Journal of Engineering and Applied Sciences 13 (Special Issue 10): 8159-8166, 2018
ISSN: 1816-949X

Object Detection Based on Fusion of Multi-Feature for Video Surveillance System

Tawfiq A. Al-Asadi and Fanar Ali Joda
Department of Software, Faculty of Information Technology, University of Babylon, Hillah, Iraq

Abstract: Object detection is the most important stage due to its great impact on next stages. Moving objects detection in dynamic environment has several challenges such as dynamic background, illumination changes (gradual, sudden), noise and etc. Traditional techniques of background modeling do not have the ability to face these challenges since, they can deal with a static background. Based on spectral, spatial and temporal features, we have proposed multi-modal method to model background and detecting moving object for video surveillance system it can deal efficiently with these challenges. The proposed system consists of two stages, the first one is construction of background model by select N frames to create group of histograms for each pixel. The second one is foreground detection and maintenance scheme for each pixel two histograms are constructed: histogram of ACS-LBP descriptor to extract local texture patterns and histogram of hue channel to extract color information. Temporal values fused within ACS-LBP histogram of intensity. Then maintenance scheme for background model is implemented by using two learning rate. Experiments on video challenging sequences illustrate the efficiency of the proposed method compared to the benchmark methods.

Key words: Background modelling, foreground detection, texture, histogram, local binary pattern,benchmark methods

INTRODUCTION

With the great development of capture and storage devices in recent years ago, video surveillance has become an important subject related to city security, home, traffic analysis and etc., The amount of video data has increased dramatically and automatic video analysis has become necessary (Hadi and Sabah, 2015). Background modelling and foreground detection is essential step for video processing applications such as intelligent video surveillance, optical motion capture, content based video retrieval, human machine interaction, activity recognition and many other applications (Bouwmans *et al*., 2014).

Methods of traditional background modelling focus on the temporal changes between consecutive frames to detect moving objects. Average, median and standard deviation represent traditional methods that depend on thresholding the difference between reference modeland current frame such as these methods fail to deal with challenging situations such as illumination changes, dynamic background and noise due to it do not have the ability to construct multi-modal background. Background model must be robust to overcome a number of challenges by building a multi-model and be adapting with challenging situations (Bouwmans *et al*., 2014).

The methods of background model are classified into pixel based methods and region (block) based methods, the classification depends on the method is used to construct background model. In pixel based methods each pixel remains independent when building the model. This case is the most common but does not take into account temporal and spatial features. These methods are not effective for environments that have rapid changes in the background (Maddalena and Petrosino, 2008).

In region (block) based methods, the frames is divided into overlapped or non-overlapped blocks, the features associated with the block are calculated such as correlation, covariance and histogram to model the block. The block based methods provide high effectiveness against dynamic background. In non-overlapping methods the moving object is segmented and produced coarser shape than the overlapping methods. In overlapping methods each pixel is modeled based on features of block, the detection of moving object is done at pixel level, so, the object's shape is better than non-overlapping methods (Moeslund *et al*., 2006; Xue *et al*., 2011). Features can be categorized by their essential properties into the following types (Li *et al*., 2004):

Spectral features can be obtained directly from the images such as intensity and color. Spectral features are

Corresponding Author: Fanar Ali Joda, Department of Software, Faculty of Information Technology, University of Babylon, Hillah, Iraq

possible to identify changes easily if the difference in color value between the background and foreground is large, so, the spectral features lead to detect FP (False Positive) due to illumination variations and to detect FN (False Negative) when the object has same color with background, this is called camouflage problem. Spectral features ignore the relationships between the adjacent pixels to the target pixel.

Spatial features (texture and edge features) help to extract the moving object that has camouflage problem with background in addition, these features have the ability to handle shadow problem. They are not applicable with dynamic background for pixel level as long as the spatial features are variable over time. Temporal feature represents the relationship between consecutive frames.

