Object Detection based on Combination of Visible and Thermal Videos using A Joint Sample Consensus Background Model

Guang Han
College of Information Science and Engineering, Northeastern University, Shenyang, China
Email: a00152738@sohu.com

Xi Cai and Jinkuan Wang*
Northeastern University at Qinhuangdao, Qinhuangdao, China
Email: cicy_2001@163.com, wjk@mail.neuq.edu.cn

***Abstract*—In uncontrolled video surveillance environments, performing efficient foreground segmentation is very challenging. In order to improve robustness and accuracy of object detection, we take advantage of spectral information of both visible and thermal videos. This paper presents a novel joint background model combining visible and thermal videos for foreground object detection in complex scenarios. Different from traditional methods that first detected moving objects in either domain respectively and then fused the detection results, we provide a joint sample consensus background model with four channels (red, green, blue and thermal) to accomplish the object detection and fusion of complementary information simultaneously, which lowers the computational cost of our method. Raw foreground segmentation is obtained in the thermal domain, making initial foreground more accurate. Meantime this can enhance the efficiency of further steps. Time out map (TOM) is utilized to deal with the problem that a newly exposed background is wrongly marked as foreground for a long time. In the updating phase, unlike most sample-based methods using first-in first-out policy, we intentionally employ a random update policy to reserve some older samples. That is, when a pixel is classified as background, we randomly pick up one of the background samples stored for the corresponding pixel to discard. In this manner, the backgrounds, occluded by slow moving foreground or temporally still foreground, can be recovered promptly when they reappear. Experimental results show that the proposed method can achieve accurate and precise detection results.**

***Index Terms*—Object Detection, joint sample consensus, visible and thermal videos, background model, time out map, random update**

I. INTRODUCTION

Video surveillance is a very active research area in computer vision applications owing to rapidly increasing number of surveillance cameras. To make video surveillance systems more intelligent, there is a strong demand to automatically analyze their output videos. Object detection in video streams is to segment foreground (objects of interest) from background, and it is an important initial step for further high level processing, such as object tracking, object recognition, and activity analysis. Persistence (i.e. 24 hours operation in day and night) is the most desirable quality of a video surveillance system [1]. Therefore powerful object detection algorithms must be robust and effective at all times.

Recently, much effort has been devoted to detection algorithms in the visible domain which is abundant in color and texture information [2]-[13]. However, in real world scenarios, object detection for only visible cameras are generally not effective in sudden or gradual changes of illumination, visibility and weather, generating poor segmentation of actual objects and many false positives. There are also some researches on object detection algorithms in the thermal domain [14]-[17]. Since thermal cameras can capture information of the thermal energy emitted/reflected from objects in the scene, they are independent of illumination. This makes detection algorithms in the thermal domain more effective than that in the visible domain under poor lighting conditions. However, the thermal videos have their unique inherent challenges due to low signal-to-noise ratio, uncalibrated white-black polarity changes, and the "halo effect" that appears around very hot or cold objects, which leads to noise pollution, inaccurate location and enlarged scale of the detection results [1].

A promising solution in non-ideal environment is to segment foreground objects relying on combination of multimodal sensors. Combining the thermal sensors and the visible sensors can be advantageous in two ways. First, thermal and visible information has complementary nature, which would remedy the opposite side's default. Second, the redundancy of information captured by different sensors of different spectrums increases the accuracy and robustness of detection algorithms.

A few methods have combined information from both sensors to robustly detect and track the moving objects

* Corresponding author.

doi:10.4304/jsw.8.4.987-994

[1], [18]-[21]. These methods all extracted initial foreground in the visible domain and the thermal domain respectively, and then fused them to further improve the detection results. However, implementing object detection twice was so complicated that it caused very high computational cost, and the fusion step at the feature level only exploited limited complementary information of original images.

In this study, we propose a method for object detection in the thermal and the visible domains using a joint background model (JBM-TV for short) to achieve the detection and fusion in the meantime. The joint background model is nonparametric. Each pixel has a set of background samples and each background sample has four channels (red, green, blue and thermal). Inspired by sample consensus method with three channels originally presented in [8], we make some improvement by using four channels joint sample consensus to combine the thermal and the visible videos. It is different from aforementioned traditional methods in combining information from both sensors at the pixel level. More specifically, each observed pixel is classified as foreground or background according to its similarity to the joint background model.

Since a joint background model of augmented "Red-Green-Blue-Thermal" video input is built, object detection and fusion can be executed simultaneously, which lowers the complexity of detection method. The thermal video is used to extract raw foreground pixels to get fewer miss detections and more accurate regions of interest (ROIs). A memoryless update policy is employed. Different from the traditional first-in first-out (FIFO) policy, we randomly update the background samples by allowing a few old samples to remain in the background model to improve the time relevance. Experimental results demonstrate that this framework allows us to fuse the complementary information from both domains, making it a robust algorithm with accurate and precise detection results.

The organization of the remainder of this paper is as follows: in Section II, we present a short review of previous related work on object detection. Section III describes our technique and details our major innovations. In Section IV, experimental results show the advantages of our method qualitatively and quantitatively. Finally, we conclude the paper in Section V.

