i,j,b} = \frac{\mathbf{S}_{i,b} \cdot \mathbf{S}_{j,b}}{\left|\mathbf{S}_{i,b}\right| \cdot \left|\mathbf{S}_{j,b}\right|}, \tag{1}$$

$$\mathbf{S}_{i,b} = [s_{i,b,1}, s_{i,b,2}, \cdots, s_{i,b,63}], \tag{2}$$

where $-1 \leq \sigma_{i,j,b} \leq 1$, and $s_{i,b,n}$ ($n = 1, 2, \ldots, 63$) represents the sign of the n-th AC coefficient, in the zigzag scan order, in the b-th DCT block of the i-th frame. The denominator of the fraction in Eq. (2) is the number of coefficients which are nonzero in both frames, and the numerator is the sum of 1, -1, 0 which respectively represents that the sign of the DCT coefficients in the same place of the two frames are "identical," "different," or "zero at least one." When the sign of all nonzero coefficients are the same in the b-th DCT block of the two frames, $\sigma_{i,j,b}$ reaches its maximum value 1.

If this similarity $\sigma_{i,j,b}$ is smaller than user-defined positive threshold σ_{th}, i.e.,

$$\sigma_{i,j,b} < \sigma_{\text{th}} \tag{3}$$

is satisfied, it is decided that a MO exists in the b-th DCT block of the i-th frame.

For further robust MOs detection, extended similarity between the b-th DCT block of the two frames is defined [5] as

$$\hat{\sigma}_{i,j,b} = \frac{\sum_{b' \in \mathbf{R}_b} \mathbf{S}_{i,b'} \cdot \mathbf{S}_{j,b'}}{\sum_{b' \in \mathbf{R}_b} \left|\mathbf{S}_{i,b'}\right| \cdot \left|\mathbf{S}_{j,b'}\right|}, \tag{4}$$

where $\mathbf{R}_b$ is a set of DCT blocks centering on the b-th DCT block, and b' represents the location of a DCT block in $\mathbf{R}_b$. For 3×3 blocks shown in Fig. 2, $\mathbf{R}_b$ is represented as

$$\mathbf{R}_b = \{b-4, b-3, b-2, b-1, b, b+1, b+2, b+3, b+4\}. \tag{5}$$

If $\mathbf{R}_b = \{b\}$, Eq. (4) comes down to Eq. (2). Then, if

$$\hat{\sigma}_{i,j,b} < \sigma_{\text{th}}, \tag{6}$$

as Eq. (3), it is decided that a MO exists in the b-th DCT block of the i-th frame.

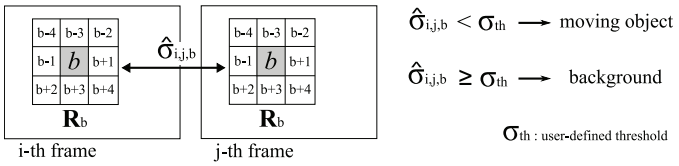

Fig. 2. MOs detection by comparison of two frames with comparison region $\mathbf{R}_b$ (3×3 blocks) for the b-th DCT block.

When i is positive odd number and $j = i - 1$, it corresponds to adjacent frame subtraction [5]. On the other hand, if background frame is used in place of the j-th frame, it corresponds to background subtraction.

III. Proposed Method

The proposed method models a multi-modal background by multiple DCT sign sets per DCT block, and constructs a DCT sign background frame in which each DCT sign set is respectively most appropriate as a background in a DCT block. The method detects MOs by comparing this DCT sign background frame to a DCT sign frame which consists of DCT sign sets extracted from an input frame.

A. Proposed Multi-Modal Background Model

The model for the b-th DCT block in the i-th frame consists of $\kappa_{i,b}$ components where $\kappa_{i,b} \leq K$ and K is given by a user. the k-th component ($k = 0, 1, \ldots, \kappa_{i,b} - 1$) has 2 items:

- $\mathbf{M}_{i,b,k}$: DCT sign set
- $w_{i,b,k}$: weight for $\mathbf{M}_{i,b,k}$ ($0 \leq w_{i,b,k} \leq 1$).

