Exploring Contextual Redundancy in Improving Object-Based Video Coding for Video Sensor Networks Surveillance

Tsung-Han Tsai *Member, IEEE*, and Chung-Yuan Lin

***Abstract*—In recent years, intelligent video surveillance attempts to provide content analysis tools to understand and predict the actions via video sensor networks (VSN) for automated wide-area surveillance. In this emerging network, visual object data is transmitted through different devices to adapt to the needs of the specific content analysis task. Therefore, they raise a new challenge for video delivery: how to efficiently transmit visual object data to various devices such as storage device, content analysis server, and remote client server through the network. Object-based video encoder can be used to reduce transmission bandwidth with minor quality loss. However, the involved motion-compensated technique often leads to high computational complexity and consequently increases the cost of VSN. In this paper, contextual redundancy associated with background and foreground objects in a scene is explored. A scene analysis method is proposed to classify macroblocks (MBs) by type of contextual redundancy. The motion search is only performed on the specific type of context of MB which really involves salient motion. To facilitate the encoding by context of MB, an improved object-based coding architecture, namely dual-closed-loop encoder, is derived. It encodes the classified context of MB in an operational rate-distortion-optimized sense. The experimental results show that the proposed coding framework can achieve higher coding efficiency than MPEG-4 coding and related object-based coding approaches, while significantly reducing coding complexity.**

***Index Terms*—operational rate-distortion theory, contextual redundancy coding, object based video coding, intelligent video surveillance, visual sensor network.**

I. Introduction

Intelligent video surveillance systems are attempted to incorporate content analysis processing tasks (e.g. motion detection [1], object tracking [2], behavior analysis [3], [6]) to understand events happened in a site. Content analysis processing tasks can be further embedded or distributed within a camera network. Such camera network is called intelligent video sensor networks [5] (VSN) as shown in Fig. 1. In the intelligent VSN, each sensor node is tasked to capture video data and is capable to perform specific content analysis tasks to extract information from the video. The captured video and the extracted information are delivered to an aggregation node (AN). The role of the AN is to process the collected data and deliver important information to the base station (BS) [4], [5]. Intelligent VSN has been envisioned for a wide range of important applications, including security monitoring, environment tracking, and assisted living [5]. A design with distributed interactive video arrays for situation awareness of traffic and incident monitor was presented in [7]. Multicamera tracking system with a fully decentralized handover procedure between adjacent cameras was proposed in [8]. A distributed accident management system for assisted living was presented in [9].

In the aforementioned VSNs, the intelligence toward automated surveillance is succeeded by various content analysis tasks in different nodes. The content analysis tasks extract specific features from visual object data and generate semantic information by the extracted features. Visual object data and semantic information, thus, are transmitted from one node to another node over the network. However, the raw data transmission of visual object could be very expensive. For example, transmission bandwidth of a visual object which occupies a 100*100 pixels region with 4:2:0 format in one node requires about 3Mbits per second. Moreover, the rest visual objects are probably transmitted for storage of a complete frame which leads to more than 3Mbits for that node. The transmission bandwidth of VSN can further increase in direct proportion to the number of sensor nodes due to many-to-one network architecture (see Fig. 1).

Consequently, frame-based video coding technique (e.g.

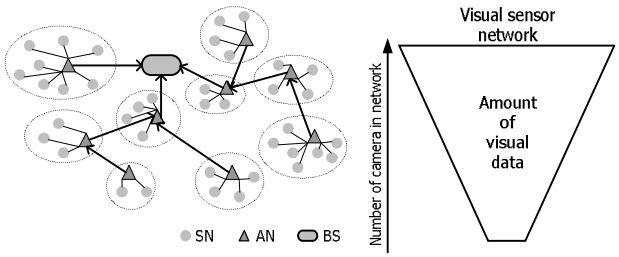


Fig. 1. Topology of intelligent video sensor networks. The many-to-one network requires large amounts of visual data to be compressed.

Manuscript received December 15, 2010; revised June 20, 2011. Date of publication November 30, 2011. This work was supported in by the National Science Council, Taiwan, under Grant 100-2220-E-008-001. This paper was recommended by Associate Editor N. Thinh.

T.-H. Tsai is with the Department of Electrical Engineering, National Central University, Jhongli City, Taoyuan 3201, Taiwan (e-mail: han@ ee.ncu.edu.tw).

C.-Y. Lin is with the Department of Electrical Engineering, National Central University, Jhongli City, Taoyuan 3201, Taiwan (e-mail: yashiro@dsp.ee.ncu.edu.tw).

MPEG-2, H.263 or H.264/AVC) can be applied to alleviate the burden on transmission by reducing the size of the visual object data with small quality loss. However, frame-based techniques may lack flexibility on transmission. Since the content analysis tasks are almost based on processing visual object data (e.g., object tracking), transmission of entire frame occupies network bandwidth when only few visual objects are required by a node. Additionally, the received frame has to been fully decoded to acquire the data needed for a content analysis task. Therefore, each visual object is desired to originate its own bitstream in such VSN scenario.

Many object-based coding approaches concentrated on encoding of surveillance video have been proposed in past few years. These object-based coding approaches can be broadly classified by the encoding architecture. The first kind of coding approach consists of a single-open-loop in which an intraframe-only coding scheme is involved. The typical coding architecture of single-open-loop is illustrated in Fig.2. In [12], a JPEG2000 based transcoder is proposed for transmission of visual object data which is formed by rectangle-specified formation as example shown in Fig. 3. Since the visual data associated with a rectangular window is compressed and transmitted, the precision of the visual object is insufficient to meet the requirement of the content analysis tools. [13] improves the precision of visual object by using pixel-specified formation (see Fig. 3) to facilitate the requirement of the content analysis tools. Here, region-of-interest coding tool standardized by Motion-JPEG2000 is applied to encode the visual object. The drawback of the single-open-loop is that limited coding efficiency is improved since only intraframe redundancy is exploited.

The second kind of coding approach consists of a single-closed-loop in which the motion-compensated technique is involved. The typical coding architecture of single-closed-loop is illustrated in Fig.2. Frame compressed by motion search is referred to as motion-compensated coding. Due to interframe redundancy is exploited, the coding

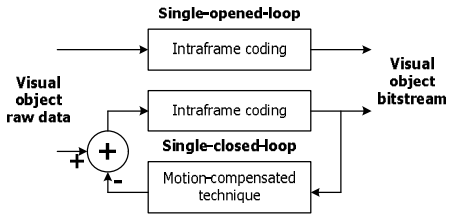


Fig.2. Illustration of single-opened-loop and single-closed-loop block diagrams.

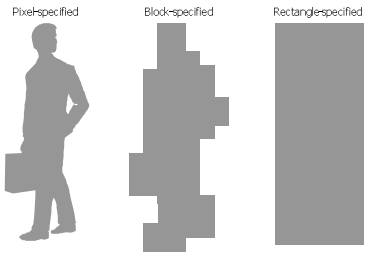


Fig. 3. Illustration of object formations in the object-based surveillance video coding approaches.

TABLE I
Classification of existed approach and the proposed approach

Category	Example	Applied tool	Visual object formation
Single-open-loop	[12]	Motion JPEG	Rectangle-specified
	[13]	Motion JPEG2000	Pixel- specified
Single-closed-loop	[14]	H263	Block- specified
	[15], [17]	MPEG-4	Pixel- specified
	[16]	MPEG-4	Block- specified
	[18]	MPEG-4 and chain code	Pixel- specified
One closed loop and one opened loop	[24]	MPEG-4 and JPEG2000	Pixel- specified
Dual-closed-loop	This work	MPEG-4	Pixel- specified

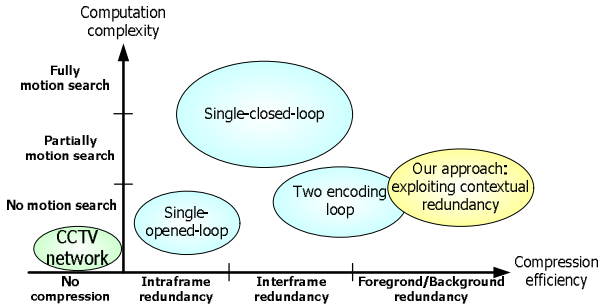
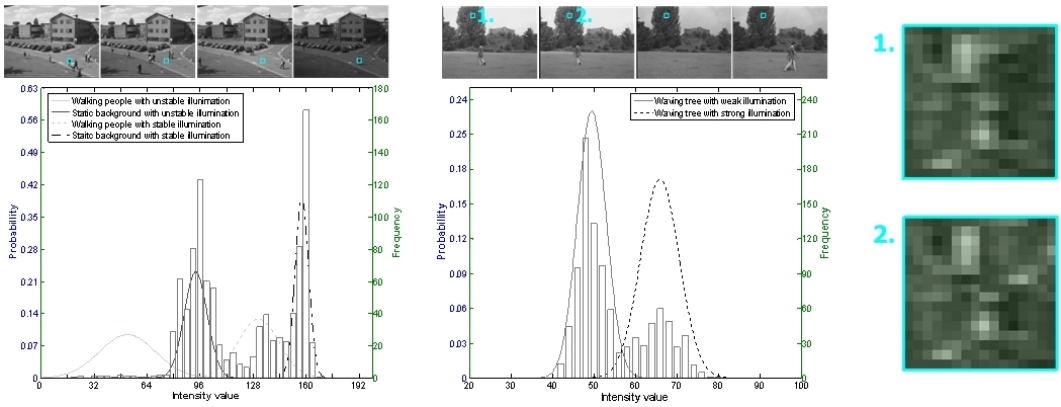


Fig.4. Classification of object-based video surveillance coding approaches with respect to compression efficiency and computation complexity.

efficiency of single-closed-loop can be much better than the single-open-loop. [14] proposed a single-closed-loop approach in which the visual object is formed by block-speficied formation (see Fig. 3) and encoded by H.263. [15] presented a system where visual object is formed by pixel-specified formation and encoded by MPEG-4. Shape information associated with an arbitrary visual object is coded by MPEG-4 standardized shape coding tool. Similar approach can be found in [17]. In [16], the visual object is further classified as foreground object and background object for different bitrate allocation. The approach in [18] is concerned with shape coding efficiency in video surveillance. The chain coding is utilized as the visual object shape coding tool and combined with shape adaptive DCT for encoding texture of the visual object.