II. Related Work

Some works have been published on detection methods appropriate for joint thermal-visible video surveillance applications. Davis et al. proposed a background subtraction algorithm fusing contours from thermal and visible images [1]. Binary contour fragments in both domains were extracted from initial ROIs respectively using contour saliency map, and then combined into a better and less broken contour image. However, since contours were first extracted from cluttered or incomplete foregrounds and then fused, further post-processing was required to clear the erroneous contour lines and complete the missing contour segments. A belief model was developed in [19] to determine the validity of a foreground region detected in either sensor respectively. The improved segmentation of foreground objects was achieved through fusion of reliable sensor measurements. In [20], Ulusoy et al. implemented background modeling using a single Gaussian and foreground detection respectively in infrared image, intensity image and color image, and then fused the detection results. The resulting foreground regions were used as a mask on the infrared image, and snake algorithm was applied to detect object boundaries.

All these methods first detected objects in both domains respectively, and then fused the detection results to utilize the complementary information. They can gain good detection results, but their frameworks are very complicated.

III. The Proposed Detection Method

We present a detection method based on a joint background model to simultaneously detect and fuse visible and thermal videos. The proposed framework is shown in Fig. 1. At first, we use the thermal video to segment the raw foreground region (i.e. possible candidate foreground pixels). This can lower the amount of processed data significantly. Candidate foreground pixels and their corresponding background samples are fed into the joint sample consensus algorithm, and then the pixels from the dynamic background and other uninteresting moving objects (e.g. shadows and halos) are suppressed. Some post processing methods are utilized to improve the integrity and quality of the detected foreground. In order to suppress "ghosts", a technique named time out map (TOM) is employed in our method to assign lasting foreground pixels to be background. To

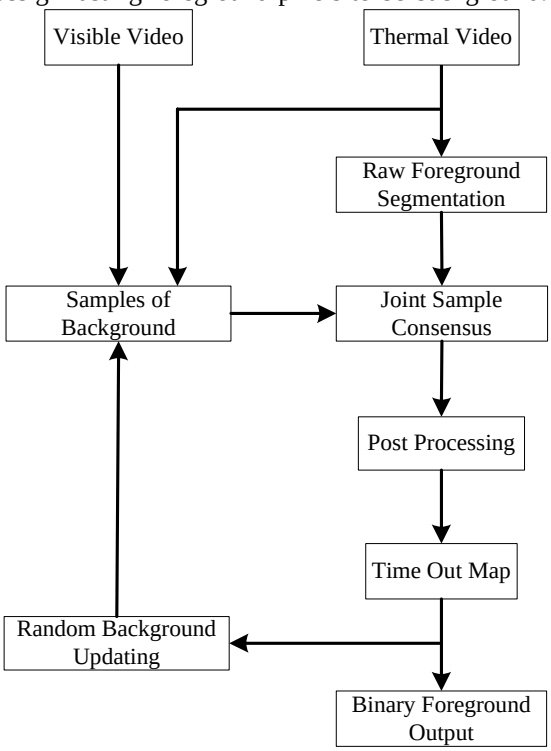


Figure 1. Framework of the proposed method.

make the background model adaptive, random update policy of the background samples is employed every single frame for pixels which are labeled as background. Finally, binary foreground output is achieved.

A. Raw Foreground Segmentation

Raw foreground regions are segmented in the thermal domain because there are fewer miss detections and false positives in the thermal domain compared to the visible domain. Adjacent frame differencing used in [8] is not adopted to extract candidate foreground pixels, for it yields holes in the foreground region. Instead, pixel-wise median values are computed during a training phase, and pixels of test image which exceed a fixed threshold of the median are considered foreground. The median and threshold method can get more accurate raw foreground.

B. Joint Sample Consensus

As a nonparametric background model, *N* background samples are recorded for each pixel. The values of each background sample are obtained from both visible and thermal videos, denoted by $x_{\mathbf{m}}(i) = \{x_{\mathbf{m}}^{R}(i), x_{\mathbf{m}}^{G}(i), x_{\mathbf{m}}^{B}(i), x_{\mathbf{m}}^{T}(i)\}$, where R, G, B for the red, green, blue channels, T for the thermal channel, and $x_{\mathbf{m}}(i)$ indicates four channel values of the *i*th sample for pixel **m**.

For current frame at time t (t>*N*), we define $x_{\mathbf{m}}(t)$ as the observed four channel values for pixel **m**. According to similarity between the current observation and background samples, *N* votes are generated for each channel as follows,

$$Vote_{\mathbf{m}}^{c}(i,t) = \begin{cases} 1, & \left| x_{\mathbf{m}}^{c}(i) - x_{\mathbf{m}}^{c}(t) \right| \le T_{\mathbf{m}}^{c}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $T_{\mathbf{m}}^{c}$ is a threshold, proportional to standard deviation $\sigma_{\mathbf{m}}^{c}$ of background samples. Moreover, considering that $\sigma_{\mathbf{m}}^{c}$ may be overestimated when the background samples are multimodal distributed, we also set a constant $T_{\max}$ to limit the maximum value of $T_{\mathbf{m}}^{c}$. So

$$T_{\mathbf{m}}^{c} = \min(T_{\max}, \eta\sigma_{\mathbf{m}}^{c}), \tag{2}$$

where η is usually set as 2.5.

