Dynamic Background Segmentation for Remote Reference Image Updating within Motion Detection JPEG2000

Théodore Totozafiny
Computer Science Lab LIUPPA
University of Pau
IUT, Chatêau Neuf
64100 Bayonne, France
t.totozafiny@estia.fr

Olivier Patrouix
Laboratory for Industrial
Process and Services
ESTIA, Technopole Izarbel
64210 Bidart, France
o.patrouix@estia.fr

Franck Luthon
Computer Science Lab LIUPPA
University of Pau
IUT, Chatêau Neuf
64100 Bayonne, France
Franck.Luthon@univ-pau.fr

Jean-Marc Coutellier
Magys
Technopole Izarbel
64210 Bidart, France
jm.coutellier@magsys.net

***Abstract*— We present in this paper a new system based on Motion JPEG2000 intended for road surveillance application. The system uses a reference image and consists in 4 processing steps, namely initialization phase where the first reference image is built, reference estimation, motion segmentation (foreground extraction, ROI mask), and JPEG2000 coding. A first order recursive filter is used to build a reference image that corresponds to the background image. The obtained background is sent to the decoder once for all. The reference image at the coder side is estimated according to a Gaussian mixture model. The remote reference image is updated when specific conditions are met. The updating remote reference is triggered according to the states of mobile objects in the scene (no, few or lot of mobiles). The motion detection given by classical background subtraction technique is performed in order to extract a binary mask. The motion mask gives the region of interest of the system. The JPEG2000 image coded with a ROI option is sent towards the decoder. The decoder receives, decodes the image and builds the implicit binary ROI mask. Then, the decoder builds the displayed image using the reference image, the current image and the mask.**

I. Introduction and Previous Work

Our study addresses the problem of road surveillance for safety purposes. Recently [1] developed an integrated system for smart encoding in video surveillance. A part of the system is based on Motion JPEG2000 (Part3 of the JPEG2000 standard) in which the mobile objects in the scene are extracted automatically with a background subtraction method. The reference image is estimated from each frame with a mixtures of Gaussian technique. The segmentation results are used as input region of interest masks for the JPEG2000 encoding of each frame. The frame is coded with ROI option of JPEG2000. The Motion JPEG2000 codes each frame independently and consists in a concatenation of JPEG2000 images. One of the requirements in JPEG2000 is the support of ROI coding, where ROI of the image can be coded with better quality than the background. The JPEG2000 image encoding standard defines two kinds of ROI coding techniques: the general scaling based method and the Maxshift method. Several methods [2] were proposed to improve the ROI coding. With the Maxshift method, no extra information about the shape of the mask is required for the decoder. The authors in [1] combined the two above techniques to encode the ROI mask. Their system allows in real-time capturing an image, performing a motion detection, encoding with JPEG2000. The data is stored in a server. Then, the data will be transmitted toward a decoder in which a final image is reconstructed. The transmission is made with two layers. The first layer contains the data, the second contains the reference image. But the authors do not explain how to obtain initially the reference. Indeed, in the mixtures of Gaussian technique, some images are necessary to obtain a correct reference image. In order to obtain better results for reconstructed image at the decoder, an initialization phase should be done.

[3] gives a definition of the initialization of background and proposed a solution to build initially a reference image.

[4] proposed a scheme for region-based channel adaptive source coding scheme. The scheme proposed is compared with JPEG2000 encoding with ROI option. In video surveillance application with a transmission channel of very low bandwidth such as GSM network, the use of the reference image in Motion JPEG2000 encoding presents tree problems: obtaining a reference, updating and transmission towards the decoder of the reference regularly.

We have developed a system based on the Motion JPEG2000 standard. In an initial stage, the system builds a reference image and transmits it towards a decoder through the GSM network. This initialization is necessary in order to reconstruct the final image during the next transmission. When the initialization is achieved, the reference image is estimated according to a mixtures of Gaussian model. The motion detection is performed with classical background subtraction technique. The motion mask gives the region of interest for the system. The current image, containing only data linked with the mask, is coded using the ROI option of JPEG2000 encoding and transmitted towards the decoder. The remote reference image is updated when the conditions are required. The updating of remote reference is triggered according to the states of mobile objects in the scene (no, few or lot of objects).