The total drawback of the single-closed-loop approach is that both background and foreground objects are coded in the same loop. Since induced redundancies in background and foreground are quite different, single-closed-loop approach therefore cannot fully exploit these redundancies. The third kind of coding approach, which consists of two encoding loops for foreground object and background object respectively, can further improve the encoding efficiency. In [24], an efficient coding strategy for background is reported where the background is treated as an entire frame instead of an arbitrary shape of visual object. The background frame is generated by a low-pass filter and encoded by JPEG2000 at time intervals. A classification of the existing approaches based on the coding architecture is categorized in Table I. Due to different design philosophies between conventional video compression (i.e. one-to-many for broadcast application) and surveillance video compression (i.e. many-to-one as shown in Fig.1) [10],



(a) (b) (c)

Fig. 5. Averaged intensity distribution histogram of a specified MB over a period. (a) and (b) show the scene with illumination changes and background variations respectively. Bin represents the frequency of intensity, and the line represents the fitted Gaussian model with respect to bins. (c) shows the detailed pixel values in the block in (b).

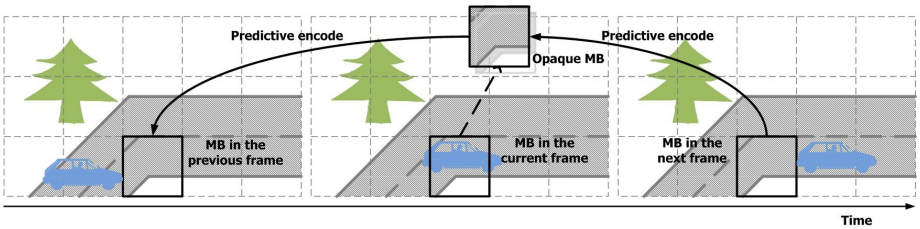


Fig. 6. Illustration of the purpose of the opaque MB.

applying aforementioned object-based approaches related to closed encoding loop architecture will increase the cost of VSN when a mass deployment of end cameras is needed. This is because high computation complexity of near fully motion search produces for entire frame (see Fig. 4).

In this paper, a novel approach with contextual redundancy associated with background and foreground objects in a scene is explored. Our approach aims to go beyond general object-based video coding techniques by exploiting the contextual redundancy of macroblocks (MBs) as shown in Fig. 4. The context of MB, here, is defined based on the temporal distribution of a MB in the surveillance video sequence. An improved object-based coding architecture, namely dual-closed-loop coding is proposed to encode the context distinct MB in an operational rate-distortion-optimized scene. The object-based coding tools standardized by MPEG-4 are utilized in this work. The rest of this paper is organized as follows. Section II formulates scene analysis for context MB classification. Section III details the proposed framework. Section IV conducts experimental results and comparison. Section V concludes this paper.

II. Context of Scene Analysis

This paper targets on the surveillance scene captured by the fixed sensor node since it is very common in VSN. The captured surveillance video is a special kind of video sequence in which induced redundancies in background and foreground are quite different. A current pixel in a common surveillance scene can be associated with static background, moving background and moving foreground with/without illumination change according to observation on several recent frames. Such information is called the context of scene in this paper. The context of scene corresponds to the distribution of temporal pixels over several recent frames. For example, the context is likely to be static background without illumination change if the temporal distribution of a pixel is constant for a long time. By specifying the context of scene of pixels in a MB, a most representative context for the MB can be determined. Such information is called the context of MB. In the following, the aspect on modeling context of MB is discussed and the redundancy for different type of context of MB is analyzed.

A. Exploiting Context of Scene for Coding

The value of each pixel represents a measurement of the radiance in the direction of the sensor of the first object interested by the pixel's optical ray. Suppose that a background is static and the illumination is static too, temporal distribution of intensity value of a background pixel is constant. Since Gaussian noise is incurred in the procedure of image acquisition, intensity value of background pixel can be modeled by a single Gaussian distribution. However, most surveillance sequences involve illumination changes and moving objects. If illumination changes occurred in a static background, it would be necessary for the Gaussian to adapt to those changes. Besides, background variation occurs if moving background objects are present in the scene. In general, most of background motions are periodic, there should be repeated pixel values supporting Gaussian distribution. These factors suggest a need for multiple adapted Gaussian distributions to model context of scenes in several recent frames.

Fig. 5 shows the distribution of averaged intensity of a specific MB over 1000 frames and its fitted Gaussian distributions. Each fitted Gaussian distribution is associated

with a specific context of the scene. Here, MB is defined as a set of non-overlapping block of 16×16 pixels. In the scene shown in Fig. 5a, the illumination frequently changed within first 600 frames, and then it became stable. Accordingly, the Gaussian for static background is tight and so have high probabilities, whereas the Gaussians for foreground are dispersed and so have low probabilities. As the illumination changes occurred, the Gaussian is slightly separated and so have probabilities lower than the probabilities of Gaussian of static background. Therefore, the probability of Gaussian implies that how many redundancies in the current foreground/background MB can be predicted from the previous foreground/background MB. A static background MB usually has highest probability because the pixels in the MB are repeated, whereas a foreground MB has lowest probability because the pixels therein are not repeated. As illumination changes occurred in a background MB, the probability is degraded such that a certain part of pixels in the current MB are slightly different from the pixels in the previous MB. The same situation can be observed from the MB which contains moving background pixels as that shown in Fig. 5b. The detailed pixel values in two consecutive dynamic background MBs are shown in Fig. 5c. It shows that some pixels are indeed slightly different while the rest pixels are remained the same.

The implicated redundancy associated with context of MB is called contextual redundancy and is used to improve the coding efficiency in facts: 1) the contextual redundancy, related to foreground object motion, is low. Such MB should be inter-predicted from previous frame by motion search. 2) the contextual redundancy, related to the illumination change in background and the background object motion, is partial. Such MB can be directly predicted from its previous MB. 3) the covered background pixels can be exposed by parameters of its Gaussian so that the next MB can be predicted from the current one (see Fig. 6 as example). To achieve this, opaque MB is introduced. Another advantage of introducing opaque MB is that bits spent on shape coding can be saved. Therefore, each current MB is classified into *object MB* (*oMB*), *opaque background MB* (*pMB*), *static background MB* (*sMB*) or *dynamic background MB* (*dMB*) to exploit contextual redundancy. Table II summaries the relationship between context of MB and context of scene.

TABLE II
Evaluation of context distinct MB classification accuracy.

Context of MB	Handled context in surveillance scene
Object MB	Any object with noticeable motion, typically the moving foreground object
Static background MB	Background with camera noise
Dynamic background MB	Background with illumination change, background with unnoticeable motion
Opaque background MB	Any background covered with moving object

B. Context of Scene Analysis

Determining the context for a current MB directly by its fitted Gaussian distributions (e.g. Fig. 5a and Fig. 5b) is difficult due to two reasons. First, the type of context of scene appeared in sequence cannot be known in prior. For example, the tightest Gaussian in Fig. 5a should be classified as *sMB*, whereas the tightest Gaussian in Fig. 5b should be classified as *dMB*. Suppose that the contexts of scene for Fig. 5b are not known in prior. The MB would probably be misclassified as *sMB*. Second, the Gaussians associated with the same type of context of MB are varied with sequence. Examples are the Gaussian for unstable illumination (Fig. 5a) and the Gaussian for waving tree (Fig. 5b). These two reasons raise the difficulties on directly modeling each type of context of MB.

A solution to overcome this problem is to indirectly model the context of MB by an estimated background image (*EB*). By comparing the diversity between the MB in current frame and the MB in the *EB*, the context of each MB can be well determined. This concept is illustrated in Fig. 7. Each pixel value of the *EB* is the mean value of the most representative Gaussian. The most representative Gaussian is the Gaussian that accounts for a maximum portion of temporal pixels over several recent frames. Let a set $\{X_{t-Np+1},\ldots,X_t\}=\{I_{x,y,i}:t\text{-}Np+1 \le i \le t\}$ contained Np number of intensity of consecutive temporal pixels be denoted as consecutive temporal pixels set of a given pixel, (x,y), and let $I_{x,y,t}$ be denoted as the intensity value of pixel (x,y) at time t. Then, the most representative Gaussian is selected as B-th Gaussian if it accounts for a maximum portion of pixels in $\{X_{t-Np+1},\ldots,X_t\}$. Taking the Fig. 5a as example, the most representative Gaussians is the Gaussian for static background with unstable illumination. The mean value of such Gaussian is, therefore, the average value of the maximum portion of pixels. Accordingly, this average value best describes the background pixel about the consecutive temporal pixels set.