Equation (1) means that, when current observation value is similar to one of the background samples, one "agree" vote would be produced to suggest the current observation as a background pixel. Counting the number of "agree" votes in each channel, the proposed method obtains background decision as follows,

$$Bg_{\mathbf{m}}(t) = \begin{cases} 1, & \sum_{i=1}^{N} Vote_{\mathbf{m}}^{c}(i,t) \ge T_{N}^{c} \ \ \forall c \in \{R,G,B,T\}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where T_{N}^{c} is a threshold for the total amount of "agree" votes, and $Bg_{\mathbf{m}}(t)$ is a binary value. When the total amount of "agree" votes is lager than T_{N}^{c}, $Bg_{\mathbf{m}}(t)$ equals one which represents a background pixel; Otherwise, $Bg_{\mathbf{m}}(t)$ equals zero that signifies a foreground pixel. Obviously, T_{N}^{c} is in direct proportion to sample size *N* and threshold $T_{\mathbf{m}}^{c}$. Namely,

$$T_{N}^{c} = \tau T_{\mathbf{m}}^{c} N, \tag{4}$$

where τ is a constant and chosen empirically.

C. Post Processing

Morphological operations are employed in post processing. First, morphological close operation to connect foreground part is implemented. We use vertical line as structure element, because human has many vertical edges. Then, morphological fill operation is used to fill the holes in the foreground region. At last, connected component analysis is utilized for each resulting image, eliminating the foreground pixels whose 4-connected region is less than 20.

D. Time Out Map

After post processing, there still exists a problem that some pixels belong to foreground continuously. This is abnormal in real-world scenario though the speed of some moving objects is slow. This phenomenon is called "ghosts" and it is always caused by the moved or inserted background objects. We use a simple counter TOM to solve this problem, and $TOM_{\mathbf{m}}(t)$ is the counter's value at pixel **m** at frame t. This map is incremented every frame at the pixels that have been classified as foreground. Once a pixel is classified as a background pixel, $TOM_{\mathbf{m}}(t)$ is set to zero. Namely,

$$\begin{cases} TOM_{\mathbf{m}}(t) = TOM_{\mathbf{m}}(t-1) + 1, & Bg_{\mathbf{m}}(t) = 0, \\ TOM_{\mathbf{m}}(t) = 0, & \text{otherwise}. \end{cases} \tag{5}$$

When $TOM_{\mathbf{m}}(t)$ is larger than a threshold, pixel **m** should be assigned to the background.

E. Random Background Updating

As the background scene is always changing, the background samples should be updated to fit the changes of the environment. We employ conservative update policy by selectively adding only pixels marked as background to the background model.

Furthermore, we randomly choose the *i*th（i=1,…,N）background sample to update as follows,

$$\begin{cases} x_{\mathbf{m}}^{c}(i) = x_{\mathbf{m}}^{c}(t), & Bg_{\mathbf{m}}(t) = 1, \\ \text{no updating}, & \text{otherwise}. \end{cases} \tag{6}$$

In this way, older samples are not discarded directly like FIFO, but reserved partly. According to the valuable information in the older samples, we can recover the backgrounds when they reappear after once occluded by slow moving foreground or temporally still foreground.

IV. Experimental Results

In this section, we shall present results of our JBM-TV method while challenging real-world visible and thermal videos. We take a publicly available OTCBVS Benchmark Dataset Collection (Dataset 03: OSU Color-Thermal Database) [1], which has six sequences in the dataset. The first two sequences (Campus A and Campus B) are used in our test and they have color and thermal images recorded in a daytime situation where a group of clouds is passing in the sky, causing abrupt illumination changes in various parts of the observed scene. We empirically set the number of background samples N=50.

A. Qualitative Analysis

The aim of this experiment is to demonstrate the efficacy of our algorithm for object detection. For comparison, we also introduce Zivkovic method [22], a very promising object detection method recommended by [23], to perform on visible and thermal domain respectively.

Examples of silhouettes extracted are shown in Fig. 2-Fig.7. In Fig. 2-Fig. 5, we can see that in the visible domain, Zivkovic method detected a lot of false positives because of the sudden illumination changes, and also detected shadows as foreground. The proposed JBM-TV method can eliminate the false positives and shadows by using the thermal video in initial segmentation. As seen from Fig. 6 and Fig.7, in the thermal domain, Zivkovic method detected halos around the person, but JBM-TV method can still eliminate the halos by combining visible information.

B. Quantitative Analysis

We use two CLEAR metrics [24], multiple object detection accuracy (MODA) and multiple object detection precision (MODP), to quantitatively analyze the proposed method.

MODA is to access the accuracy of the performance at the object level. Assuming that the number of misses is denoted by m_t and the number of false positives is denoted by fp_t for each frame t, MODA is defined as

$$MODA(t) = 1 - \frac{m_t + fp_t}{N_G^t}, \tag{7}$$

where N_G^t is the number of ground truth objects in frame t. The normalized MODA (N-MODA) for the entire sequence is denoted as

$$N - MODA = 1 - \frac{\sum_{t=1}^{N_{frames}} (m_t + fp_t)}{\sum_{t=1}^{N_{frames}} N_G^t}, \tag{8}$$

where N_{frames} is the total frame number of the sequence.