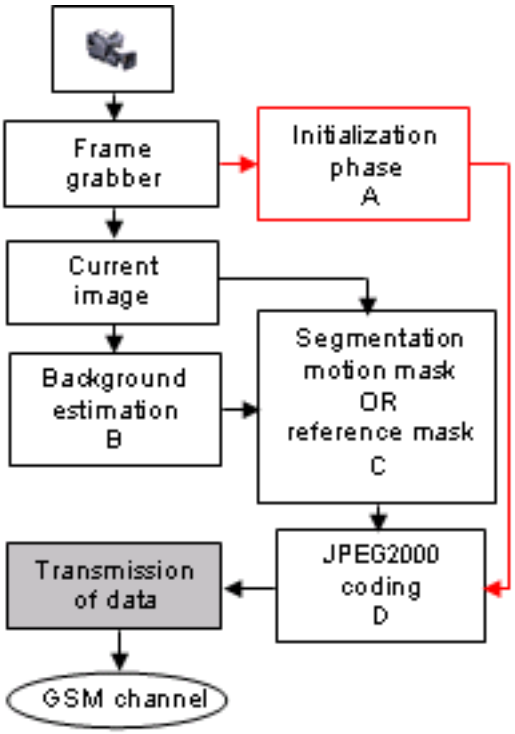

Fig. 1. Coder MRIJ2K.

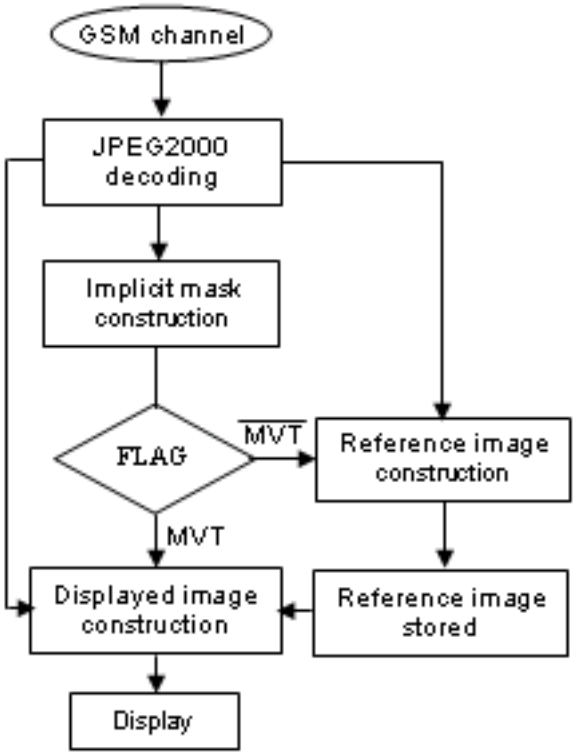

Fig. 2. Decoder MRIJ2K.

This paper is organized as follows: we describe our approach in section II and updating of the remote reference is given in section III. Finally experimental results and conclusions are given in sections IV and V.

II. Description our system

In our approach, the embedded device is the one that is directly connected to the camera and the remote one is the ground station which receives the encoded images. The embedded device is mainly based on a camera, a CPU and a wireless transmission device.

A. Coder and decoder

Fig.1 shows our coder system Motion Reference Image JPEG2000 (MRIJ2K). The system is based on four main processing blocks:

- Block A: initialization phase where the first reference image is built (i.e. the background image) and transmitted towards the decoder. This is done only one time at the starting of the system,
- Block B: reference image estimation,
- Block C: segmentation in order to extract a motion mask or reference mask,
- Block D: JPEG2000 encoding with ROI option,

Fig.2 shows our MRIJ2K decoder system. The decoder receives, decodes the image and builds the implicit ROI mask. Then, it checks (indicated as FLAG in block diagram) whether the decoded image is a motion image or an updated remote reference image. If it is an updated reference, then the image is stored as the remote reference image otherwise the image is displayed.