Literature review: Heikkila and Pietikainen (2006) proposed local texture descriptor based on Local Binary Pattern (LBP) to model the background and foreground detection. Each pixel is modeled as set of LBP histograms that are calculated for a region with a specific radius around the pixel. This method work on block level therefore the shape of object is not accurate.

Shimada and Taniguchi (2009) proposed Spatial Temporal-Local Binary Pattern (ST-LBP) consists of two types of background model, the first one is background model of pixel level which is strong against long term illumination changes. The second one is spatio-temporal background model which is used to short term illumination changes.

Liao *et al*. (2010) proposed method consists of three stages, first stage, scale invariant local ternary pattern is calculated to handle illumination changes effectively. Second stage, Kernel Density Estimation (KDE) is used to model probability distributions of local pattern in pixel level, then uses one single LBP to represent pattern instead of histogram feature. Third stage, multi-modal background model is used to deal with complex dynamic background using multi-scale fusion.

Hadi and Sabah (2015) presents spatial color feature extraction descriptor called Spatial-Color Binary Pattern (SCBP). The features are extracted include color and texture information. Additionally, refine model is constructed to enhance the boundary of moving objects. For each pixel, multi SCBP histograms are calculated for a region with radius around the target pixel, then the pixel in new frame is labeled into foreground or background according to matching degree between SCBP histogram and background model, then the label is refined and the model is updated.

Yuan *et al*. (2011) proposed Uniform Local Binary Pattern (ULBP) with new method for thresholding to deal with noise interfere adaptively. The background model is built by combine enhanced LBP with color information by adopting a Mixture of Gaussian (MOG) method that uses several models to represent the background. Finally, LBP features vector is simplified to reduce time complexity.

Wu *et al*. (2014) present a method to combine center symmetric scale invariant local ternary pattern with Kernel Density Estimation (KDE) for constructing background model. Then the background model is used to segment the foreground for each new frame by comparing new model with background model. This method execute in pixel level to produce accurate shape.

Charles and Bilodeau (2014) present training-less, single model, single update method and spatio-temporal based background modelling and foreground detection method which is in part consequent from the non-parametric ViBe technique. The main characteristic to spatio-temporal neighborhood analysis at this point is that instead of solely using pixel level intensities as the primary component to generate the reference model (in ViBe-like model), then complement it with the output of a modified Local Binary Similarity Pattern (LBSP) descriptor.

Panda and Meher (2015) proposed a method to model the background and segmented of foreground which depends on oriented gradient and local binary pattern descriptor to construct histogram for difficult dynamic scene. The pixels of background model are modeled as set of vector of feature, each vector of feature is calculated using pixel neighborhood. This method can be handled multi-modal of background subtraction based Mixture of Gaussian (MOG).

Zhang *et al*. (2016) present new model consists of two layers based on codebook technique combined with local binary pattern descriptor to solve dynamic background and illumination variations problem. The first layer is constructed by block based codebook merging with local binary pattern histogram and average values of RGB color channels. Due to of the stability of the LBP features with related to monotonic illumination variations, this layer can create block-wise detection results with considerable tolerance of illumination changes. The pixel based codebook is used to support the accuracy from the results of the first layer which is to remove false positive further.

MATERIALS AND METHODS

Proposed system: Background modeling and object detection are the most important stages in the video processing applications such as multimedia applications, optical motion capture, video-surveillance and content based video retrieval which aims to build and update a

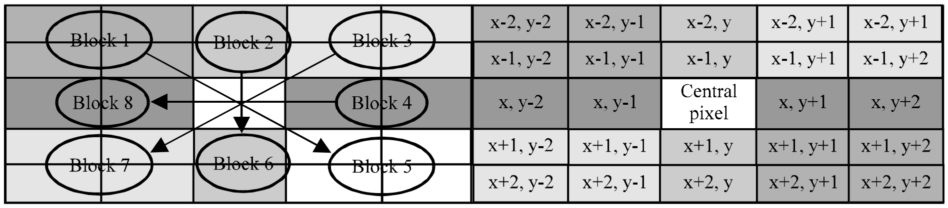

Fig. 1: ACS-LBP descriptor

statistical representation of a particular scene captured by a camera when the camera is stationary. In proposed system, each pixel is modeled identically and this leads to high parallel execution speed if necessary. The main components of proposed system are explained in following sections.