MODP is to evaluate the precision of the performance at the pixel level. We first compute the mapped overlap ratio (MOR) as follows

$$MOR(t) = \sum_{i=1}^{N_{mapped}^t} \frac{\left| G_i^t \cap D_i^t \right|}{\left| G_i^t \cup D_i^t \right|}, \tag{9}$$

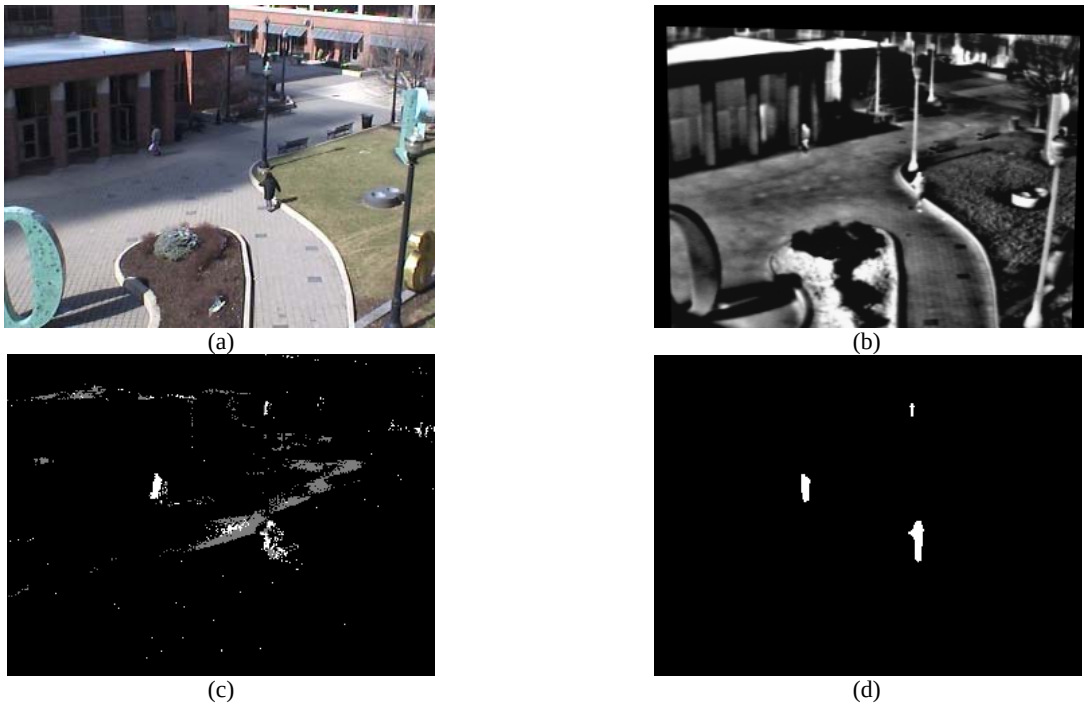


Figure 2. Original frames and detection results of frame 131 of Campus A: (a) original frame in the visible domain. (b) original frame in the thermal domain. (c) using Zivkovic method in the visible domain. (d) using JBM-TV method in both domains.