B. Initialization phase

The initialisation phase consists of the construction of reference image. In a static camera context, several methods in the computer vision literature were proposed to obtain a reference image. In [5], [6], [3], [7], the authors evaluate the advantages and drawbacks of some techniques. The common techniques used to extract the reference image are : median filter**,** frame differencing, Kalman filter, Gaussian mixtures, non-parametric method, wallflower, etc.

1) Constructing a reference image: Considering the constraints of the final custom application (lack of memory and embedded system working on a PC104 format board), we use the following filter transfer function form:

$$G(Z) = \frac{b}{1 + aZ^{-1}} \quad (1)$$

where a and b are real. The recursive equation of the filter is given by:

$$I_{refinit}(p, t+1) = \alpha_p I(p,t) + (1 - \alpha_p) I_{refinit}(p,t) \quad (2)$$

where $I_{refinit}(p,t)$ and $I(p,t)$ are the intensity values of pixel p in the reference image and in the current image at time t respectively. $\alpha_p \in [0,1]$ is the learning rate that gives the training speed, and p is the pixel location in the image ($b = \alpha_p$,

$a = \alpha_p - 1$). Of course when large object moves slowly in the scene, it will be included in the constructed reference image.

2) Learning rate filter: Since the reference image is the static part of the scene, the value of α_p for all pixels belonging to a moving object must be 0. In order to determine α_p we have to know all pixels that belong to the background. When p is a background pixel element, the α_p value must be in the $]0, 1]$ interval, otherwise α_p is 0. The binary stability charts of the three consecutive images $I(t-2)$, $I(t-1)$ and $I(t)$ are used in order to know the object states in the scene. To compute the binary stability chart, we use the edge information and the noise of the acquisition camera is attenuated by an average filter (3×3). The Canny edge detector is used to calculate the gradient of each frame in the sequence.

$$d_1 = |I_g(p, t-2) - I_g(p, t-1)| \quad (3)$$

$$d_2 = |I_g(p, t-1) - I_g(p, t)| \quad (4)$$

where I_g is the spatial gradient of the processed image in the sequence.

A logical AND between d_1 and d_2 followed by an entropic thresholding technique indicates whether an object is moving or not: when the condition $((d_1 < \lambda)\, AND\, (d_2 < \lambda))$ is true, the object is a background element. λ can be obtained according to an entropy power threshold selection method [9]. The model (2) gives good results when mobile objects are of small sized, in the opposite case, the quality of the regions occupied by a large object is bad into the reference image obtained. It will be useful to improve these regions first.

C. Segmentation

In computer vision literature, several techniques were proposed to perform a segmentation of mobile object in the scene [5], [6], [7]. The motion detection is a binary labelling problem. It consists in attributing to each pixel p of an image I at time t one of the two following label values:

$$e_p = \begin{cases} 1 & \text{if } p \in \text{moving object} \\ 0 & \text{if } p \in \text{static background} \end{cases} \quad (5)$$

In order to carry out the binary labelling task, we use two observations. For each pixel p at time t, we calculate:

- the difference between the reference image and the current image:
$$o_{dr}(p, t) = |I(p, t) - I_{ref}(p, t)| \quad (6)$$
- the difference between the two successive frames:
$$o_{dt} = |I(p, t) - I(p, t-1)| \quad (7)$$

To find the most probable label field with these two observations, we can use the conditional probability (Bayes theorem). [8] proposed a MAP criterion to obtain the most probable configuration. In our application, we used a logical AND in order to cope with the embedded target. We introduce a thresholding on both images o_{dr} and o_{dt} and then we compute the logical AND. So the pixel label is given by the following equation.

$$e_p = ((o_{dr} > T_f)\, AND\, (o_{dt} > T_f)) \quad (8)$$

The moving object threshold T_f is determined according to an automatic entropy based threshold selection. The gradient mathematic morphology filter is used to improve the binary mask obtained. To determine a threshold T_f, we used an entropy power for threshold selection [9]. T_f is given by:

$$T_f \approx 2^{H_{bit}} \quad (9)$$

where H_{bit} is the entropy of an information source (observations) that is classicaly defined as follows:

$$H = -\sum_{i=0}^{N=255} pi \log pi \quad (10)$$

where pi is the probability that the observation at any site takes the value i, and $\log$ denotes binary logarithm.