The goal of the scene analysis method presented here is to select the most representative Gaussian, the B-th Gaussian, from mixture of Gaussians (MoG) [1], [20]. Specifically, the set $\{X_{t-Np+1},\ldots,X_t\}$ is modeled with a mixture of K Gaussians, and each Gaussian represents a context of scene. The probability of observing the current pixel value X_t, given K context of scenes is

$$p(X_t)=\sum_{i=1}^{K}\omega_{i,t}\cdot\eta_i(X_t,\mu_{i,t},\sigma_{i,t}^2) \quad (1)$$

where $\eta_i(.)$ is the i-th Gaussian distribution and $\omega_{i,t}$, $\mu_{i,t}$, and $\sigma^2_{i,t}$ are the weight, mean, and variance of i-th Gaussian,

$$\eta_i(X_t,\mu_{i,t},\sigma_{i,t}^2)=(2\pi\sigma_{i,t}^2)^{-1/2}\exp(-0.5(X_t-\mu_{i,t})^2\sigma_{i,t}^{-2}) \quad (2)$$

parameters of K Gaussians are estimated in order to fit the mixture distribution to the set $\{X_{t-Np+1},\ldots,X_t\}$ at any time t.

A standard method for maximizing the likelihood of the observed data in $\{X_{t-Np+1},\ldots,X_t\}$ is *expectation maximization* (*EM*). However, implementing an exact *EM* algorithm on the set $\{X_{t-Np+1},\ldots,X_t\}$ would be costly due to iterative computation for the set. Instead, the on-line K-means approximation presented by Stauffer and Grimson [1] is implemented. The on-line K-means approximation is performed in a manner

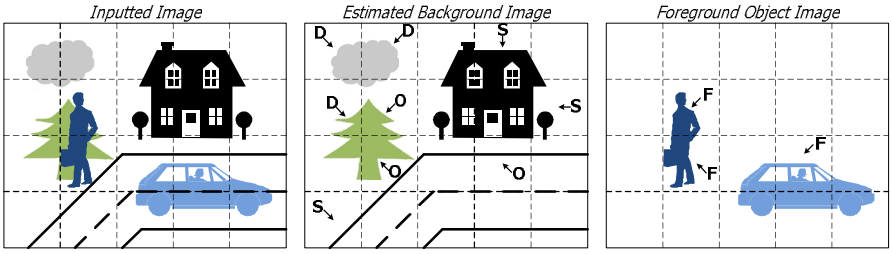


Fig. 7. Example of context distinct MBs. F represents the *object MB*, O represents the *opaque background MB*, D represents the *dynamic background MB*, and S represents the *static background MB*.

similar to *incremental EM* algorithm [27]. By the on-line K-means approximation, the posterior probability of Gaussians q for an incoming pixel is recalculated in a manner consistent with the *expectation* and only one Gaussian with the maximum posterior probability q is updated in a manner consistent with the *maximization*. The only different thing between them is that the on-line K-means approximation does not iterate for the set. Instead, it treats each incoming pixel X_t as a sample set of size 1 and uses learning rules to integrate the sample set with the previous set $\{X_{t-Np},\ldots,X_{t-1}\}$ at time t-1 [1]. Therefore, the on-line K-means approximation is computationally efficient. The estimated posterior probability q_i for i-th Gaussian is given as

$$q_i = \frac{\omega_{i,t-1}\cdot\eta_i(X_t,\mu_{i,t-1},\sigma^2_{i,t-1})}{\sum_{j}^{K}\omega_{j,t-1}\cdot\eta_j(X_t,\mu_{j,t-1},\sigma^2_{j,t-1})} \tag{3}$$

Note that the posterior probability of i-th Gaussian is estimated by its parameters at time t-1 because its parameters at time t have not maximized yet.

Utilizing the on-line K-means approximation to estimate parameters of MoG will encounter two problems. First problem is related to the winner-take-all update for overlapping Gaussians. This is the case, for instance, when a tree is shaken by the wind (e.g. Fig 5b) or when the background includes wavy water. Selection of only a maximum expected Gaussian will lead to starvation where one Gaussian stretches increasingly more expected posterior probability to over-dominate the others. Such situation would cause imprecise model for the context of scene and be wrong to determine the most representative Gaussian. To overcome this problem, soft-partition strategy [21], which updates rest Gaussians by an amount proportional to its respectively posterior probability q, is introduced [20]. The estimated posterior probability q_i for i-th Gaussian is redefined as

$$q_i = \begin{cases} 1, & \text{if } i = \arg\max_j\{q_j\} \\ q_i, & \text{otherwise} \end{cases} \tag{4}$$

Second problem is related to the update with a fixed learning rate. Once a new Gaussian is reassigned, the learning rate is desired to be weighted proportionally to the number of observed pixel fitted to the new Gaussian [20]. Such adapted learning rate allows parameters of the new Gaussian quickly converge to the expected values at initial. In our scenario, the adapted learning rate is particularly suited to update Gaussian for background object motion and Gaussian for illumination change, such that the *EB* can more reflect the true background image. In this paper, the learning rate of i-th Gaussian is adapted to its posterior probability q and its weight ω at time t and is derived as

$$\alpha_{i,t} = q_i\cdot\left(\frac{Np-1}{Np+Np\cdot\omega_{i,t}}+\frac{1}{Np}\right) \tag{5}$$

the learning rate is content adaptive. If i-th Gaussian is reassigned, the learning rate is followed as $\alpha_{i,t}\approx q_i/(1+\omega_i)$ since the weight ω_i is very small at initial. After more than Np pixels are observed, the learning rate then approximates to $\alpha_{i,t}\approx q_i/(0.5\cdot Np)$.

Given an incoming pixel (x,y) at time t, the procedure of scene analysis starts with computing the posterior probabilities q_i for each Gaussian by (4). The posterior probability describes how correspondence between the incoming pixel with a Gaussian in the mixture at time t-1. Once all expected posterior probabilities q are approximated to zero, this situation means that a new context of scene appears in a pixel (x,y) and a new Gaussian associated with the new context should be reassigned. In such case, the Gaussian with the minimum weight value is replaced with the new one. Let L-th Gaussian be denoted as the Gaussian with the minimum weight and be defined as follows

$$L = \arg\min_{i}^{K}\{\omega_{i,t}\} \tag{6}$$

mean, variance and weight of L-th Gaussian are replaced with the current pixel $I_{x,y,t}$, an initial high variance [1] (in our case, $\sigma^2_{L,t}$=400), and $1/Np$.

After posterior probability of each Gaussian is recalculated, weight of each Gaussian is updated by its expected posterior probability to achieve the soft partitioning [20], [21].

$$\omega_{i,t} = \frac{Np-1}{Np}\cdot\omega_{i,t-1}+\frac{1}{Np}\cdot q_i \tag{7}$$

such update manner removes contribution from the foremost pixel X_{t-Np} to the sum with exponentially decay and adds contribution from the current pixel X_t. Similar to the updating weight, mean and variance of each Gaussian are updated by the respective learning rate as follows

$$\mu_{i,t} = (1-\alpha_{i,t})\cdot\mu_{i,t-1}+\alpha_{i,t}\cdot X_t \tag{8}$$

$$\sigma^2_{i,t} = (1-\alpha_{i,t})\cdot\sigma^2_{i,t-1}+\alpha_{i,t}\cdot(X_t-\mu_{i,t-1})^2 \tag{9}$$

After parameters of Gaussians at time t are maximized, the most representative Gaussian (i.e., B-th Gaussian) is selected

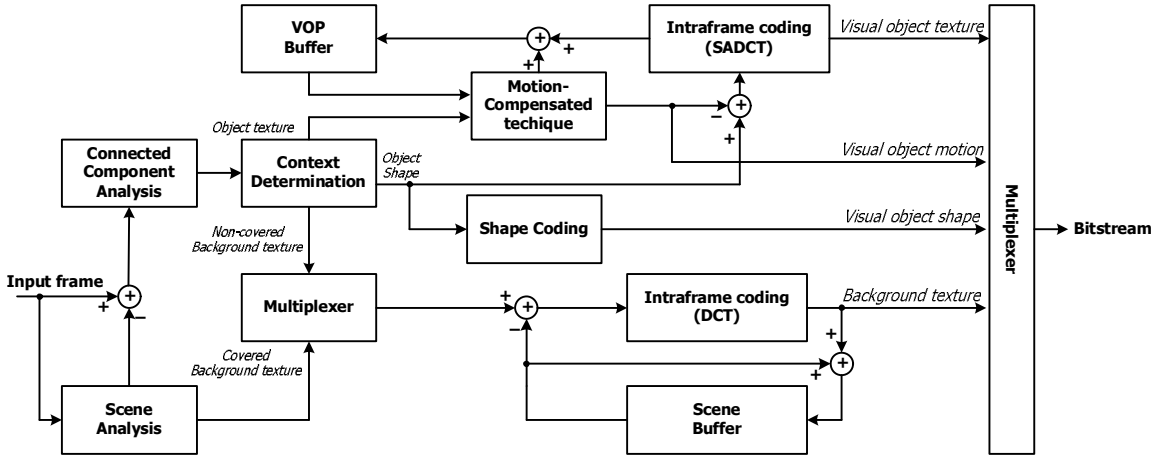


Fig. 8. The dual-closed-loop encoding approach.

as the Gaussian with the maximum weight value in the mixture. B-th Gaussian is defined as follows

$$B = \arg\max_{i}^{K}\{\omega_{i,t}\} \tag{10}$$

Exploiting B-th Gaussian of each pixel, the EB is composed by the mean value of each B-th Gaussian. The method that uses the EB for the context of MB determination is described in the Section III.