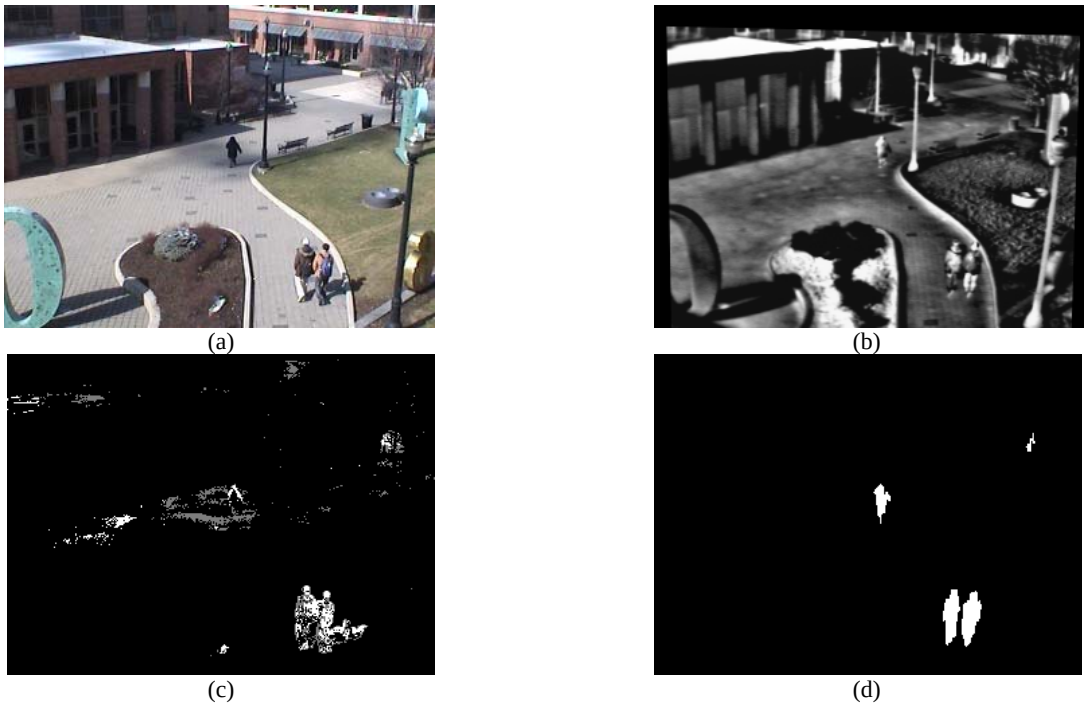


Figure 3. Original frames and detection results of frame 301 of Campus A: (a) original frame in the visible domain. (b) original frame in the thermal domain. (c) using Zivkovic method in the visible domain. (d) using JBM-TV method in both domains.

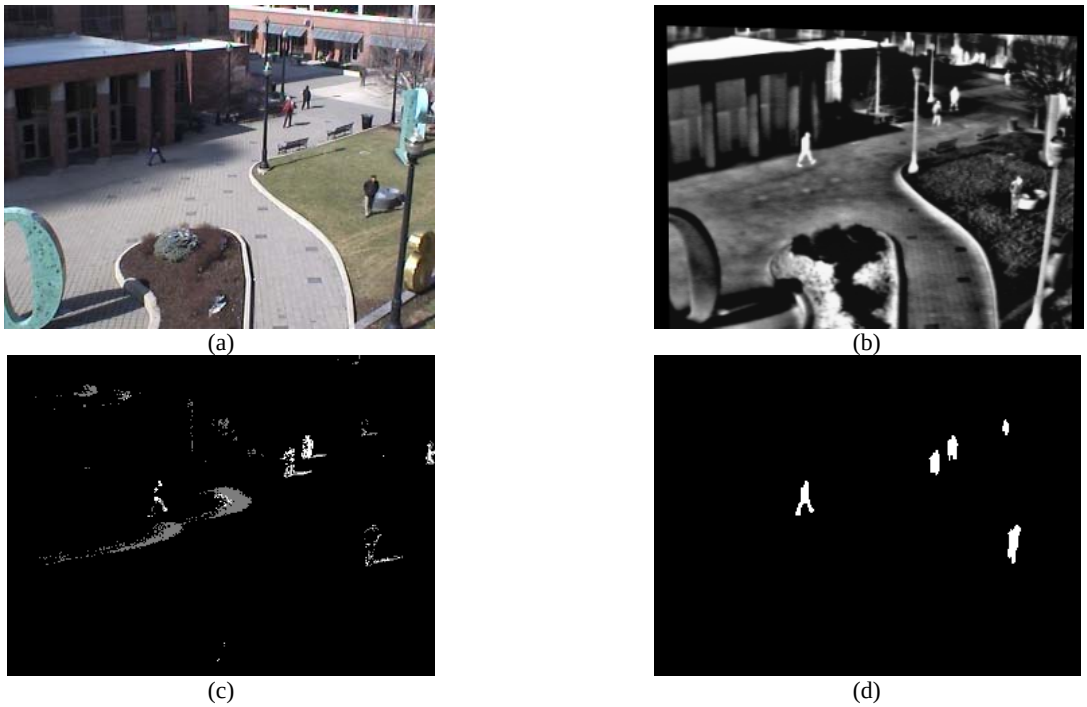


Figure 4. Original frames and detection results of frame 341 of Campus B: (a) original frame in the visible domain. (b) original frame in the thermal domain. (c) using Zivkovic method in the visible domain. (d) using JBM-TV method in both domains.

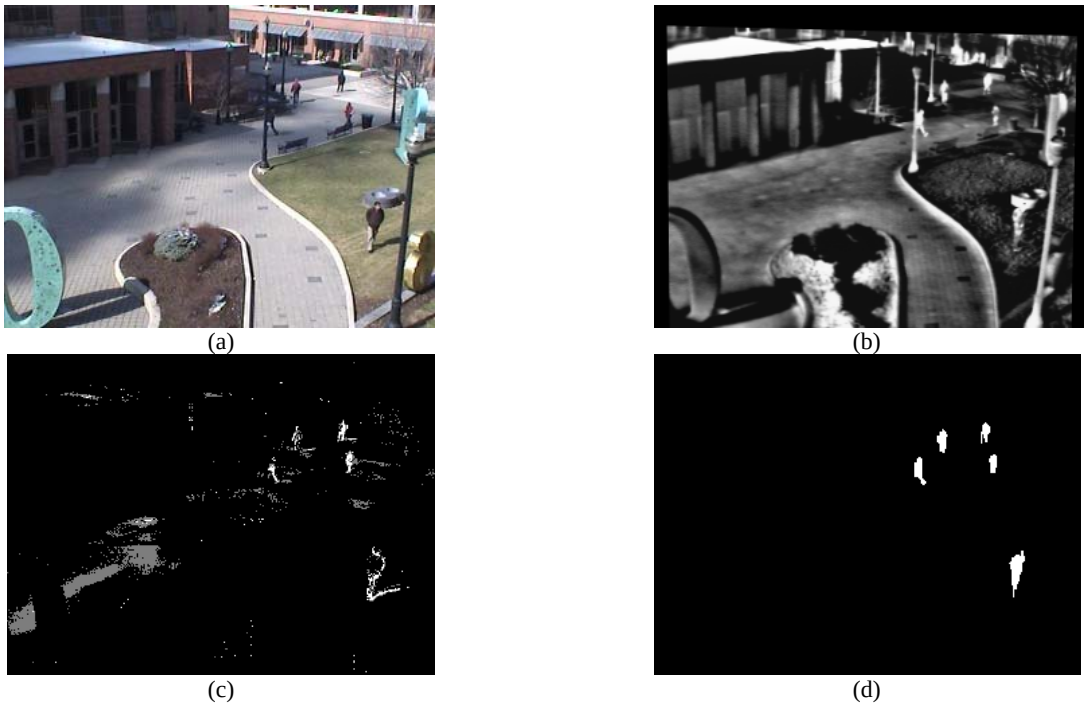
where G_i^t denotes the *i*th ground truth object in frame t, D_i^t denotes the detected object for G_i^t, and N_{mapped}^t is the number of mapped object pairs in frame t. Then, the MODP is computed as