D. Estimation of the reference image

We used the same scheme as developed in [5] to estimate a reference image in embedded device. For each grabbed frame, the background image is updated. The method consists in the modelling of each pixel with K Gaussian distributions. $P(x_t)$ represents the probability for a pixel to have the intensity x_t at time t. This probability is estimated by:

$$P(x_t) = \sum_{j=1}^{K} \frac{\omega_j}{(2\pi)^{\frac{d}{2}} \left|\sum_j\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_t-\mu_j)^T \sum_j^{-1} (x_t-\mu_j)} \quad (11)$$

where ω_j is the weight of distribution j, μ_j is the mean of distribution j and $\sum_j$ is the covariance for the distribution j and $d = 3$. For computational reasons, the covariance matrix is assumed to be of the form $\sum_j = \sigma_j^2 I$. The first B distributions are used as a model of the background of the scene where:

$$B = \arg\min_b (\sum_{j=1}^{b} \omega_j > T) \quad (12)$$

T is the fraction of the total weight given to the background model and $3 \leq K \leq 5$. The parameters $(\omega_j, \mu_j, \sigma_j^2)$ of the matched and unmatched components are updated according a specific method.

E. JPEG2000 coding

JPEG2000 Part 1 has adopted a specific implementation of the scaling based ROI approach. This algorithm called Maxshift method consists in scaling down the background coefficients by 2^s. With the Maxshift method, no extra information about the shape of the ROI is required for the decoder [10]. This property is very significant in our scheme.

The ROI mask is automatically extracted from motion detection, and the image to be coded will only contain the

useful data (no background data). In spatial domain, the pixels belonging to the ROI mask are set to the current image pixel value and for the others, the value is set to 128. This value is chosen because it is the dynamic mean value of an image coded with JPEG2000. This choice is done in order to control the transition effects between background and region of interest when the image is decoded.

Also, now the current image is compressed using a ROI option with a very low bit rate which will give a rate of 1 img/s with GSM channel (size of data 9600 bits around).Then, the compressed current image is transmitted towards the decoder.

F. Displayed image construction

The decoder receives an image. This image could be a reference image (i.e. background image) or a motion image (i.e. motion data). A *flag* is needed to identify those two cases and it is added in the file header during the image encoding. In the reference case, the received image is stored as background image on the decoder, otherwise the image is only displayed.

When a motion image or a reference image is decoded, the ROI binary mask is implicitly obtained. Due to the Maxshift technique property, the decision of whether or not a wavelet coefficient c belongs to the background (i.e. $\overline{ROI}$) is stated using the following conditions:

$$\begin{cases} c \in ROI \text{ if } c \geq dwt \\ c \in \overline{ROI} \text{ else} \end{cases} \tag{13}$$

where dwt is a downshift parameter and dwt is set to 2^s.

The image to be displayed is built using the reference image, the motion image and the mask reconstructed. In spatial domain, a simple pixel substitution is used to build the displayed image.

III. Updating of the remote reference

Once the initialization phase is finished, the coder is able to send to the decoder the entire background image. Thus, the decoder can construct a final image : moving mask over the background. At the coder side, the reference image is updated for each frame according to a Gaussian Mixture Model (GMM). The updating of the reference image at the decoder should be done regularly.

The intuitive technique for updating the remote reference image (RRI) is based on the transmission of the RRI towards the decoder with a specific period. For example every 4 seconds or 10 images like in [1]. The RRI must be coded with no ROI option and with a moderate bit rate compression. In our system, the transmission of the complete RRI slows down the transmission rate.