III. Video Surveillance Coding Using Context Of Scene

To encode different type of context of MB in an operational rate-distortion-optimized scene, the dual-closed-loop coding architecture is presented. The block diagram of dual-closed-loop coding is shown in Fig. 8. It performs one loop by inter prediction with motion estimation and the other loop by inter prediction with difference only. The scene analysis block periodically estimates the most representative Gaussian and generates the estimated background image EB. The context determination block classifies each MB by the image EB. Each classified MB is encoded with its associated loop depending on type of the determined context. Accordingly, foreground objects and background objects are individually encoded and originate its own bitstream that facilitates the flexible transmission in intelligent VSN.

A. Context Determination of MB

The difference value between pixels in the current frame and pixels in the EB is selected as the measurement of diversity between the current context of scene and the most representative Gaussian. The difference values are processed by the context determination procedure which involves a two-step examination. The oMB is obtained in the first step because the difference values therein are clearer than differences in the sMB and the dMB. Let a MB be defined as a non-overlapping spatial block of 16×16 pixels starting at (x,y).

$$MB_{x,y} = \{(i,j) : 0 \le i - x \le 15;\ 0 \le j - y \le 15\} \tag{11}$$

A MB is regarded as an oMB if there exist pixels that are dissimilar with pixels in EB at the same position (see Fig. 7). It is obtained by thresholding the absolute difference value with a pre-defined threshold value TH_D. In order to encode each individual object, the connected component analysis is subsequently applied to label connected pixels whose difference values exceed the threshold TH_D. Let O_m, $0 \le m \le M$, be the objects in the given scene, and each object

$$O_m = \{(x,y) : I_{x,y} - EB_{x,y} \ge TH_D\}, \tag{12}$$

contains connected pixels and the number of connected pixel is large than a pre-defined threshold TH_O. Setting the object size threshold TH_O is to avoid wasting bits on encoding shape of small object because such object is visually discerned. The value of threshold TH_O is simply given as an amount of 0.1% of total pixels in a frame. Then, a MB is determined as oMB if there exist pixels corresponding to m-th object O_m

$$oMB_{x,y} = \{(i,j) : (i,j) \in MB_{x,y},\ MB_{x,y} \cap O_m \ne \varnothing\} \tag{13}$$

A pMB contains the background pixels covered with object. Therefore, a pMB should be determined if an oMB at the same position is determined

$$pMB_{x,y} = \{(i,j) : (i,j) \in oMB_{x,y}\} \tag{14}$$

note that the oMB and the pMB are conjugated, and both of them are concurrently encoded.

Once a MB is not determined as oMB in first pass, it must belong to sMB or dMB. The discrimination between sMB and dMB is further obtained in the second pass. A MB is regarded as a sMB if it is very similar to the MB in EB at the same position, whereas a MB is regarded as a dMB if the similarity lies between sMB and oMB. However, there would exist difference values in dMB which are closed to the difference values in sMB. Thus, making decision by directly comparing difference values with a threshold value such as (12) would be too sensitive. Since the difference values in the sMB are associated with camera noise (the definition in Table II), the discrimination between dMB and sMB then is turned to discriminate differences belonging to camera noise or not. This is achieved through the significant test [25].

Specifically, the determination for a MB as a sMB or a dMB corresponds to choosing one of two competing hypotheses. The null hypothesis H_0 implies that the absolute differences obeying

zero-mean Gaussian distribution by camera noise, i.e. s*MB*. The alternative hypothesis is that if the error rejects zero-mean Gaussian distribution, i.e. *dMB*. Aach et al. [19] proposed a significant test method where pixel-based normalized squared sum is used to select the hypothesis based on *chi-squared* distribution. In this paper, the normalized squared sum is extended to MB-based. The normalized squared sum $\Delta^2_{MB_s}$ of an undetermined MB is defined against variance associated with the most representative Gaussian as

$$\Delta^2_{MB_{x,y}} = \sum_{(i,j)\in MB_{x,y}} \sigma^{-2}_{B_{i,j}} (I_{i,j} - EB_{i,j})^2, \forall MB_{x,y} \notin oMB \qquad (15)$$

where $B_{i,j}$ is the most representative Gaussian of a pixel (i,j). Then a threshold TH_B is specified to the normalized squared sum of the MB. As the normalized squared sum $\Delta^2_{MB_{x,y}}$, given the H_0, is known to obey a *chi-squared* distribution χ, the threshold TH_B is determined from the false alarm rate β [19].

$$\chi(TH_B \mid H_0) = \beta \qquad (16)$$

Once an acceptable false alarm rate β has been chosen, an uncovered background MB is regarded as *sMB* if its normalized squared sum is smaller than TH_B

$$sMB_{x,y} = \{(i,j) : (i,j) \in MB_{x,y}, \Delta^2_{MB_{x,y}} \le TH_B\} \qquad (17)$$

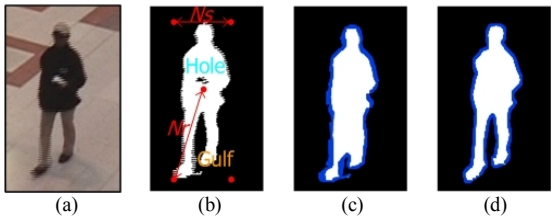
otherwise, it is regarded as *dMB*

$$dMB_{x,y} = \{(i,j) : (i,j) \in MB_{x,y}, \Delta^2_{MB_{x,y}} > TH_B\} \qquad (18)$$

B. The operational rate-distortion-optimized encoding approach

The dual-closed-loop approach encodes context distinct MBs based on two separate motion-compensated loops. The *oMB*s corresponding to a same object are encoded by the loop involved motion estimation and the shape coding block. With a block-based motion-compensated coding approach, video frame are encoded by taking the DCT of 8×8 blocks of the intensity or the displaced frame difference. However, there are 8×8 blocks which are partially occupied by an object when shape information is used. Although the actual number of pixels belonging to the object would be smaller, we would still need to transmit 64 coefficients by using 8×8 DCT for such blocks. Shape-adaptive DCT (SADCT) provides for a way of encoding such blocks using a number of coefficients that is equal to the number of the object pixels in the block [22]. Accordingly, the loop for coding *oMB*s is accomplished by quantization of the SADCT coefficients followed by variable length coding as the entropy coding.

However, small but distinct features of original boundary may disappear (be "cut off") in its approximation for low bit rate. Besides, if an object contains holes or cluttered shape; more bits will be wasted for shape encoding [23]. To facilitate object texture coding by SADCT, it is beneficial with a smooth shape and a shape approximation in areas of high gradient magnitude. This is due to that the coefficients of *oMB* with a broken shape will result in discontinuity and leads to encode them with more bits. Several works considered the encoding of



(a) (b) (c) (d)

Fig. 9. A comparison of shape information between the precise object and the smooth object. (a) shows the original sub-image. (b) shows the segmented object by eq.(12). (c) shows the resulted object by applying the content adaptive reconstruction by closing to object in (b). (d) shows the precise object by sophisticated segmentation algorithm [31].

the boundary as thought with the smallest possible distortion satisfying a bit budget constraint [23], [26]. However, these approaches are computation extensive because minimizing a cost function with the shape is involved.

To deal with such shape-oriented problem in a more efficient way, the mathematical morphology technique is introduced. This set theoretic, shape oriented approach treats the image object as a set and a kernel of operation, commonly known as structuring element *se*, as another set [28]. The basic operations of mathematic morphology are erosion, dilation, opening and closing. In general, erosion shrinks image objects, while dilation expands them. Opening and closing are the cascades of dilation and erosion with different orders. Opening cuts out narrow isthmuses, suppresses island smaller than the used *se*, whereas closing fills in small holes and thin gulfs. Among them, closing is utilized at two aspects. First, it can be used as optimal approximation of shape by removing cluttered object boundary where it may be expensive in terms of coding but not visually important. Second, holes in object can be filled. Therefore, undesired boundaries of holes are removed and bits for encoding those shapes are saved. Applying this operation to binary image involve only logic AND and OR operations, and so it is considered as a low cost solution here. However, closing does not allow a perfect preservation of the contour information [29]. In order to improve the contour preservation properties, closing by reconstruction [29] is used.