Average Center Symmetric Local Binary Pattern (ACS-LBP) descriptor: Recently, Center Symmetric Local Binary Pattern (CS-LBP) proposed by Heikkila *et al.* (2009) to describe the information of local texture depending on the size of block. It has the ability to face illumination variations, additionally it use to model pixel level for moving object detection. In this study, we have proposed Average Center Symmetric Local Binary Pattern (ACS-LBP) descriptor which are considered as extended and improved to CS-LBP descriptor. The average values of pairs are compared for pixels in center symmetric direction as shown in Fig. 1.

The average of pixels in block 1 is compared with the average of pixels in block 5 and the average of pixels in block 2 compared with the average of pixels in block 6 and so, on. We can calculate ACS-LBP descriptor by Eq. 1:

$$\left.\begin{array}{l} S_1 = S\,(\text{Avg.}\,(x<0 \text{ and } y<0)\text{-Avg.}\,(x>0 \text{ and } y>0)\ 2^0+ \\ S_2 = S\,(\text{Avg.}\,(x<0 \text{ and } y=0)\text{-Avg.}\,(x>0 \text{ and } y=0)\ 2^1+ \\ S_3 = S\,(\text{Avg.}\,(x<0 \text{ and } y>0)\text{-Avg.}\,(x>0 \text{ and } y<0)\ 2^2+ \\ S_4 = S\,(\text{Avg.}\,(x=0 \text{ and } y>0)\text{-Avg.}\,(x=0 \text{ and } y<0)\ 2^3 \end{array}\right\} \tag{1}$$

The value of S_i can be equal to 1 or 0 based on Eq. 2:

$$S_i(p)\begin{cases} 1 \text{ if } p\geq\tau \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

The value of τ is set by user, usually the value of τ is small value it produces 2^4 (16 binary patterns). ACS-LBP can address noise problem at pixel level better than CS-LBP descriptor, Fig. 2 shows how ACS-LBP descript orcan obtain same result even with noise,on the contrary

(a)

68	66	72	50	49
64	64	68	60	47
66	60	64	68	55
67	61	63	61	61
74	71	70	67	63
CS-LBP: 1100				
ACS-LBP: 1100				

(b)

68	66	72	50	49
64	64	65	65	47
66	60	64	52	55
67	61	63	63	61
74	71	70	67	63
CS-LBP: 0110				
ACS-LBP: 1100				

Fig. 2: Comparison between CS-LBP and ACS-LBP descriptors; a) Original data and b) Red pixels are changed with noise, the value of τ is 2

from CS-LBP descriptor it obtains a different result with noise. This method is also characterized by ability to solve monotonic and linear illumination changes and robustness on flat regions of image using thresholding the differences based on τ value.

In this study we have proposed a method to fuse temporal information with spatial information at the same time, the Eq. 1 is updated as follows:

$$\left.\begin{array}{l} S_1 = S\,(\text{Avg.}\,(x<0 \text{ and } y<0)\text{-Avg.}\,(x>0 \text{ and } y>0)\ 2^0+ \\ S_2 = S\,(\text{Avg.}\,(x<0 \text{ and } y=0)\text{-Avg.}\,(x>0 \text{ and } y=0)\ 2^1+ \\ S_3 = S\,(\text{Avg.}\,(x<0 \text{ and } y>0)\text{-Avg.}\,(x>0 \text{ and } y<0)\ 2^2+ \\ S_4 = S\,(\text{Avg.}\,(x=0 \text{ and } y>0)\text{-Avg.}\,(x=0 \text{ and } y<0)\ 2^3+ \\ S_5 = f\,(p(x,y,t)\text{-}\bar{\mu}(x,y,t\text{-}1))\ 2^4 \end{array}\right\} \tag{3}$$