In our scheme, the image transmission rate should be done at 1 img/s. Since we only have at our disposal one single GSM transmission channel, the complete RRI update takes more than 4 seconds. This latency is not acceptable. To solve this problem and in order to keep the image transmission rate close to 1 img/s, we propose a new technique in order to update the RRI by pieces. In this scheme, the RRI will be coded like the motion image with a ROI mask. We have chosen a square pattern in order to build the ROI mask and the information in this area will be coded and sent to the decoder for the updating of the background image.

A. Regions to be updated

The background image is composed of Nb blocks and each block will be used as an ROI mask. In order to improve the efficiency of our strategy, we define a parameter qc for each area. The qc represents a quality coefficient and $\overline{qc}$ non quality coefficient with respect to the initialization phase (background image creation) or the GMM background estimation. qc_i and $\overline{qc}_i$ ($0 \leq qc_i, \overline{qc}_i \leq 1$) are used to compute a refreshing priority for the region i. The region with the highest priority will be updated first.

1) Local quality coefficient calculation: For a given frame, the local quality coefficient for a given block is based on the intersection of the motion ROI mask and the block area. The occupation percentage of mobile object is computed as follows. Lets $Imask$, Ir be the mask of motion object and current reference image respectively and i be the current index of treated block. We can write:

$$Ir = \bigcup_{i=0}^{Nb-1} Irb_i \tag{14}$$

$$Imask = \bigcup_{i=0}^{Nb-1} Imaskb_i \tag{15}$$

where Irb , $Imaskb$ are sub-images of Ir and $Imask$ respectively. For each block i, we compute the local quality coefficient lqc using the following equations :

$$\overline{lqc}_i = \frac{NonZero(Imaskb_i)}{Nbsize^2} \tag{16}$$

$$lqc_i = 1 - \overline{lqc}_i \tag{17}$$

where the $NonZero$ is an operator that counts the number of pixels set to 1 in the current $Imaskb$ and $Nbsize$ is the size of the square block.

2) Global quality coefficient calculation: At frame k, the quality coefficients are updated according to the local quality coefficients and the previous values of the quality coefficient at frame $k-1$. A threshold T_{qc} is introduced to determine a changing of given region:

$$\overline{qc}_i(k) = \begin{cases} MAX(\overline{qc}_i(k-1), \overline{lqc}_i(k)) \text{ if } \overline{lqc}_i(k) \geq T_{qc} \\ (1-\gamma)\overline{qc}_i(k-1) \text{ else} \end{cases} \tag{18}$$

where γ is the learning rate of the background estimation (ex. during the initialization phase, model (2) $\gamma = \alpha_p$).

B. Remote reference updating decision

The region to be updated in remote reference must correspond to the block with the highest $\overline{qc}$. We introduce two thresholding values T_{high} and T_{down} for observing the occupation of the mobile objects in the scene. The high threshold T_{high} indicates that a lot of mobile objects is observed in the scene and T_{down} indicates that few mobile objects are observed in the scene. The updating of the remote reference strategy is run according to the state of object mobile in the scene (much or less). Three cases are considered:

1) case 1: $T_{state} < T_{down}$,
2) case 2: $T_{state} > T_{high}$,
3) case 3: $T_{down} < T_{state} < T_{high}$.

where T_{state} is the global percent state of mobile objects together and can be obtained by:

$$T_{state} = \frac{NonZero(Imask)}{dim} \tag{19}$$

where $Imask$ is a mask of moving object and dim represents the size of the image.

C. Pulling of blocks

The blocks to be updated are chosen according to the three cases considered above.

In the first case, we can consider that no interesting object is present in the scene. We propose to use the block with the highest $\overline{qc}$ as the ROI mask for the JPEG2000 encoding. Then, the non-quality coefficient of this block is forced to zero ($\overline{qc} = 0$, $qc = 1$) in order to flag it as "processed".

In the second case, too many moving objects are detected in the scene, then the updating of the remote reference should not be triggered. But if this configuration lasts a too long time (for example more than 2 hours), maybe the updating should be performed like in the previous case.

