Gaussian Mixture Model in Dynamic Background of Video Sequences for Human Detection

Bimo Haryo Setyoko
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
bimo_hs@uinsalatiga.ac.id

Edi Noersasongko
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
edi-nur@dsn.dinus.ac.id

Guruh Fajar Shidik
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
guruh.shidik@dsn.dinus.ac.id

Fikri Budiman
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
fikri.budiman@dsn.dinus.ac.id

Moch Arief Soeleman
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
arief22208@gmail.com

Pulung Nurtantion Andono
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
pulung@dsn.dinus.ac.id

Pujiono
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
pujiono@dsn.dinus.ac.id

***Abstract*—Background subtraction is the initial key process of human detection. One of the problems faced by background subtraction is various changes or dynamic backgrounds. It can be caused by the illumination or color so the colors of the background and the foreground look similar. To adjust the background changes, a different Gaussian Mixture Model was proposed in this paper. Performance evaluation was conducted qualitatively and quantitatively by comparing GMM-based background subtraction against standard background subtraction. Measurement is carried out using PSNR (Peak Signal Noise Ratio) and MSE (Mean Square Error). The Average MSE of GMM-based was 3,40 and the average PSNR was 11,80. The Average MSE of standard Background Subtraction was 3,98 and the average PSNR was 8,54. The MSE value of GMM-based background subtraction based is lower than the MSE of standard Background Subtraction and the PSNR value is higher. The experiment result proved that GMM-based background subtraction is promising to segment and extract the foreground in detection.**

Keywords—background subtraction, human detection, gaussian mixture model, dynamic background, GMM-based

I. INTRODUCTION

Moving object segmentation, especially for human detection is the initial key of the basic step for Human Behavior Recognition (HBR) or Human Activity Recognition (HAR) in video processing and computer vision research [1]. Nowadays, Human Activity Recognition (HAR) continues to develop. HAR is increasingly being explored for its use in various ways to assist human activities, including security or surveillance [2], education [3], health [4] [5], sport [6], etc. Because of the many benefits that can be obtained from HAR, it makes HAR still become one of the research trend topics in Computer Vision [7]. Another application of HAR is in the automation of the development of smart environments in buildings, [8], namely smart houses and smart offices [9] [10].

Recognition from video processing has several stages, and each stage is an important process because the results of each stage greatly affect the final result of the recognition. However, the initial key to the recognition process is how to get accurate and reliable segmentation results of each part of the detection process [11]. In other words detection of the objects accurately plays a very important role. Success in detection and segmentation will affect the subsequent processes such as extraction of the features, classification, and recognition.

Background subtraction can be the first step of the whole series of recognition processes. Background subtraction works by isolating and extracting the foreground for further analyzing the video sequence [12]. In applications that require a background as an input to separate the background object and foreground object, a re-initialization is required if there is a change in the background. If not re-initialized, it will cause inaccurate foreground. The principle of background subtraction is the modeling of the background. To update the model of the background, a new frame in the sequences is used by background modeling. But background subtraction faces a problem in various changes or dynamic backgrounds. The changes in light intensity suddenly, illumination, or similar color of the background to the foreground make the background dynamic. The method of Background Subtraction used will have many disadvantages to detect the moving object. To cover this shortcoming, there is an adaptive background models method, namely the Gaussian Mixture Model (GMM).

We proposed GMM-based background subtraction to segment the moving objects and extract the foreground from the dynamic background of video sequences in this paper.

For a more detailed discussion, this paper is organized to provide further explanation of our ideas as follows. Section 2 presented the review of some related studies in the field. Section 3 explains the methodology proposed in this research. The experimental and evaluation results will be shown in

Section 4. And the last, the conclusion and suggestions for future work will be addressed in Section 5.

II. Related Work

Various techniques and methods were developed to obtain the best results for segmentation. However, the results of the methods that have been developed depend on the cases being handled. A method can get good results for certain cases, but not necessarily can get the same results for other cases. This makes it a challenge in itself.

The background subtraction method using a statistical model was proposed by B. N. Subudhi et al. [13]. This model works well with multimodal backgrounds and adaptive parameters.

Another study about the method of threshold determining dynamically and automatically has been proposed by Ka Ki Ng and Edward J. Delp [14]. In this study, they use the intensities of the pixels to determine the threshold. The pixels intensities used are the pixels in the current frame. And the background model will be obtained. A learning rate method is used to update the model of the background.

M. Babaee et al. [15] proposed the Neural Network Model based on background subtraction. With this model, unstable lighting, bootstrapping, and dynamic background issues can be resolved.

Soeleman M. A. Et al. [16], proposed a method to get the foreground. Paralel process for getting adaptive threshold are performed in this study. Mixture of gaussian, Otsu, and gray threshold used in this method.

A Robust foreground segmentation algorithm was proposed by X. Song et al. [17]. This algorithm is used to make the classification between foreground and background. The processes of morphological and multiple thresholds are used in this method. The background subtraction uses mean and standard deviation.