The function f(p) set zero or one based on the value of current pixel p(x, y, t), mean $\bar{u}(x, y, t\text{-}1)$ and variance $\bar{\sigma}^2(x, y, t\text{-}1)$ as Eq. 4.

$$f(x) = \{(0 \text{ if} \left|P(x,y,t)\text{-}(\mu)^{-}(x, y, t\text{-}1)\right|<2\sigma^{-2}(x, y, t\text{-}1)@1 \tag{4}$$

After this change is added for ACS-LBP descriptor it produces 2^5 (32 binary patterns). Estimation method is adopted to estimate variance and mean values for each pixel. ACS-LBP descriptor is more efficient in dynamic scenes.

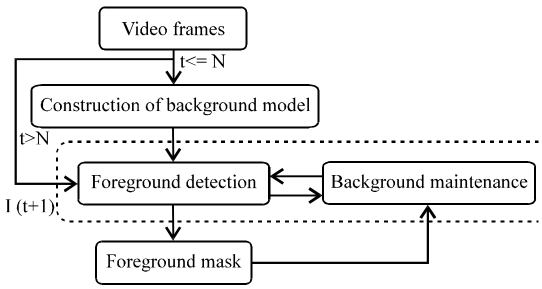

Fig. 3: Block diagram of proposed system

$$\bar{\mu} = (1-\varepsilon)\bar{\mu}_{t-1}+\varepsilon I_t \tag{5}$$

$$\bar{\sigma}_t^2 = (1-\varepsilon)\bar{\sigma}_{t-1}^2+\frac{\varepsilon}{2}(I_t-I_{t-1})^2 \tag{6}$$

where, ε is a learning rate I_t and I_{t-1} represent current and previous pixel values, respectively.

The approach: The main stages of background modelling and moving object detection is showed in Fig. 3. The stages in Fig. 3 are explained as following:

Construction of background model: Is achieved by investment of N clean frames at the first sequence of frames when the frames are consecutive, logically, this suggestion of (N clean frames) is not often obtain in real situations because of existence continuous clutter. In general, the model is prepared by means of set of training frames which contain or do not contain foreground objects. When the N frames processing is finished, the result will be a group of modals. For each pixel, background model consists of group of adaptive weighted histograms k user predefined when k is increased, this leads to better results in term of {m→0, ..., m→k}accuracy but the time complexity and memory requirements will be raised, the appropriate number of k is 3 which represents tradeoff between the accuracy on one hand and time complexity and memory requirements on the other hand.

Foreground detection: The pixels are classified as foreground or background by comparing the similarity measure between current frame and background model.

Background maintenance: In this stage the background model is updated by using current background model, current frame foreground mask and classification of previous frame pixels. It is incremental online. The background maintenance and foreground detection stages are done frequently during the processing time.

In the following sections, the proposed method will be presented in detail: After finishing the construction stage of the background model, set of procedures are achieved for each new frame:

- Convert RGB color space to HSV color space
- ACS-LBP histogram is constructed for V channel
- Hue channel histogram is constructed

In addition, set of matrices must be existed before starting with the processing procedure, first one is binary matrix of foreground mask for previous frame, due to the learning rate of background model is updated based on pixels classification of previous frame, the other two matrices are mean and standard deviation which represent history of pixel to handle global illumination changes.

HSV (Hue, Saturation, Value) color space is used, the images are obtained from video are RGB color space which are converted to HSV. The motivation of using HSV color space is high ability to separate intensity (represented by V channel) from chrominance (represented by H and S channels). Three component values (RGB) decrease or increase together as the brightness decreases or increases, respectively while one component value (V) decrease or increase together as the brightness decreases or increases whereas the other two components (H and S) are less affected when the lightness changes. It is suitable for algorithms that rely on intensity and color information. In addition, HSV color space has a better possibility of detecting moving object and removing cast shadow from detected object.