In the third case, there is no easy decision so we choose to force the updating every n frames in order to improve the global quality of the displayed image. In order to limit the background update, we define a threshold T_{qc} and only the blocks with a $\overline{qc}$ greater than T_{qc} will be updated. Then, the non-quality coefficient of this block is forced to zero ($\overline{qc} = 0$, $qc = 1$) in order to flag it as "processed".

IV. Experimental results

For the JPEG2000 encoding, we used the Kakadu [12] codec SDK. Our goal is to obtain a 9600 bits image file in order to reach the image rate of 1 img/s with the GSM channel. At the ground station, the image reconstruction is done using the background image and the current JPEG2000 image. While decompressing the current image, we are able to retrieve the ROI mask which contains the spatial information (Maxshift method property). In spatial domain, the current finale image is computed by copying the background image and replacing the background pixel by the current pixel if it belongs to the ROI. For experimental tests, we are using 2 PCs. The transmission tests are done using a serial port, RS232 restricted to 9600 bps). The values of parameters are given in Tab.I.

$Nbsize$	T_{down}	T_{high}	T_{qc}	α_p	n
32	10%	70%	15%	0.1	10

TABLE I
VALUES OF THE PARAMETERS.

In order to evaluate the quality of final image obtained, we have used an objective classical criterion: PSNR. For each reconstructed image, the PSNR is calculated, yielding an average value of $32dB$. Our experimental results are shown on Fig.3, Fig.4 and Fig.5. The PSNR results are shown on Fig.6. During of the initialization phase, the training frame number to extract the background image is set initially to 50. Our tests have been done with recorded video sequences of typical highway scenes[1]. In order to check the quality of the background reference image, a frame without any moving object is grabbed manually in the sequence and the PSNR is evaluated. The value is $30dB$ which leads to an average quality level. The quality is linked to the number of frames and also to the tuning of the first order parameters according to typical speed and size of the moving objects in the scene. As shown in left of Fig.3, the quality of the regions occupied by large moving object is bad (a lapse of memory). Those regions having high values in the boxes of Fig.3 bottom, should be updated first.

Then, the reference image is compressed to a JPEG2000 file at a moderate bitrate of 0.4 bpp, and sent once for all towards the decoder.

In Fig.5, tree images are lost with the use of intuitive technique (here every 5 images) to update the remote reference image, whereas only one image is lost with our method. We have run tests with potential customers of our product and their global point of view is that our strategy gives a more fluid image flow than the intuitive method.

V. Conclusion

A video coding approach with Motion JPEG2000 using a reference image has been developed in the context of road surveillance. The complete scheme is implemented and it reaches the expected performances. We also showed how the image reference is built during initialization phase. The classical background subtraction technique is used to perform the segmentation of mobile objects. Instead of updating the remote reference with a specific period, we presented a new technique to update the remote background image by pieces. The updating of the remote reference is triggered when some specific conditions are met, depending on the amount of moving areas. The implementation of our system runs at 4 frames per second on a 1.6 GHz AMD processor for 320x240 color images. The transmission, using a restricted serial port, has been tested and the rate of 1 img/s has been reached. These

[1]Thanks to LAPS laboratory (http://www.u-bordeaux1.fr) for providing an image sequence; We have also tested our algorithm with the sequences available at http://i21www.ira.uka.de/image_sequences/

Fig. 3. Reference image. From left to right: reference image built initially; reference estimated with a GMM (at frame 250); Bottom: represents the no quality coefficients values associed to each box.

Fig. 4. Final images ; from left to right: current image at coder, final image at decoder.

results are close to the transmission which can be achieved with GSM channel.

Our future work will be focused on two axis. On one hand, we will build the first prototype using PC104 modules for Magys company[2] and on the other hand, we will study the effect of LUX (Logarithmic hUe eXtension)[11] on the rendering quality of the image in road surveillance context.

Notes that by using GPRS or future UMTS, one might expect a video rate of 5 img/s and 25 img/s respectively.