III. Proposed Method

The method for the moving object detection that we proposed, is illustrated in Fig. 1. This method is composed using several steps, first, the video is extracted into sequence frames. Then, the extracted frames will be performed using GMM-based background subtraction.

In the GMM-based background subtraction, Gaussian Mixture is used to reduce inaccuracy that occurs due to changes in the background caused by changes in illumination or contrast intensity. We used a different method for the operation of morphological, erosion, and dilation. Operation of morphological is performed to fill the holes in the foreground object area. Then to reduce the noisy pixel, the erosion technique is performed. The result of this process is the detected binary moving object in the current frame (foreground). Then the binary moving object detected is evaluated qualitatively and quantitatively.

A. Background Subtraction

In videos or images, the background is an image with objects that do not move (static objects). While the moving object on the image is called the foreground. So that an image or video can be analyzed further for certain cases, it is necessary to separate the background and foreground. Background subtraction is one of the methods used to split the background and foreground.

The basic concept of this method is to subtract the background reference with the chosen frame. The steps to segment the foreground using this technique are, the first step is to extract the frames from the video. Then determine the background frame by calculating the median value of several frames that were randomly chosen. Convert the current frame and background frame to grayscale and convert the grayscale to a binary image. After the conversion, subtract the current frame with the frame of the background.

Non-recursive technique in background subtraction that is used in this paper is frame difference. Let's determine the binary foreground as BF.

$$BF\ (x,y,n) = \begin{cases} 1, if\ |I\ (x,y,n) - I\ (x,y,n-1)| \geq \alpha \\ 0, otherwise \end{cases} \quad (1)$$

The threshold (α) can be determined to classify between foreground and background.

B. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a kind of density model. The model consists of Gaussian function components [18]. The components of this function have different weights in producing a multi-model density. GMM is used to model the color of each pixel of the background.

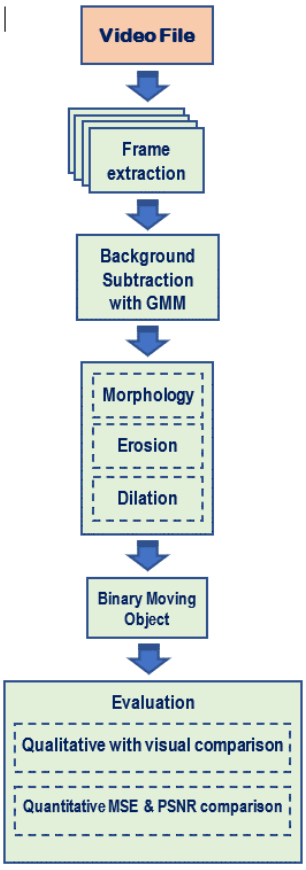


Fig. 1. Proposed Method

Each pixel has its GMM and will process some data. The data that will be processed is obtained from the input in the form of pixel color. The models of GMM are constructed from time-based pixel color data. Then the pixels are divided into 2 groups, the model that reflects the background, and the model that reflects non-background. The number of GMM models will affect the number of background models. In other words, the background model that a pixel has will increase if the amount of GMM models used is larger [19].

The process stages of this method are, matching the input to the distribution and the stage of selecting a distribution that refers to the background. The mathematical formula of the GMM equation model is as follows (2).

$$P(X_t) = \sum_{i=1}^{K} \omega_{it}\, \eta\big(X_t, \mu_{i,t}, \Sigma_{i,t}\big) \quad (2)$$

In formula (2), the number of distributions is denoted as K, while the average value of gaussian at term unit t is denoted as μ, and Σ is the covariance matrix at i threshold on gaussian, and ω is the weight. The below formula is a function of gaussian probability density.

$$\eta\big(X_t, \mu_{i,t}, \Sigma_{i,t}\big) = \frac{1}{(2\pi)^{\frac{n}{2}}\, \Sigma^{1/2}}\, exp^{\frac{1}{2}(X_t-\mu_t)^T\, \Sigma^{-1}(X_t-\mu_t)} \quad (3)$$

$$\omega_{i,t} = (1-\alpha)\omega_{i,t} + \alpha M_{i,t} \quad (4)$$

The Gaussian distribution dimension is denoted as n. The value of n will be grouped into 3 categories of the background model. First, the model is a gray image when n is equal to 1. Second, the model is a color image when n is equal to Σi,t, and third, the model is an RGB image when n is equal to 3. It will match when the value M is 1, and vice versa, the value of M is 0. To update the values of μ and α, the following formula is used.