For each pixel, ACS-LBP histogram for intensity (V) channel is computed through circular region with R radius around the pixel, ACS-LBP histogram is used as feature vector. The Hue channel is quantized into u levels to build Hue histogram as show in the Fig. 4, the aim of quantization of Hue channel is easy to calculate and reducing the size of histogram that leads to decrease the time complexity. Similar to feature vector of intensity texture, an N×N pixels structuring element is used to gather the statistic of local color of a pixel.

The motivation for using color feature represented by histogram of quantized Hue channel to solve the problem of texture feature Furthermore, a method based on texture only may cause detection errors in regions of heterogeneous texture and blank texture. The pair of (ASC-LBP and Hue histograms) has weight symbolized with w_k and its value between [0, ..., 1], the sum of all weights for each k model histograms equal to 1. ACS-LBP and Hue channel histograms are calculated using N frames at beginning of the frames sequence and the frames are consecutive.

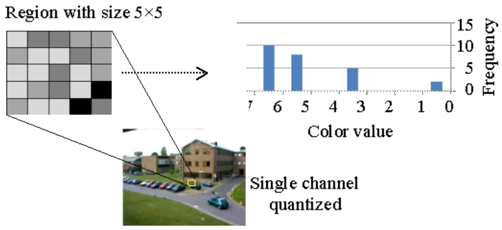

Fig. 4: Example of histogram of quantized (8 levels) Hue channel

The new histogram is compared with K model histograms that related with corresponding pixel using histogram intersection as a measure of distance as follows:

$$\cap(h, m) = \frac{\sum_{t=0}^{L-1} \min(h_i, m_i)}{\min(\sum_{i=0}^{L-1} h_i, \sum_{i=0}^{L-1} m_i} \quad (7)$$

Where:

h = Current histogram
m = Existing histogram
i = Index of histogram bins
L = Number of bins of histogram

There are many distance measures for comparing between two histogram such as log likelihood, Chi squared and intersection. The aim behind selects histogram intersection is to decrease the time complexity due to it easy to implement. Since, background model is represented by ACS-LBP and Hue histograms, the final similarity measure is obtained by using weighted averaging as follows:

$$S_{T,H} = \gamma.S_T + (1-\gamma).S_H \quad (8)$$

where S_H and S_T represent Hue and ACS-LBP similarity measures, respectively that is obtained using Eq. 7 while γ variable denotes to the weight of the importance of color and texture features. When γ variable takes high value, this indicates to importance of texture feature instead of color feature.

The pixel is classified as foreground when $S_{T,H}<T_d$ for all histograms model, new histogram model is created instead of histogram model with least weight. Small initial weight is assigned for new histogram model. In our method, the value of initial weight is 0.005 and then no further processing is needed. If new histogram matches one of histograms model in this case the processing is necessary. Histogram model with best matching is adopted with new value by changing its bins.

In most common methods of background modelling and foreground detection are used background maintenance scheme to update the histogram model with same learning rate. The main problem for these methods is some of pixels which are misclassified as background in current frame while are classified as foreground in previous frame, these pixels will be used to update the values of background histograms model and its weights faster than required. In proposed method, the maintenance method is used to update best matching of histogram model employs two learning rates. The selection of learning rate is based on the classification of pixels in previous frame into background or foreground as follows:

$$\overrightarrow{m_k} = \alpha\vec{h} + (1-\alpha)\overrightarrow{m_k} \quad (9)$$

If the value of corresponding pixel is 0 in binary matrix. While:

$$\overrightarrow{m_k} = \beta\vec{h} + (1-\beta)\overrightarrow{m_k} \quad (10)$$