$$MODP(t) = \frac{MOR(t)}{N_{mapped}^t}. \quad (10)$$

The normalized MODP (N-MODP) that gives the detection precision for the entire sequence is defined as

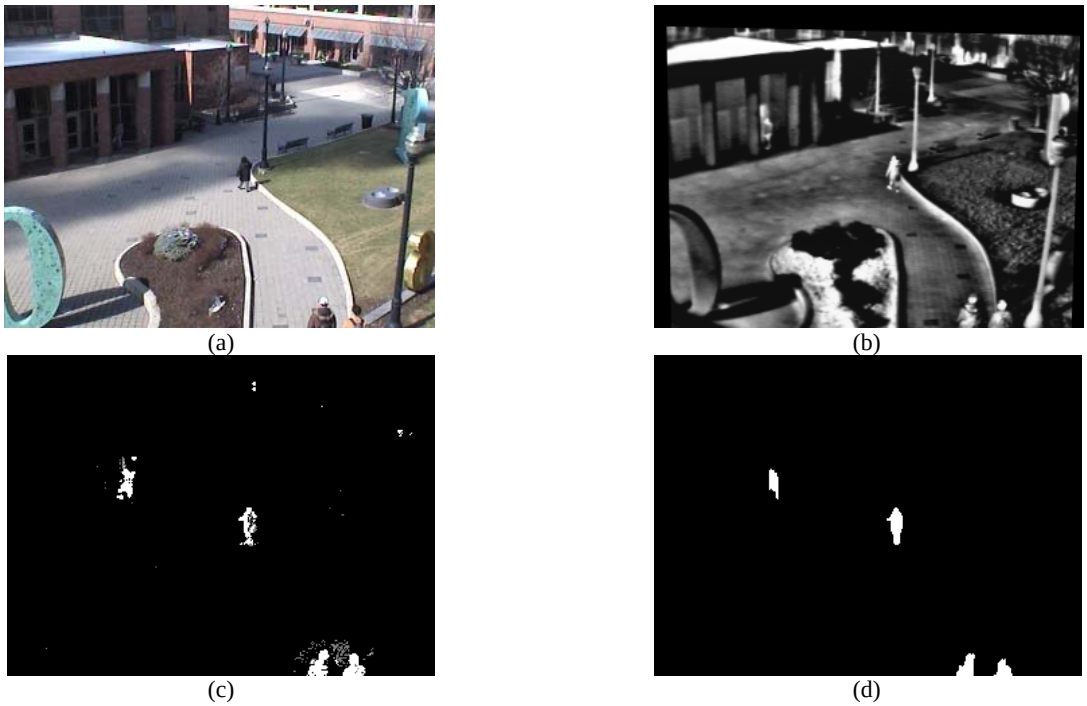
$$N-MODP = \frac{\sum_{t=1}^{N_{frames}} MODP(t)}{N_{frames}}. \quad (11)$$

In our experiments, we randomly pick up 40 frames in



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

Figure 5. Original frames and detection results of frame 491 of Campus B: (a) original frame in the visible domain. (b) original frame in the thermal domain. (c) using Zivkovic method in the visible domain. (d) using JBM-TV method in both domains.