Closing by reconstruction with a flat structuring element *se* of size *Ns* is defined as

$$\varphi^{Nr}(\delta^{Ns}(f), f) = \cdots \varepsilon_g(\cdots \varepsilon_g(\delta^{Ns}(f), f) \cdots, f) \qquad (19)$$

where f is the targeted binary image, $\delta(f)$ is dilation, and the superscript *Ns* denotes that the size of applied structuring element *se* is a *Ns*×*Ns* window. The geodesic erosion ε_g, which involves geodesic transforms image $\delta^{Ns}(f)$, is always defined with respect to a reference function f and given as $\varepsilon_g(\delta^{Ns}(f),f) = \varepsilon(\delta^{Ns}(f)) \cup f$. The superscript *Nr* denotes the number of iteration where geodesic erosion is obtained for reconstruction, such that $\varphi^{Nr}(\delta^{Ns}(f),f) = \varphi^{Nr-1}(\delta^{Ns}(f),f)$ is satisfied. Applying (19) to entire binary image raises the problem that the same parameters are applied for objects with different size and different shape. That is, in such shape oriented processing for each object, the number of reconstruction iteration *Nr* and the size of structuring element *Ns* should depend on the shape complexity

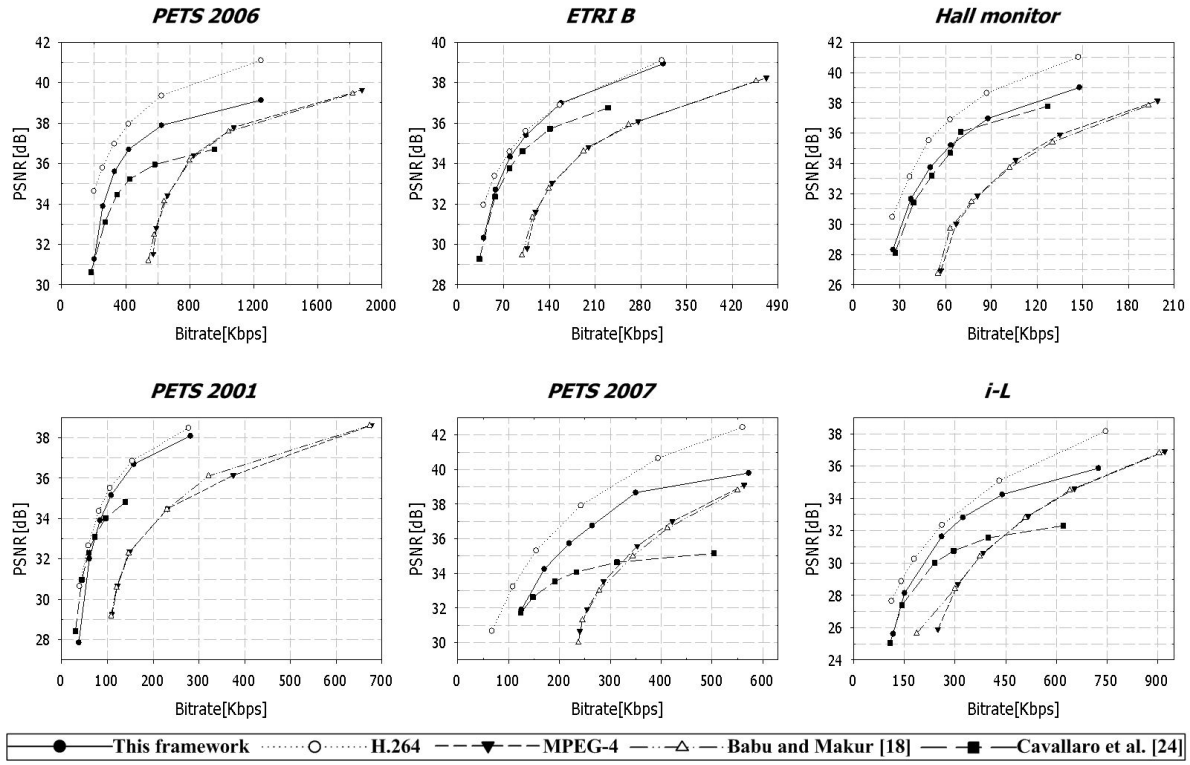


Fig. 10. Rate-distortion performance comparison for 6 test sequences.

and the size of the object.

In this paper, content adaptive closing by reconstruction is proposed. It automatically determines *Nr* and *Ns* based on the geodesic of shape. This method proceeds as follows. First, size information of an object O_m is extracted. Let a set V contain four pixels $V=\{(x_{min}, y_{min}), (x_{min}, y_{max}), (x_{max}, y_{min}), (x_{max}, y_{max})\}$ corresponding to a rectangle that encompasses *m*-th object, and one pixel (x_c, y_c) corresponds to the mass center of the object. Then, the size of structuring element for *m*-th object is estimated as the minimum distance between the object height and the object width

$$Ns=\left\lfloor \min\left(x_{max}-x_{min}, y_{max}-y_{min}\right)\right\rfloor \tag{20}$$

the number of reconstruction iteration is estimated as the maximum distance between four vertices of rectangle and the mass center

$$Nr=\left\lfloor \max\left(\sqrt{(x-x_c)^2+(y-y_c)^2}, \forall (x,y)\in V\right)\right\rfloor \tag{21}$$

Inserting the values computed by (20) and (21) to (19), and replacing the entire binary image to an object O_m, the content adaptive closing by reconstruction is defined as $\varphi^{Nr}(\delta^{Ns}(O_m),O_m)$. Fig. 9 shows the effect on the content adaptive closing by reconstruction. By comparing the boundary of the smooth object (Fig. 9c) with the boundary of the precise object (Fig. 9d) by sophisticated segmentation algorithm [31], the precise object usually involves longer boundary. Accordingly, more bits would be required for coding such precise object.

sMBs, *dMBs*, and *pMBs* reconstruct a pure background image (*PB*). The covered pixels in *p*MBs are referred to estimated background *EB* as

$$PB_{x,y}=\begin{cases} I_{x,y}, & \text{if } (x,y)\in sMB \text{ or } dMB \\ EB_{x,y}, & \text{if } (x,y)\in pMB \end{cases} \tag{22}$$

the *PB* now can be considered as an entire frame instead of an object with arbitrary shape so that the shape coding can be removed. The advantages of this improvement are twofold: 1) the background is still encoded individually to originate its own bitstream. 2) the bits related to shape coding are saved in comparison with the single-closed-loop approaches where the bits related to shape coding are still required. Although the covered pixels related to *pMBs* are additionally encoded, the required bits for motion-compensated coding of these pixels are much fewer than those for the arbitrary shape coding.

In this loop, the inter prediction by motion search is simplified as the inter prediction by direct difference. The motion compensation is simplified by adding reconstructed data with the data stored in scene buffer. The simplification of motion search and compensation is significant because the amount of visual data associated with background are much more than the amount of visual data associated with foreground in the surveillance video sequence. Also note that, a zero vector has to be assigned in this loop for the encoding of motion vector.

Context of MBs are coded according to the Group of Picture (GOP) structure in which an I-picture is normally followed by several P-pictures and B-pictures. As to the encoding of I-picture in GOP, all *sMBs*, *dMBs*, *pMBs*, and *oMB*s are intraframe coded. As to the encoding of P-picture and B-picture, *dMBs*, *pMBs*, and *oMBs* are interframe coded by motion search or direct difference according to the coding loop. The *sMB*

Fig. 11(a) Hall monitor (QCIF) coding results at 62Kbps.

Fig. 11(b) PETS2006 (D1) coding results at 630Kbps.

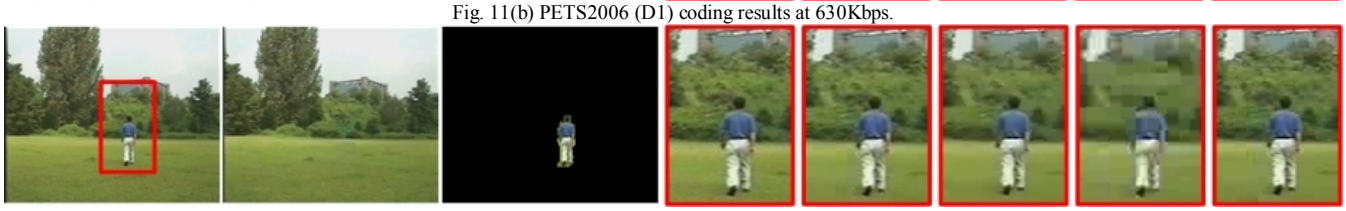


Fig. 11(c) ETRI B (CIF) coding results at 103Kbps.

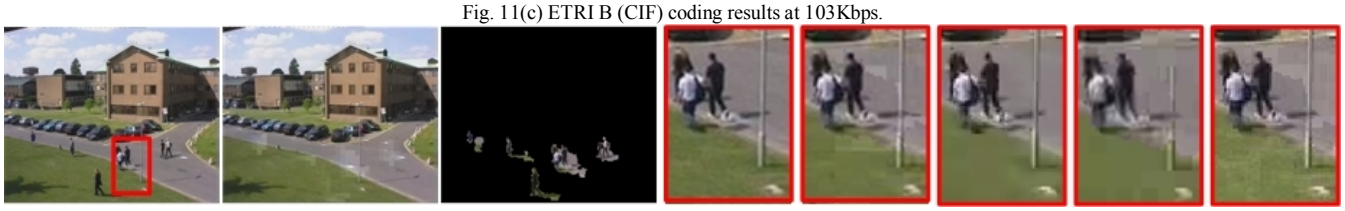


Fig. 11(d) PETS2001 (CIF) coding results at 108Kbps.

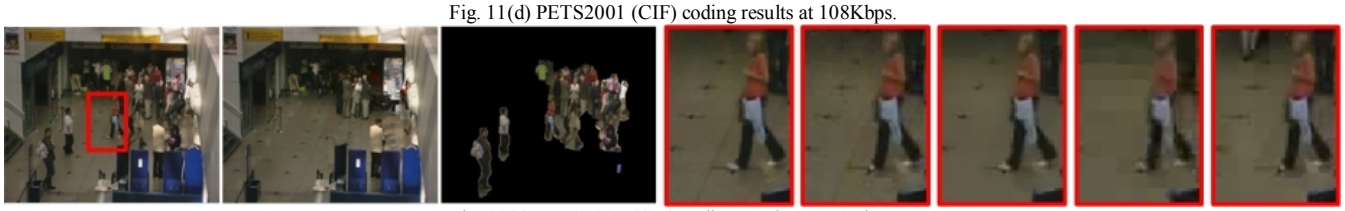


Fig. 11(e) PETS2007 (CIF) coding results at 250Kbps.