If the value of corresponding pixel is 1 in binary matrix. α and β are defined by user, the idea of using two learning rates is to reduce the effect of the misclassification of pixels in previous frame, therefore, the β value is much smaller than α value. The weights of histograms model also are updated as follows:

$$w_k = \alpha_w M_k + (1-\alpha_w) w_k \quad (11)$$

where, $\alpha_w \in [0, 1]$and α_w is defined by user. For the best matching histogram, M_k = 1 otherwise equal 0. After finishing update the weights of histograms model, then the weights are normalized to become equal to 1. The parameters α_w, α and β are controlled on speed of adaptation. The larger values of parameters lead to faster adaptation. In last step of background modelling, the histograms model are reordered according to its weights with descending order, then select first B_H histograms to be the model of background as follows:

$$\text{Select } B_H \text{ where } w_0 +, ..., +w_{B-1} > T_B \quad (12)$$

where, $T_B \in [0, 1]$ and T_B is defined by user. To address sudden global illumination change, the current frame will be divided into equal blocks with size N×N, then the intensity value of centers of all blocks are compared with corresponding location in previous frame, if all locations are changed with fixed ratio in terms of increasing or decreasing this indicates to exist global illumination change, if the change value larger than threshold th_l and the values of H and S channels of blocks center have change less than threshold th_2, then the learning rates will be updated according to ratio of change in illumination. The aim of selecting only center of blocks instead of all pixels for comparing process is decreasing the time complexity.

RESULTS AND DISCUSSION

Sequences of two videos are tested to prove the effectiveness of proposed method, as shown in Fig. 5-7. Gaussian of Mixture (MOG), Kernel Density Estimation (KDE) and codebook methods are employed to compare with the proposed method. For the experimentations, these parameter values: N = 1000 frames, size of block = 5×5 = 0.005, u = 32 levels, $\gamma = 0.7$, $T_d = 0.65$, $\alpha = 0.01$, $\beta = 0.002$, $\alpha_k = 0.01$ and $T_b = 0.7$.

To evaluate the performance of our method, Precision Rate is used, True Negative (TN)/True Positive (TP) are defined as the number of background/foreground pixels that are correctly classified as background/foreground pixels. False Negative (FN)/False Positive (FP) are defined as the number of background/foreground pixels that are incorrectly classified as foreground/background. So, the precision rate can be computed as following:

$$PR = \frac{TP}{TP+FP} \tag{13}$$

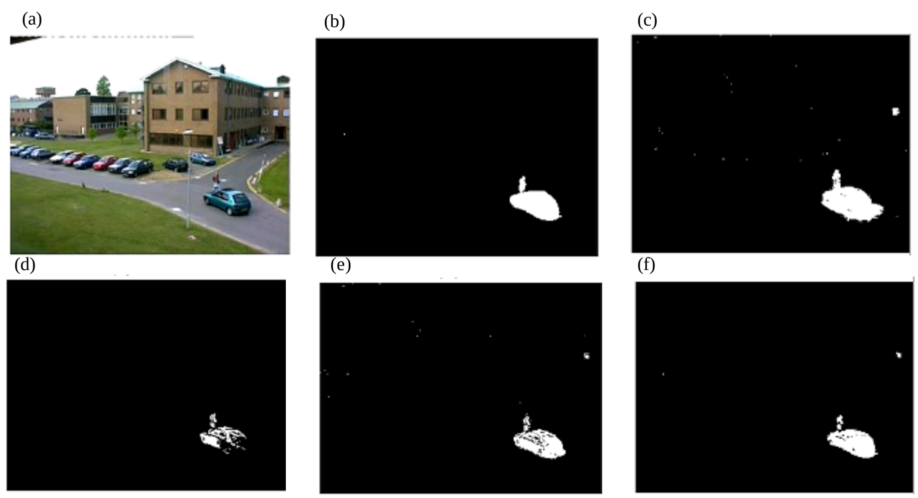

Fig. 5: a) PETS2001 results of; b) Ground truth; c) Codebook; d) MOG; e) KDE and f) Proposed method