REFERENCES

[1] J. Meessen, C. Parisot, C. Lebarz, D. Nicholson, and J. -F. Delaigle. Smart Encoding for Wireless Video Surveillance. In SPIE Proc. Image and Video Communications and Processing 2005, San Jose, CA, January 2005.

[2]Magys is a SME involved in developing video devices mainly for video surveillance.

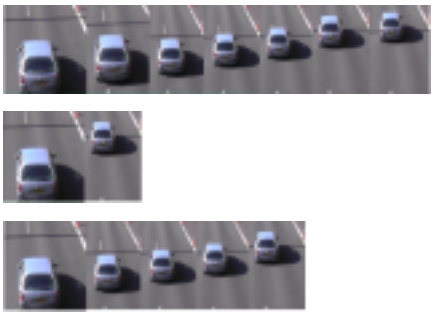

Fig. 5. RRI updating artefact. From top to bottom: image sequence of a scene; effect of simple updating of the RRI (every 5 images); our method for updating the RRI.

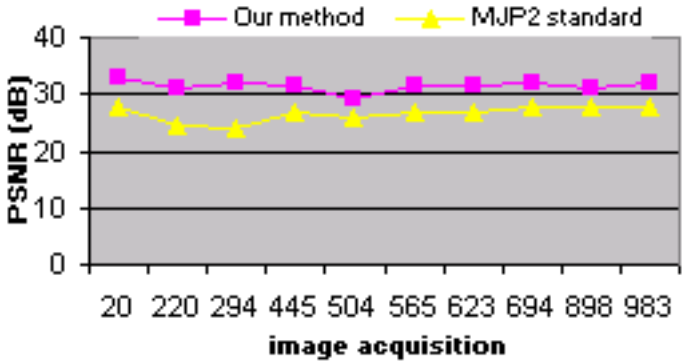

Fig. 6. PSNR comparision: Motion JPEG 2000 (MJP2) coding and our method.

[2] Zhou Wang, Serene Banerjee, Brian L. Evans, and Alan C. Bovik. Generalized Bitplane-by-Bitplane shift method for JPEG2000 Roi coding. IEEE Conference on Image Processing, volume 3, pages 81-84, September 22-25, 2002, Rochester, NY.

[3] D. Gutchess, M. Trajkovic, E. Cohen-Solal, D. Lyons, A. K. Jain. A Background Model Initialization Algorithm for Video Surveillance. International Conference on Computer Vision, volume 1, pages 733, July 07-14, 2001, Vancouver, BC, Canada.

[4] Alec Chi-Wha Wong and Yu-Kwong Kwork. On a Region-of-Interest Based Approach to Robust Wireless Video Transmission. ISPAN, pages 385-390, 2004.

[5] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In Proceeding CVPR, volume 2, pages 246-252, 1999.

[6] Ahmed Elgammal, David Harwood and Larry Davis. Non-Parametric Model for Background Subtraction, 6th European Conference on Computer Vision, pages 751-761, July, 2000, Dublin, Ireland.

[7] Dongsheng Wang, Tao Feng, Heung-Yeung Shum and Songde Ma. A Novel Probability Model for Background Maintenance and Subtraction. The 15th International Conference on Vision Interface, May 27-29, 2002, Calgary, Canada.

[8] A. Caplier, L. Bonnaud et JM. Chassery. Robust fast extraction of video objects combining frame differences and adaptive reference image. ICIP, volume II, pages 785-788, Greece, 2001.

[9] F. Luthon, M. Liévin and F. Faux. On the use of entropy power for threshold selection. *Signal Processing*, volume 84, pages 1789-1804, 2004.

[10] D. S. Taubman et M. W. Marcellin. *JPEG2000 Image Compression fundamentals standard and practice*. Kluwer academic publishers, Netherlands, 2002.

[11] F. Luthon and B. Beaumesnil. Color and R.O.I. with JPEG2000 for wireless videosurveillance. IEEE Int. Conf. on Image Processing, ICIP'04, Singapore, October 2004.

[12] Kakadu SDK website, http://www.kakadusoftware.com.