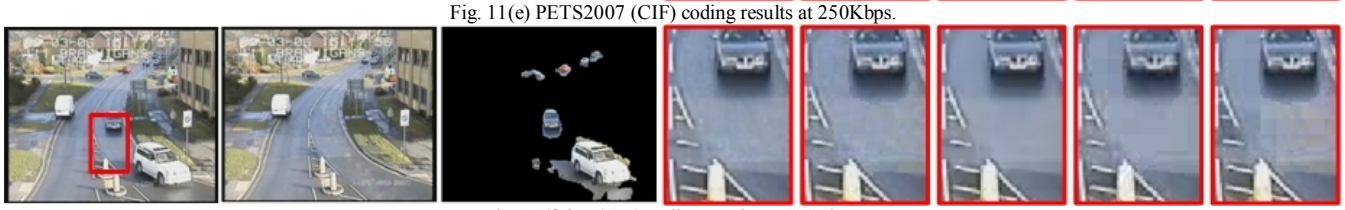


Fig. 11(f) i-L (CIF) coding results at 420Kbps.

Fig. 11. Coding results and corresponding detailed parts at similar bitrates. First column shows the original frame. Background and object coding results of this framework are shown in second and third column. The magnified details of original frame and the encoding results of this work, H.264/AVC, MPEG-4 object coding, and [24] are shown in fourth, fifth, sixth, seventh, and eighth column respectively.

always skips interframe coding until it has been classified as *dMB* or *pMB*. In this situation, the background coding loop duplicates MB in previous reconstructed image stored in scene buffer and normally skips encoding. The skipping *sMB* is based on two reasons. First, the inter prediction error of *sMB* would be not visually discernible so that the transmitted DCT coefficients of 8×8 blocks are trivial. Second, the amount of *sMB* generally dominates the amount of MB for the background coding loop in a 24/7 video surveillance.

IV. Experimental Results

The performance of the proposed dual-closed-loop coding is comparing to other object-based coding approaches, including MPEG-4 object coding [11], the single-closed-loop approach: Babu and Makur [18], and the two-coding-loop approach: Cavallaro et al. [24]. Besides, the comparison with state-of-the-art frame-based coding approach, H.264/AVC [30], is conducted.

For all of these approaches, fix quantization parameter scheme is applied through the rate distortion evaluation, the length of GOP is set as 16 with IPBBPBB structure, full search is applied to motion estimation and the search range is given as 16. Additional parameters setup for the proposed work is given as: Np=100, TH_D=10, and β=5×10^{-4}. The number of Gaussian K is selected as four in consideration that foreground pixel, static background pixel, dynamic background pixel and

TABLE III
Averaged △BD-bitrate and averaged △BD-PSNR compared to different encoding approaches by Bjontegaard metric.

	H264/AVC		Mpeg-4		Babu and Makur [18]		Cavallaro et al. [24]	
	△PSNR (dB)	△bitrate (%)	△PSNR (dB)	△bitrate (%)	△PSNR (dB)	△bitrate (%)	△PSNR (dB)	△bitrate (%)
Hall monitor	-1.859	39.495	8.5485	-50.926	8.5128	-51.205	0.47816	-9.3507
ETRI B	-0.37876	13.374	7.4736	-56.281	6.2659	-56.354	0.67741	-21.842
PETS 2001	-0.74895	25.635	6.8328	-59.336	6.1431	-57.574	0.24904	-19.817
PETS 2006	-1.7562	83.868	13.16	-53.074	11.551	-52.932	1.4048	-46.83
PETS 2007	-1.8929	56.908	7.3857	-41.23	8.0594	-44.392	2.5614	-96.68
i-L	-1.1604	30.953	5.1657	-37.952	4.5982	-36.24	1.5258	-49.878

illumination changed pixel may all present in the consecutive temporal pixels set. The mixture of Gaussian background subtraction algorithm [1], one of familiar foreground detection algorithm for video surveillance, is applied to segment objects for MPEG-4 object coding. In [24], JPEG2000 is used for background image coding at every 2 second. For the H.264/AVC coding, the number of reference frame is 3. Context-adaptive binary arithmetic coding is used for entropy coding.

Six surveillance sequences namely, *Hall monitor* (QCIF, 30Hz, 300 frames), *ETRI B* (CIF, 30Hz, 3000 frames), *PETS2001* (CIF, 30Hz, 5000 frames), *PETS2006* (D1, 30Hz, 3000 frames), *PETS2007* (CIF, 30Hz, 2800 frames), and *i-L* (CIF, 30Hz, 4000 frames), including varied object activities and weather and illumination changes in background, are selected.

A. Rate-distortion performance comparison

Fig. 10 shows the rate distortion (R-D) of six test sequences respectively. The objective visual quality is measured by the peak signal-to-noise-ratio (PSNR) of the luminance component. Compared with the single-closed-loop approaches: MPEG-4 object coding and [18], the proposed approach is improved with an additional 2.8-6 dB gain in *Hall monitor*, 2-4.6 dB gain in *PETS2006*, and 1-4.5 dB gain in *PETS2007* in indoor type sequences. In the outdoor type sequences, 2.6-5.6 dB, 3.1-6 dB and 1-4 dB gains are improved in *ETRI B*, *PETS2001*, and *i-L* respectively. These results are equivalent to 40-60% bitrate saving. This is because the improved dual-closed-loop approach can efficiently exploit the redundancy related to foreground and background than the single-closed-loop based approaches.

In comparison with two coding loop based approach in [24], the MB context determination with scene analysis can provide an additional 0-1 dB, 0-1.5 dB, and 0-2.8 dB gains in *ETRI B*, *PETS2001* and *i-L* respectively, as well as 0.2-0.4 dB, 0.8-2 dB and 0-4.3 dB gains in *Hall monitor*, *PETS2006* and *PETS2007* respectively. A noticeable quality improvement is observed in the results associated with higher bitrates, especially with bitrates higher than 200Kbps in *i-L* and *PETS2007,* and with bitrates higher than 400Kbps in *PETS2006*. The background encoding by JPEG2000 at every specified time interval will lose quality when background variation is occurred such as illumination is changed or a still object begins moving such as a parked car begins moving. On the other hand, the classified context distinct MB can improve the coding performance in the scene with background variation.

(a) (b)

Fig. 12 Different situations on coding error cause PSNR degradation by this work and [24] respectively.

In comparison with H.264/AVC, MPEG-4 object coding results in 3.8-6.6 dB loss in indoor type sequences and 3.1-6.1 dB loss in outdoor type sequence. The proposed dual-close-loop approach significantly alleviates this gap by exploiting contextual redundancies. The performance gap between the dual-close-loop approach and H.264/AVC is reduced to 0-2.6 dB in outdoor sequences and 1.4-3.3 dB in indoor sequences. Moreover, the Bjonteggard method is used to measure the average bitrate difference (Δbitrate) and the average PSNR difference (ΔPSNR) [32] between the proposed and other approaches. These results are given in Table III. In comparison with single-closed-loop object coding approach, 49% average bitrate is saving and 7.1dB average PSNR is improved. 40% average bitrate is saving and 1.1dB average PSNR is improved as comparing to the two-coding-loop approach. As comparing to H.264/AVC, 41% average bitrate is increased and 1.2dB average PSNR is decreased.

Fig. 11 shows the coding results at similar bitrates. First column shows the original frame. Background and object coding results of this framework are shown in second and third column. The magnified details of original frame and the encoding results of this work, H.264/AVC, MPEG-4 object coding, and [24] are shown in fourth, fifth, sixth, seventh, and eighth column respectively. In sub-images in Fig. 11(b)-(e), several textural features are visible by the dual-closed-loop approach and [24], while they are vague by other works event state-of-the-art, H.264/AVC. There may raise a question why the R-D performance of H.264/AVC is better than that of dual-closed-loop approach and [24]. This is due to the contain error in *pMB* in which few temporally stationary foreground pixels are included in the *pMB*. An example can be observed from Fig. 12(a), in which errors in *pMB*s are came from two previous frames. Encoding such *pMB*s can suffer seriously

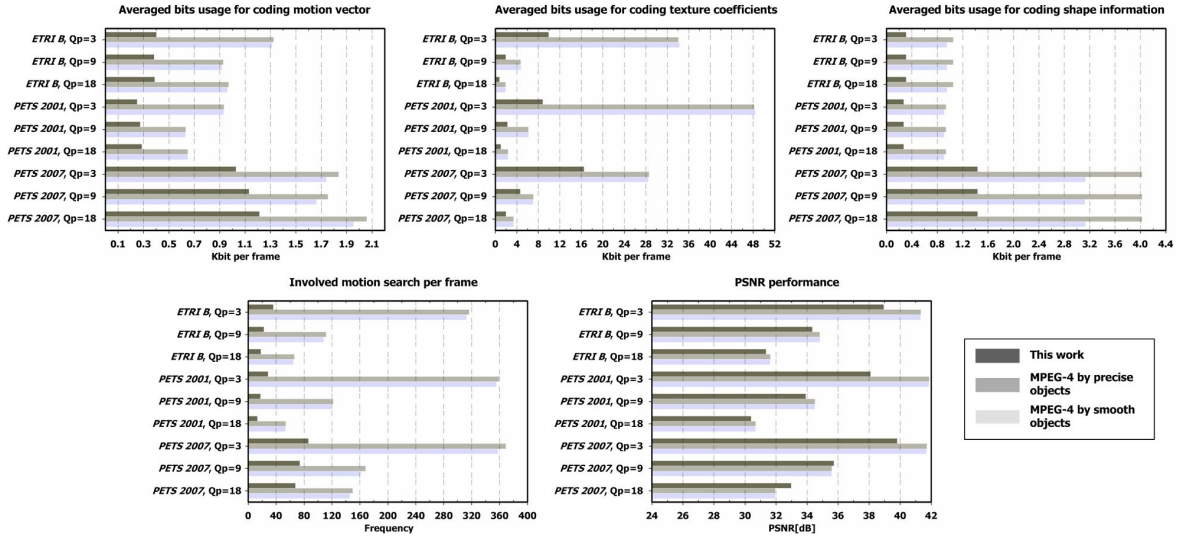


Fig. 13 Evaluation of bits usage for coding motion vector, texture coefficients, and shape information, and the involved motion search per frame, and the PSNR performance against different QP value.