DCT sign set $\mathbf{M}_{i,b,k}$ represents vector

$$\mathbf{M}_{i,b,k} = [m_{i,b,k,1}, m_{i,b,k,2}, \ldots, m_{i,b,k,63}], \tag{7}$$

where $m_{i,b,k,n}$ ($n = 1, 2, \ldots, 63$) represents the sign of the n-th AC coefficient in the zigzag scan order of the b-th DCT block for the model of the i-th frame. The initial model consists of only one component which has a DCT sign set of the first input frame and the weight value 1, i.e.,

$$\kappa_{0,b} = 1, \tag{8}$$

$$\mathbf{M}_{0,b,0} = \mathbf{S}_{0,b}, \tag{9}$$

$$w_{0,b,0} = 1. \tag{10}$$

By adding components to the model as described in Section III-B, the model can represent up to K varying background scenes in each block.

B. Proposed MOs Detection Algorithm

The proposed algorithm is divided into two parts as shown in Fig. 3: the DCT sign background frame estimation and the similarity-based background subtraction. In this section, the two parts are successively described.

1) DCT Sign Background Frame Estimation: The following procedure is applied to the b-th DCT block of the i-th frame where $b = 0, 1, \ldots, B-1$.

1. Calculate similarity $\chi_{i,b,k}$ between input DCT sign set $\mathbf{S}_{i,b}$ and DCT sign set $\mathbf{M}_{i,b,k}$ of the k-th component ($k =$

$0, 1, \ldots, \kappa_{i,b} - 1$) as

$$\chi_{i,b,k} = \frac{\mathbf{S}_{i,b} \cdot \mathbf{M}_{i,b,k}}{\left|\mathbf{S}_{i,b}\right| \cdot \left|\mathbf{M}_{i,b,k}\right|}. \tag{11}$$

2. Among the DCT sign set of $\kappa_{i,b}$ components, the $L_{i,b}$-th DCT sign set is chosen to construct the DCT sign background frame, where

$$L_{i,b} = \arg\max_{k} \{\chi_{i,b,k} \times w_{i,b,k}\}. \tag{12}$$

After this step is executed at all of B DCT blocks, the DCT sign background frame for the i-th frame is constructed as shown in Fig. 4.

3. If

$$\chi_{i,b,k} > \chi_{\text{th}} \tag{13}$$

is satisfied where χ_{th} is a user-defined positive threshold, the k-th component is updated as

$$m_{i+1,b,k,n} = \begin{cases} m_{i,b,k,n}, & m_{i,b,k,n} = s_{i,b,n} \\ m_{i,b,k,n}, & s_{i,b,n} = 0 \\ s_{i,b,n}, & m_{i,b,k,n} = 0 \\ 0, & m_{i,b,k,n} \times s_{i,b,n} = -1 \end{cases}, \tag{14}$$

$$w_{i+1,b,k} = (1-\alpha)\, w_{i,b,k} + \alpha \times \max(\chi_{i,b,k}, 0), \tag{15}$$

where α is the learning rate ($0 \le \alpha \le 1$). If any one of the components is updated, skip the next step.

4. In the case of $\kappa_{i,b} = K$, replace the component having the minimum weight with the input DCT sign set and the initial weight 0, i.e.,

$$\mathbf{M}_{i+1,b,D} = \mathbf{S}_{i,b}, \tag{16}$$

$$w_{i+1,b,D} = 0, \tag{17}$$

$$D = \arg\min_{k} \{w_{i,b,k}\}. \tag{18}$$

Otherwise, i.e., $\kappa_{i,b} < K$, create a new component as,

$$\mathbf{M}_{i+1,b,\kappa_{i,b}} = \mathbf{S}_{i,b}, \tag{19}$$

$$w_{i+1,b,\kappa_{i,b}} = 0, \tag{20}$$

$$\kappa_{i+1,b} = \kappa_{i,b} + 1. \tag{21}$$

2) Similarity-Based Background Subtraction: The following procedure is applied to the b-th DCT block of the i-th frame where $b = 0, 1, \ldots, B-1$.

1. Calculate extended similarity $\hat{\chi}_{i,b}$ between the b-th block of the input DCT sign frame and that of the estimated DCT sign background frame as

$$\hat{\chi}_{i,b} = \frac{\sum_{b' \in \mathbf{R}_b} \mathbf{S}_{i,b'} \cdot \mathbf{M}_{i,b',L_{i,b'}}}{\sum_{b' \in \mathbf{R}_b} \left|\mathbf{S}_{i,b'}\right| \cdot \left|\mathbf{M}_{i,b',L_{i,b'}}\right|}, \tag{22}$$

where $\mathbf{M}_{i,b,L_{i,b}}$ represents the DCT sign set of the b-th DCT block in the DCT sign background frame for the i-th frame.