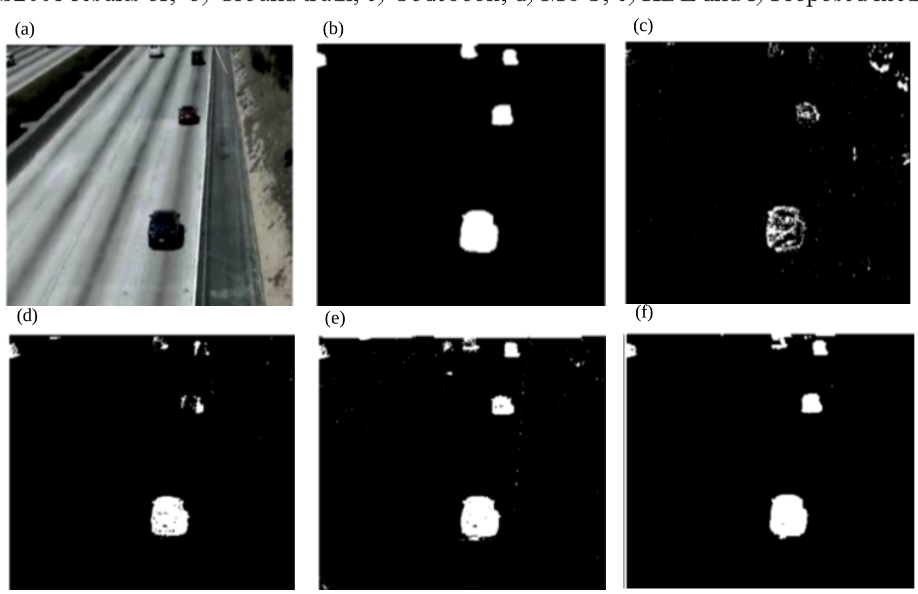

Fig. 6: a) HighwayII results of; b) Ground truth; c) Codebook; d) MOG; e) KDE and f) Proposed method

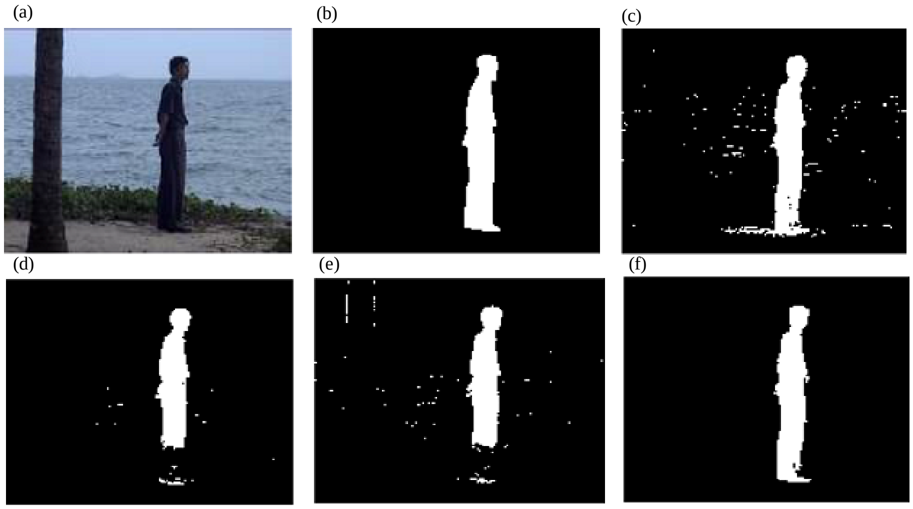

Fig. 7: a) Water surface results of; b) Ground truth; c) Codebook; d) MOG; e) KDE and f) Proposed method

Table 1 Comparison of precision rate

Sequence name	Codebook	MOG	KDE	Proposed method
PETS2001	83.56	78.09	87.38	91.64
Highway II	61.26	75.61	55.91	89.14
Water surface	81.52	87.48	89.37	88.25

CONCLUSION

ACS-LBP is proposed to extract local texture information and to address noise problem at pixel level. ACS-LBP characterize by low time complexity due to produce 2^4 patterns. Temporal value is fused with ACS-LBP to obtain motion information and color information is combined with texture information by weighted averaging to increase discrimination in case of flat region. The proposed system can be applied with pixel or block level. Frame can be divided into overlapped or non-overlapped block depend on requirements of shape accuracy and time complexity. Finally, temporal information addresses sudden global illumination change effectively.