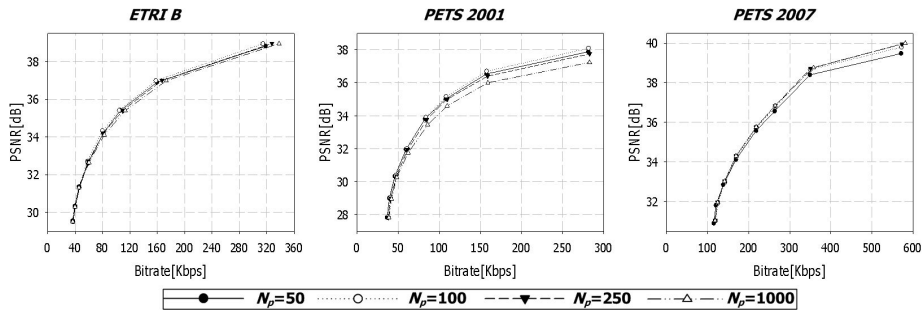


Fig. 14 Rate-distortion performance against different N_p value.

PSNR degradation. Since this kind of error almost comes from temporally stationary foreground, it therefore reflects that higher PSNR gaps appear in sequence with many temporally stationary foreground objects, e.g. *PETS2006*, *PETS2007*, and *hall monitor*. Different from aforementioned reason, the PSNR degradation of [24] is due to that entire background region is not up-to-date. This effect can be observed from Fig. 12(b) where the time information is overdue. Therefore, [24] suffers serious PSNR degradation when illumination changes, background variations, and temporally stationary foreground are presented in the scene.

B. Evaluation the advantage of exploiting contextual redundancy for object-based coding

Effects on exploiting contextual redundancy are evaluated against scene involved illumination changes, background variations, or temporally stationary foreground objects. The *PETS 2001*, *ETRI B*, and *PETS 2007* are selected to stand for each different case. Bits usages for coding motion information, texture coefficients, and shape information are evaluated. To also evaluate how the precision of object affect the coding performance, the smooth objects segmented by [1] and precise objects segmented by [31] are both applied for MPEG-4.

By the evaluation results shown in Fig. 13, several conclusions are brought out. 1) In two MPEG-4 approaches, the bits usage for coding shape of precise objects is higher than the bits usage for coding shape of smooth objects, while the rest of bits usages remain to be closed to each other. This result is intuitive because the precise object typically involves longer boundary than the smooth object involved. This work introduces the *pMB* that exposes the covered background pixels. Therefore, the bits usage for coding shape of background object is not necessary. This evaluation supports the need of the *pMB* and the usage of content adaptive closing by reconstruction. 2) the gaps of bits usage for motion vector between the proposed approach and two MPEG-4 approaches are significant. Since the bits usage for coding motion vectors of foreground objects by different approach should be closed with the same sequence, such significant gap must be due to the coding motion vectors of background. That is, the background region also generates motion vector even with the static camera and involves motion search. Since our approach only codes zero vector for the *dMB*, the bits saving is therefore significant in comparison with two MPEG-4 approaches. 3) In the evaluations of the PSNR performance and the bits usage for coding texture coefficients, a considerable amount of bits saving for texture coefficients are observed. This implies that the context of MB indeed conducts useful contextual redundancies. The cost on exploiting such redundancies is slight PSNR degradation at low quantization parameter (QP) and middle QP. At high QP, about 2dB

TABLE IV
Evaluation of context distinct MB classification accuracy.

MB classification accuracy	Simple background subtraction method [1]		Sophisticated background subtraction method[31]		This work			
Sequence	BG MB	FG MB	BG MB	FG MB	BG MB			FG MB
					dMB	*sMB*	*pMB*	*oMB*
Hall monitor	1.4%	8.7%	0%	0%	0%	1.3%	5.2%	5.2%
ETRI B	0.6%	9.2%	0.1%	0.7%	0.6%	0.2%	1.5%	1.5%
PETS 2001	0.8%	13%	0.1%	1%	6.9%	0.4%	5.6%	5.6%
PETS 2006	2.1%	8.8%	0.4%	1%	0%	0.8%	1.9%	1.9%
PETS 2007	3.5%	7.5%	1.3%	2.9%	0.2%	3.6%	8.3%	8.3%
i-L	2.2%	11.5%	0.4%	1.5%	7.4%	1.9%	7.5%	7.5%
Averaged error rate	1.8%	9.8%	0.4%	1.2%	8.9% [a]			5%

[a] Averaged error rate of BG MB is the average value of summing the error rates of *dMB*, *sMB*, and *pMB*.

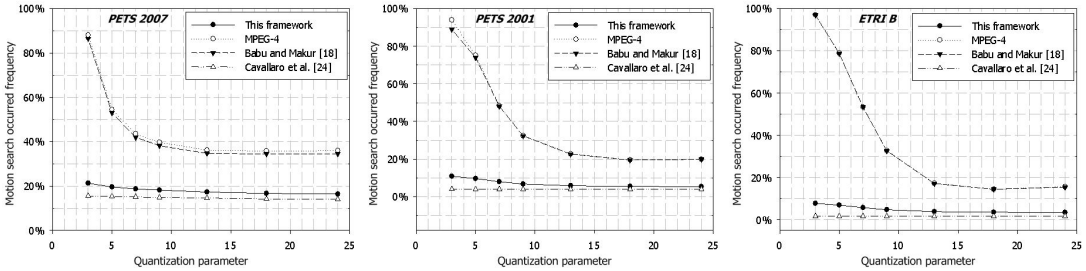


Fig. 15 Motion search occurred frequency per frame against different quantization parameters.

degradation and 4dB degradation are observed for sequence without illumination change and sequences with illumination change respectively. Also, at high QP, the bits saving of this work achieves the maximum.

Since the proposed approach essentially evaluates the context of MB and encodes in operational rate-distortion-optimized scene, the MB classification accuracy against coding performance is discussed. Table IV evaluates the MB classification accuracy by manual classifying MB in several important frames in which a considerable amount of contexts are appeared. Here the error rate is denoted as

$$\text{Error rate}=\frac{\text{number of error classified MB}}{\text{number of correct classified MB}} \quad (23)$$

In general, precise objects can improve the precision on classifying foreground MB and background MB. The erroneous classification of MB will decrease the R-D performance because the motion of that MB is not continuous and predictable and likely to be intraframe coded. However, the resulted PSNR improvement is very limited in fact. Examples are PSNR performances of MPEG-4 by precise objects and MPEG-4 by smooth object shown in Fig. 13. Furthermore, the goal of background subtraction (e.g. [1] and [31]) is to ignore background motion while to detect foreground motion. Therefore, higher MB classification accuracy does not really mean lower motion search frequency because there still are variations in background. In contrast, the MB classification accuracy is significant to this work. The erroneous classification of MB will lead to fewer contextual redundancies been exploited. The averaged error rates for both foreground MB and background MB are below 9 % for all sequences. These low error rates can support good R-D performance.

Moreover, the R-D performance against the number of consecutive temporal pixels *Np* is evaluated because *Np* is the most important parameter related to context information. The R-D performances against different *Np* are shown in Fig. 14. The higher *Np* suits to the scene with temporally stationary foreground objects but not suit to the illumination change and background variation. In contrast, the opposite results are observed by the lower *Np* value. To acquire good R-D performance for all context of scene, *Np* values from 100 to 250 are suggested because the performance gap between these two values is the smallest.

C. System level complexity analysis

Furthermore, we analyze the system level complexity of object-based approaches in term of the amount of MB involved with motion search per frame. Since the motion search dominates more than 60% of computational complexity in encoder, to reduce the number of involved motion search is then equivalent to reduce the overall computational burden. Fig. 15 shows the motion search occurred frequency per frame in percentage. Generally, lower QP will increase the number of MB involved with motion search in a general motion-compensated video coder (see MPEG-4 and [18] in Fig. 15) except [24] where background is always intraframe coded. The proposed approach can reduce 30% of motion search occurred frequency in average compared to MPEG-4 and [18], and as high as 60% of motion search occurred frequency reduction in the lowest QP. This is because the context of MB can be accurately classified with error rates of all types of MB below 9%, and therefore dual-closed-loop approach can encode context distinct MBs in an operational rate-distortion-optimized scene.

Table V further evaluates the relationship between the gain on frequency of motion search reduction and the gain on

TABLE V
Complexity analysis on motion search and execution time.

Test condition: 4000 frames (1,620,000 MBs), CIF format, QP: 22 for H.264 and 3 for rest			
Framework	#MB involved motion search	ME time	Total time
H.264	1,143,103	13432 ksec	20761 ksec
MPEG-4	1,357,240	1502 ksec	1768 ksec
[18]	1,326,485	1380 ksec	1624 ksec
[24]	153,025	439 ksec	628 ksec
This framework	296,304	630 ksec	840 ksec

TABLE VI
Complexity analysis on the dual-closed-loop approach.