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

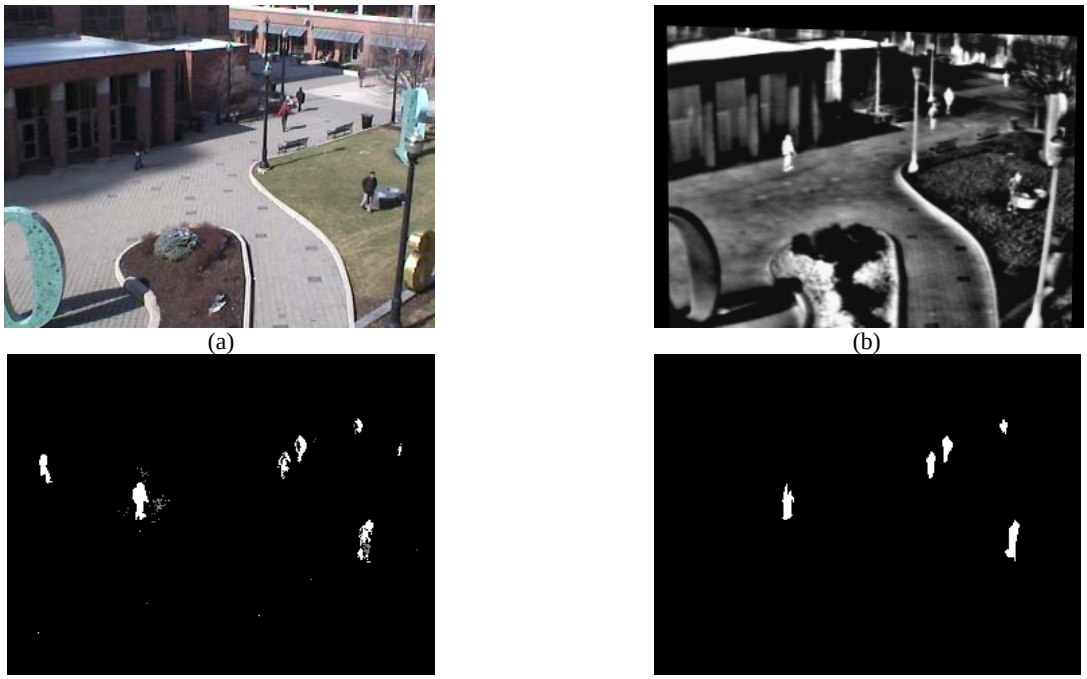
Figure 6. Original frames and detection results of frame 201 of Campus A: (a) original frame in the visible domain. (b) original frame in the thermal domain. (c) using Zivkovic method in the thermal domain. (d) using JBM-TV method in both domains.

each test sequence and their corresponding detection results. Then, we manually marked the ground-truths and the detection results using rectangle. The MODA and MODP can finally be obtained according to the parameters of these marked rectangles.

All the above considered measures attain values in [0,1], and the higher the value, the better the results.

The MODA and MODP for both sequences are shown in Fig. 8 and Fig. 9. The MODA values demonstrate that our method performs well at object level in complex environments. The MODP values show that our method is quite acceptable at pixel level.

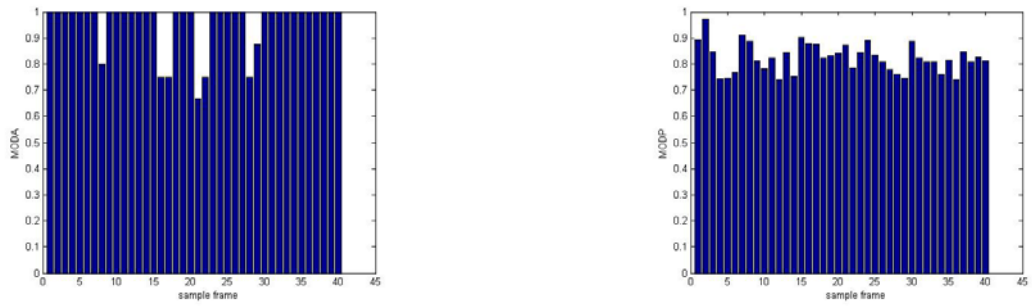
According to Fig. 8, we can calculate that, the N-MODA for Campus A is 0.96, and the N-MODP for Campus A is 0.82. As seen in Fig. 9, the N-MODA for Campus B is 0.95 and the N-MODP for Campus B is 0.81. For these two challenging sequences, our JBM-TV method attains satisfactory results.



(a) (b)

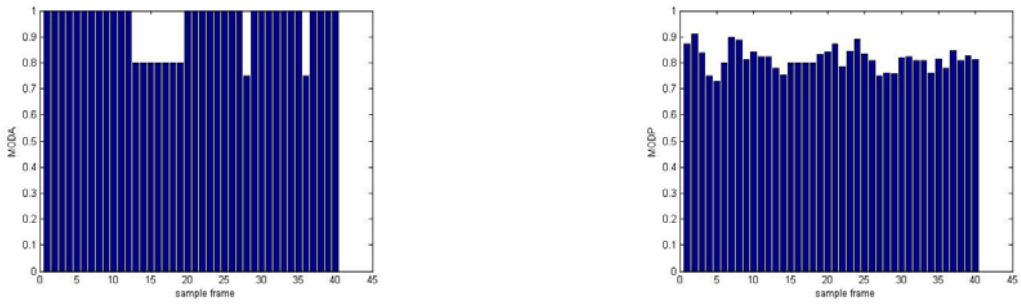
(c) (d)

Figure 7. Original frames and detection results of frame 321 of Campus B: (a) original frame in the visible domain. (b) original frame in the thermal domain. (c) using Zivkovic method in the thermal domain. (d) using JBM-TV method in both domains.