REFERENCES

Bouwmans, T., F. Porikli, B. Hferlin and A. Vacavant, 2014. Background Modeling and Foreground Detection for Video Surveillance. CRC Press, Boca Raton, Florida, ISBN:978-1-4822-0537-4,.

Charles, P.L.S. and G.A. Bilodeau, 2014. Improving background subtraction using local binary similarity patterns. Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), March 24-26, 2014, IEEE, Steamboat Springs, Colorado, ISBN:978-1-4799-4984-7, pp: 509-515.

Hadi, I. and M. Sabah, 2015. Improvement cat swarm optimization for efficient motion estimation. Intl. J. Hybrid Inf. Technol., 8: 279-294.

Heikkila, M. and M. Pietikainen, 2006. A texture-based method for modeling the background and detecting moving objects. IEEE. Trans. Pattern Anal. Mach. Intell., 28: 657-662.

Heikkila, M., M. Pietikainen and C. Schmid, 2009. Description of interest regions with local binary patterns. Pattern Recognit., 42: 425 436.

Li, L., W. Huang, I.Y.H. Gu and Q. Tian, 2004. Statistical modeling of complex backgrounds for foreground object detection. IEEE Trans. Image Proc., 13: 1459-1472.

Liao, S., G. Zhao, V. Kellokumpu, M. Pietikainen and S.Z. Li, 2010. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 13-18, 2010, IEEE, San Francisco, California, ISBN:978-1-4244-6984-0, pp: 1301-1306.

Maddalena, L. and A. Petrosino, 2008. A self-organizing approach to background subtraction for visual surveillance applications. IEEE. Trans. Image Process., 17: 1168-1177.

Moeslund, T.B., A. Hilton and V. Kruger, 2006. A survey of advances in vision-based human motion capture and analysis. Comput. Vision Image Understanding, 104: 90-126.

Panda, D.K. and S. Meher, 2015. Dynamic background subtraction using local binary pattern and histogram of oriented gradients. Proceedings of the 3rd International Conference on Image Information Processing (ICIIP), December 21-24, 2015, IEEE, Waknaghat, India, ISBN: 978-1-5090-0148-4, pp: 306-311.

Shimada, A. and R.I. Taniguchi, 2009. Hybrid background model using spatial-temporal LBP. Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance AVSS'09, September 2-4, 2009, IEEE, Genova, Italy, ISBN:978-1-4244-4755-8, pp: 19-24.

Wu, H., N. Liu, X. Luo, J. Su and L. Chen, 2014. Real-time background subtraction-based video surveillance of people by integrating local texture patterns. Signal Image Video Process., 8: 665-676.

Xue, G., L. Song, J. Sun and M. Wu, 2011. Hybrid center-symmetric local pattern for dynamic background subtraction. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), July 11-15, 2011, IEEE, Barcelona, Spain, ISBN:978-1-61284-348-3, pp: 1-6.

Yuan, G.W., Y. Gao, D. Xu and M.R. Jiang, 2011. A new background subtraction method using texture and color information. Proceedings of the International Conference on Intelligent Computing, August 11-14, 2011, Springer, Zhengzhou, China, ISBN:978-3-642-25943-2, pp: 541-548.

Zhang, Y., J. Bae and W. Kim, 2016. Multi-layer multi-feature background subtraction using codebook model framework. Intl. J. Comput. Inf. Eng., 3: 1-5.