Functionality	Time (ksec)	Percentage
Scene analysis	140	13.42
Connected component analysis	26	2.49
Context determination	38	3.64
Encoding	840	80.45
Total	1044	100

encoding time. This evaluation is run on a Pentium 2.8G PC. Unsurprisingly, H.264/AVC spends the most of encoding time on motion estimation (ME) despite the amount of MB involved motion search is fewer than MPEG-4 and [18]. In H.264/AVC, a MB involved motion search is divided into blocks of smaller sizes, and motion vector of a block is estimated by variable block-size ME technique with rate-distortion optimized coding mode decision. Therefore, it is necessary for H.264/AVC to perform ME much more times than MPEG-4 and [18] to code a MB. This result is similar to the million instructions per second (MIPS) profiling analysis reported in [33]. It states that encoding complexity of H.264/AVC baseline profile is about ten times more complex than MPEG-4 simple profile [34]. In contrast, this framework exploits the contextual redundancy to code a MB instead of exhaustively searching interframe redundancy to code a MB. Result reflects that this framework saves about 96% and 53% of encoding time in comparison with H.264/AVC and MPEG-4 respectively, and that reducing the involved motion search can lead to save the encoding time dramatically. The complexity of the proposed approach is analyzed in Table VI. The test condition is the same as that in Table V. The complexity of functionalities related to context information analysis and generation takes 19.55% of computation. Obviously it is much fewer than the complexity of motion search and reveals the low overhead in the proposed technique. This result demonstrates that the exploiting contextual redundancy can improve the coding efficiency while significantly reducing the coding complexity for the surveillance video.

V. Conclusions

Due to different design philosophies between conventional video compression and surveillance video compression, this paper proposes a low complexity object-based coding approach by exploiting contextual redundancy associated with the scene. The MB associated with distinct context is classified and encoded in an operational rate-distortion-optimized scene. Consequently the unique features of this approach are expressed as follows: 1) a scene analysis method is derived to model the context information in a surveillance scene. 2) the redundancy associated with different context is explored and a context of MB determination method is presented for context of MB classification. 3) dual-closed-loop coding approach is presented to efficiently encode the classified context distinct MB in an operational rate-distortion-optimized sense.

Objects and background are individually encoded and originate its own bitstream in our approach. Experiments have been conducted on outdoor and indoor real surveillance sequences. These results show that dual-closed-loop coding can achieve higher coding efficiency than the other object-based coding approaches and improve averaged 2dB performance gain compared to related state-of-the-art approaches. The system level complexity is evaluated in term of the number of MB involved motion search. The proposed approach can achieve 30% of motion search occurred frequency per frame compared to others object-based coding frameworks. This result is owing to the high accuracy in context MB classification at an error rate below 9%. The well coding performance demonstrates a particularly suitable approach for intelligent VSN surveillance.

References

[1] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 246–252.

[2] A. Yilmaz, O. Javed and M. Shah, "Object Tracking: A Survey," *ACM Journal of Computing Surveys*, vol. 38, No. 4, 2006.

[3] D.-Y. Chen, *et al*., "Video-based human movement analysis and its application to surveillance systems," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 372–384, Apr. 2008.

[4] F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A Survey on Wireless Multimedia Sensor Networks," *Elsevier Comp. Net.*, vol. 51, pp. 921–60, Mar. 2007.

[5] M. Valera and S. Velastin, "Intelligent distributed surveillance systems: A review," in *IEEE Proc. Vis. Image, Signal Process.*, vol. 152, no. 2, pp. 192–204, Apr. 2005.

[6] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334–352, Aug. 2004.

[7] M. M. Trivedi, T. L. Gandhi, and K. S. Huang, "Distributed interactive video arrays for event capture and enhanced situational awareness," *IEEE Intell. Syst.*, Oct. 2005.

[8] M. Quaritsch, M. Kreuzthaler, B. Rinner, H. Bischof, and B. Strobl, "Autonomous multicamera tracking on embedded smart cameras," *EURASIP J. Embed. Syst.*, 2007.

[9] S. Fleck and W. Straber, "Smart camera based monitoring system and its application to assisted living," in *Proc. IEEE*, vol. 96, no. 10, pp. 1698–1714, Oct. 2008.

[10] Y. Charfi, N. Wakamiya, and M. Murata,"Challenging issues in visual sensor networks," *IEEE Wireless Communications. Mag.*, vol. 16, no. 2, pp. 44–49, Apr. 2009.

[11] *Coding of Audiovisual Objects-Part 2: Visual*, ISO/IEC ISO/IEC 14496-2 (MPEG-4), 2001.

[12] M.-H. Hsiao, *et al*., "Object-based video streaming technique with application to intelligent transportation systems," in *Proc. IEEE Int. Conf. Networking, Sensing and Control*, pp. 315–318, Mar. 2004.

[13] J. Meessen, C. Parisot, X. Desurmont, and J.-F. Delaigle, "Scene Analysis for Reducing Motion JPEG 2000 Video Surveillance Delivery Bandwidth and Complexity," in *Proc. of IEEE Int. Conf. Image Process.*, Genova, Italy, September 2005.

[14] W. K.-H. Ho, W.-K. Cheuk, and D. P.-K. Lun, "Content-based scalable H.263 video coding for road traffic monitoring," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 615–623, Aug. 2005.

[15] A. Vetro, T. Haga, K. Sumi, and H. Sun, "Object-based coding for long-term archive of surveillance video," in *Proc. IEEE Int. Conf. Multimedia and Expo*, vol. 2, pp. 417–420, July. 2003.
[16] F. Moschetti, G. Covitto, F. Ziliani, and A. Mecocci, "Automatic object extraction and dynamic bitrate allocation for second generation video coding," in *Proc. IEEE Int. Conf. Multimedia and Expo*, vol. 1, pp. 493–496, July. 2002.
[17] Y. Yu and D. Doermann, "Model of object-based coding for surveillance video," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 693–696, Mar. 2005.
[18] R. V. Babu and A. Makur, "Object-based surveillance video compression using foreground motion compensation," in *Proc. IEEE Int. Conf. Control, Automation, Robotics and Vision,* pp. 1–6, Dec. 2006.
[19] T. Aach, A. Kaup, "Baysian algorithm for change detection in image sequences using Markov random fields," *Sig. Proc: Im. Comm*. 7(2):147-160, 1995.
[20] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 827-832, May. 2005.
[21] M.A. Sato and S. Ishii, "Online EM algorithm for the normalized Gaussian network," *Neural Computation*, vol. 12, pp. 407-432, 1999.
[22] T. Sikora and B. Makai, "Shape-adaptive DCT for generic coding of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 59–62, Feb. 1995.
[23] L. P. Kondi, G. Melnikov, and A. K. Katsaggelos, "Joint optimal object shape estimation and encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, pp. 528–533, Apr. 2004.
[24] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 1200–1209, Oct. 2005.
[25] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.
[26] K. J. Kim, C. W. Lim, M. G. Kang, and K. T. Park, "Adaptive approximation bounds for vertex based contour encoding," *IEEE Trans. Image Processing*, vol. 8, pp. 1142–1147, Aug. 1999.
[27] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," *Learning in graphical models*, MIT Press, Cambridge, MA, 1999.
[28] J. Serra and L. Vincent, "An overview of morphological filtering," *IEEE Trans. Circuits, Systems and Signal Proc.,* vol. 11, no. 1, pp. 47–108, Apr. 1993.
[29] P. Salembier and M. Pardas, "Hierarchical morphological segmentation for image sequence coding," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 639–651, Sep. 1994.
[30] *Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification*, ITU-T Rec. H.264 and ISO/IEC 14 496-10 AVC, Joint Video Team, Mar. 2003.
[31] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 1301–1306.
[32] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," ITU-T Q6/SG16, Doc. VCEG-M33, Apr. 2001.
[33] T.-C. Chen, *et al*., "Analysis and architecture design of an HDTV720p 30 frames/s H.264/AVC encoder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, pp. 673–688, Jun. 2006
[34] H.-C. Chang, *et al*., "Performance analysis and architecture evaluation of MPEG-4 video codec system," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 2, pp. 449–452, May 2000.

Tsung-Han Tsai received the B.S., M.S., and Ph.D. degrees in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1990, 1994, and 1998 respectively. From 1999 to 2000, he was a professor of electronic engineering at Fu Jen University. In 2000, he joined the Department of Electrical Engineering, National Central University, Taiwan, where he is currently a Professor. He has been an IEEE member for over 10 years, and is also a member of the Audio Engineering Society (AES) and the Institute of Electronics, Information and Communication Engineers (IEICE). Dr. Tsai has been awarded 14 patents and has published more than 120 referred papers in international journals and conferences. His research interests include VLSI signal processing, video/audio coding algorithms, DSP architecture design, wireless communication, and System-On-Chip design. Dr. Tsai received the Industrial Cooperation Award in 2003 from the Ministry of Education, Taiwan. He is a member of the Technical Committee of IEEE Circuits and Systems Society, and serves as Technical Program Committee member or Session Chair of several international conferences.

Chung-Yuan Lin received the M.S. degree from the Department of Electrical Engineering, National Central University (NCU), Taoyuan, Taiwan, R.O.C., in 2003. He is currently pursuing the Ph.D. degree at NCU. His research interests include computer vision problem on surveillance system, video processing algorithm design, algorithm optimization for DSP architecture, and embedded system design.