2. Determine that a MO exists in b-th DCT block of i-th frame, when $\hat{\chi}_{i,b}$ is smaller than threshold χ_{th}, i.e.,

$$\hat{\chi}_{i,b} < \chi_{\text{th}} \tag{23}$$

is satisfied.

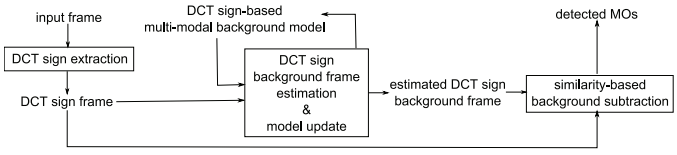

Fig. 3. proposed MOs detection algorithm.

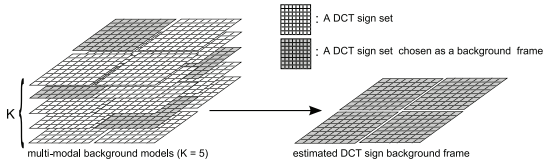

Fig. 4. background frame estimation (number of DCT blocks $B = 4$).

C. Features of Proposed Method

The proposed method robustly detects MOs by three features. 1) The method is based on the illumination-invariant similarity excluding the DC coefficient that is affected by global lighting changes. 2) The method models multi-modal background frame-by-frame to adapt time-varying background. 3) The method detects MOs per DCT block by using not only a focused block but also its surrounding blocks to reduce false detection as well as the conventional method [5].

In addition, because the sign of AC coefficients can be directly obtained from JPEG codestream as described in Sect. II-A, the proposed method doesn't require the decoding process beforehand as well as the conventional method [5].

IV. EXPERIMENTAL RESULTS

The proposed method is compared with the pixel-wise GMM method [7] and the DCT sign-based adjacent frame subtraction method [5] in terms of detection accuracy and computational time. The experimental conditions are summarized in Table I. The pixel-wise GMM method uses multiple distributions for each component of RGB, whereas the DCT sign-based adjacent frame subtraction method and the proposed method need only Y of YUV.

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) when the threshold of a system is changed. If both FNR and FPR of the system are smaller than those of other systems, i.e., the ROC curve of the system is closer to the origin than others, the system is superior to others in detection accuracy.

TABLE I
EXPERIMENT CONDITION.

Sequence	1920×1080 pixels, 24 bit color 15 fps, 120 frames Motion JPEG (Q-Factor:50)
CPU	Intel Core2 2.4GHz
RAM	2GB
OS	Linux 2.6.24
MATLAB	Version 7.5.0.338 (R2007b)

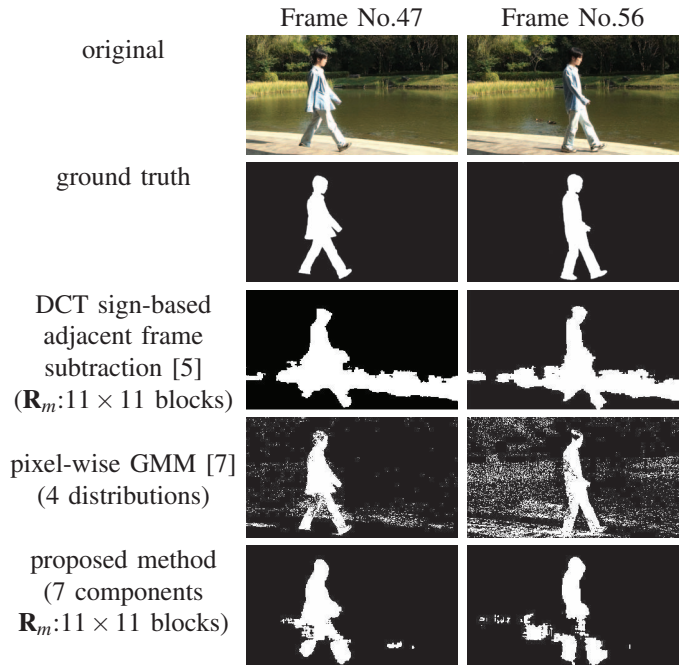

Fig. 5. detection results (each method uses the threshold which gives EER shown in Fig. 6).