(a) (b)

Figure 8. Quantitative Analysis of JBM-TV method for Campus A: (a) MODA and (b) MODP



(a) (b)

Figure 9. Quantitative Analysis of JBM-TV method for Campus B: (a) MODA and (b) MODP

V. CONCLUSION

In our study, we provide a joint sample consensus background model, which is different from the common methods that first detect foreground respectively and then fuse the detection results. In our background model, object detection and fusion can be executed simultaneously, which lowers the complexity of detection method. Also, it is helpful to extract initial ROIs in the thermal domain, employ TOM mechanism and utilize memoryless update policy. Experimental results demonstrate that JBM-TV method can handle scenes containing dynamic background, gradual and sudden illumination changes. Meanwhile it can eliminate the shadows and halos around the moving objects.

ACKNOWLEDGMENT

This work is supported by the Fundamental Research Funds for the Central Universities (N110323004) and the Natural Science Foundation of Hebei Province under Grant No.F2012501001.

REFERENCES

[1] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vis. Image Underst.*, vol. 106, pp. 162–182, May/June 2007.

[2] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, pp. 255–261.

[3] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 747–757, August 2000.

[4] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, pp. 1151–1162, 2002.

[5] K. Kim, Thanarat H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, pp. 172–185, June 2005.

[6] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 1778–1792, November 2005.

[7] M. Heikkilä and M. Pietikäinen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 657–662, April 2006.

[8] B. P. Hou and W. Zhu, "Fast Human Detection Using Motion Detection and Histogram of Oriented Gradients," *Journal of Computers*, vol. 6, no. 8, pp. 1597–1604, August 2011.

[9] J. Wu, J. Xia, J. M. Chen and Z. M. Cui, "Adaptive Detection of Moving Vehicle Based on On-line Clustering," *Journal of Computers*, vol. 6, no. 10, pp. 2045–2052, October 2011.

[10] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, pp. 1709–1724, June 2011.

[11] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, pp. 2188–2202, November 2011.

[12] J. M. Choi, H. J. Chang, Y. J. Yoo, and J. Y. Choi, "Robust moving object detection against fast illumination change," *Comput. Vis. Image Underst.*, vol. 116, pp. 179–193, February 2012.

[13] N. Armanfard, M. Komeili, and E. Kabir, "TED: A texture-edge descriptor for pedestrian detection in video sequences," *Pattern Recognit.*, vol. 45, pp. 983–992, March 2012.

[14] J. W. Davis and V. Sharma, "Background-subtraction in thermal imagery using contour saliency," *Int. J. Comput. Vis.*, vol. 71, pp. 161–181, February 2007.

[15] M. Bertozzi, et al., "Pedestrian detection by means of far-infrared stereo vision," *Comput. Vis. Image Underst.*, vol. 106, pp. 194–204, May/June 2007.

[16] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Comput. Vis. Image Underst.*, vol. 106, pp. 288–299, May/June 2007.

[17] J. -T. Wang, D. -B. Chen, H. -Y. Chen, and J. -Y. Yang, "On pedestrian detection and tracking in infrared videos," *Pattern Recognit. Lett.*, vol. 33, pp. 775–785, April 2012.

[18] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection," *Pattern Recognit.*, vol. 40, pp. 1771–1784, June 2007.

[19] P. Kumar, A. Mittal, and P. Kumar, "Addressing uncertainty in multi-modal fusion for improved object detection in dynamic environment," *Inf. Fusion*, vol. 11, pp. 311–324, October 2010.

[20] I. Ulusoy and H. Yuruk, "New method for the fusion of complementary information from infrared and visual images for object detection," *IET Image Process.*, vol. 5, pp. 36–48, February 2011.

[21] A. Torabi, G. Massé, and G. -A. Bilodeau, "An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for vide surveillance applications," *Comput. Vis. Image Underst.*, vol. 116, pp. 210–221, February 2012.

[22] Z. Zivkovic and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, pp. 773–780, May 2006.

[23] S. Brutzer, B. Höferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1937–1944.

[24] R. Kasturi, et al., "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 319–336, February 2009.

Guang Han received his B.Eng. and M.Eng. degrees from the School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2005 and 2008, respectively. Now he is a Ph.D. candidate at College of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests include object detection and object tracking based on video sequences.

Xi Cai received her B.Eng. and Ph.D. degrees from the School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2005 and 2011, respectively. Now she works at Engineering Optimization and Smart Antenna Institute, Northeastern University at Qinhuangdao, Qinhuangdao, China. Her research interests include image fusion, image registration, object detection and object tracking.

Jinkuan Wang received the M.Eng. degree from Northeastern University, Shenyang, China, in 1985, and the Ph.D. degree from the University of Electro-Communications, Chofu, Japan, in 1993.

In 1990, he joined the Institute of Space and Astronautical Science, Sagamihara, Japan, as a special member. He was an Engineer with the Research Department, COSEL Company, in 1994. He is currently the President of the Northeastern University at Qinhuangdao, Hebei, China, where he has been a Professor since 1998. He has been a main researcher in several National Natural Science Foundation research projects of China. His main interests are in the areas of intelligent control, adaptive array, wireless sensor networks and image processing.