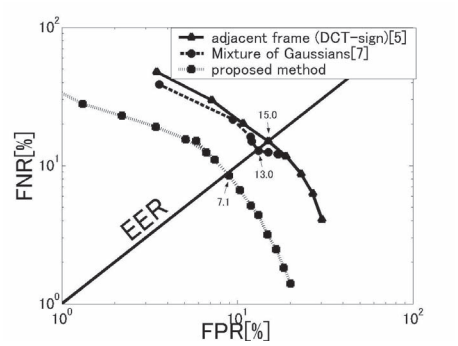

Fig. 6. ROC curves.

Also, equal error rate (EER) that is the rate at FNR = FPR for quantitative evaluation is used.

Fig. 5 shows detection results of each method in two example test frames. The proposed method adapts to waves of the pond by the multi-modal background model, whereas the DCT sign-based adjacent frame subtraction method [5] does not adapt to it. The proposed method also adapts to global lighting changes by DC coefficient elimination in the similarity calculation, whereas the pixel-wise GMM method [7] cannot adapt to it. ROC curves in Fig. 6, also, show that the proposed method is most accurate of three methods.

From Fig. 7, the proposed method achieved the best EER (7.1%) by using seven components, whereas the pixel-wise GMM method achieved the best EER (13.0%) at four distributions. Under these conditions, the proposed method is three times faster than the pixel-wise GMM method, c.f, Fig. 8. Moreover, the proposed method is five times faster than the pixel-wise GMM method at similar EER level (about 13%).

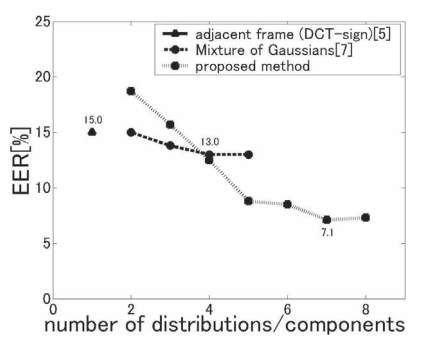

Fig. 7. Comparison detection accuracy (numbers indicate the best EER's).

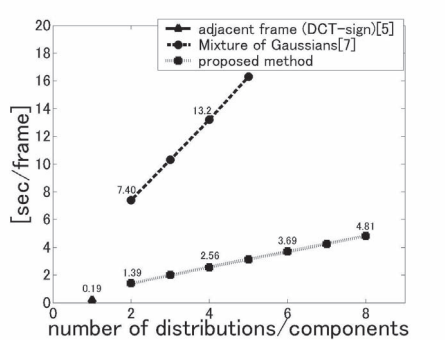

Fig. 8. Comparison computational time (computational time of the pixel-wise GMM method doesn't include decode time of a Motion JPEG movie).

V. Conclusions

This paper has proposed a robust MOs detection method for Motion JPEG movies. By describing a time-varying background with an adaptive DCT sign-based model, the proposed method can adapt to non-stationary scenes. Also, the multi block-based processing enhances the detection accuracy. Moreover the method doesn't require decoding Motion JPEG movies beforehand. Experimental results show that the MOs detection of the proposed method is more accurate and fast than the pixel-wise GMM method [7].

References

[1] A. Makarov, "Comparison of background extraction based intrusion detection algorithms," in *Proc. IEEE ICIP*, 1996, pp.521–524.
[2] A. Leykin and M. Tuceryan, "A vision system for automated customer tracking for marketing analysis: low level feature extraction," in *Proc. Human Activity Recognition and Modelling Workshop*, 2005.
[3] K. Toyama, J. Krumm, B. brumitt, and B. Meyers, "Wallflower: Principle and practice of background maintenance," in *Proc. IEEE ICCV*, 1999, pp.255–261.
[4] M. Piccardi, "Background subtraction techniques: a review," in *Proc. IEEE SMC*, 2004, pp.3099–3104.
[5] M. Fujiyoshi, K. Kuroiwa, and H. Kiya, "A scrambling method for Motion JPEG videos enabling moving objects detection from scrambled videos," in *Proc. IEEE ICIP*, 2008, pp.773–776.
[6] I. Ito and H. Kiya, "DCT sign-only correlation with application to image matching and the relationship with phase-only correlation," in *Proc. IEEE ICASSP*, 2007, pp.1237–1240.
[7] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE CVPR*, 1999, pp.246–252.
[8] *Information technology — digital compression and coding of continuous-tone still images — requirements and guidelines*, ISO/IEC International Standard 10918-1, 1994.