$$\mu_{i,t} = (1-\rho)\mu_{i,t-1} + \rho X_{i,t} \quad (5)$$

$$\sigma_{i,t}^2 = (1-\rho)\sigma_{i,t-1}^2 + \rho(X_t - \mu_{i,t})^2 \quad (6)$$

$$\rho = \alpha\eta(X_t, \mu_{i,t}, \sigma_{i,t}) \quad (7)$$

$$B = \arg min_b \left(\sum_{k=1}^{b} w_k > C_f\right) \quad (8)$$

Using (4) in the stage of parameter update, the next input is processed using the values of GMM parameters. The values that will be updated are weight, average, and variants. The value of weight will be updated at all times. After the normalization of the weight value, the total weight of all distributions is not more than 1. Then the mean value of the distribution will be updated at all times, and there is a value of the pixel that matches the value of the distribution. Formulas (6) and (7) are used to update the standard deviation value of the distribution at all times, and there is a value of the pixel that matches the distribution. Finally, by using (8) the pixels which match the gaussian model probabilistically, will be classified as background pixels if they are against the background, and their opponents as the foreground.

IV. Experiment and Result

We presented the experimental implementation of our proposed method in this section. Our experiments are conducted on a series of videos, taken from public databases. We use UCF101 datasets that are taken from UCF databases. UCF101 is a human actions database that consists of realistic action videos for an action recognition data set. UCF101 has 101 action categories as an extension of the UCF50 data set. The objective of the system is to segment the human object.

In our experiment, we used Python Anaconda Juptyter Notebook and ran on a mini server computer Dell PowerEdge T30 with Intel Xeon E3-1225 v5, 3.30GHz processor, RAM 16384 MB, and AMD Radeon R5 200 Series Graphic card. To test our proposed method, we use a video of Playing Flute that consists of 235 frames and a video of Playing Violin that consists of 168 frames.

Performance evaluation was conducted qualitatively and quantitatively. Evaluations are carried out by comparing GMM-based background subtraction against standard background subtraction. The measurement of the performance is carried out using PSNR and MSE.

A. Qualitative Evaluation

Qualitative evaluation, is provided using a visual assessment. The assessment is carried out on the detected binary object mask for the test video series [11]. Subjective measurements are provided from the results of foreground detection between standard background subtraction and GMM-based background subtraction. A qualitative comparison between standard background subtraction and GMM-based background subtraction in two scenes of a person's activity is shown in Fig. 2-3. Comparisons are also performed with ground truth images.

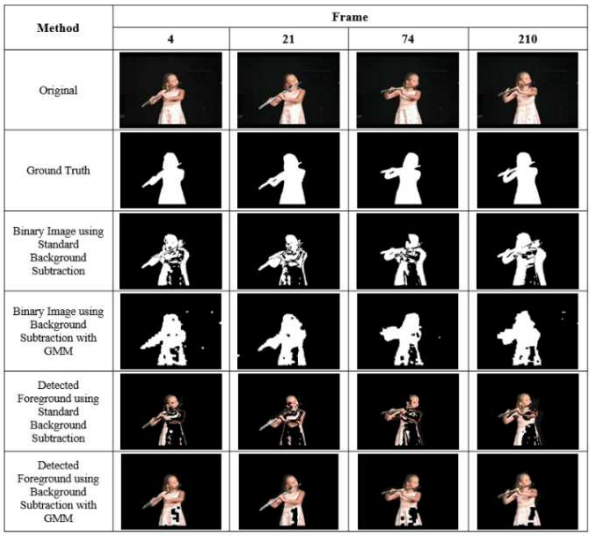


Fig. 2. Qualitative evaluation, comparative result of detection for sequence dataset 1 (playing flute)

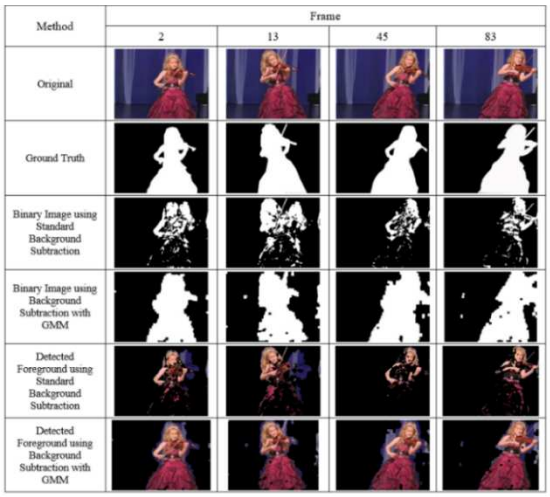


Fig. 3. Qualitative evaluation, comparative result of detection for sequence dataset 2 (playing violin)

B. *Quantitative Evaluation*

The measurement of the performance as quantitative evaluation of our human object detection that we proposed, we use MSE and PSNR. We use MSE and PSNR to calculate the difference value between the segmented foreground and the ground truth image. The use of PSNR can show similarity or homogeneity in the distribution area. The segmentation results will be better if the PSNR value is higher [20]. To calculate the MSE and PSNR, using (9) and (10) respectively could be done.

$$MSE(L,K) = \frac{\sum_{i=1}^{F}\sum_{j=1}^{X}[L(i,j)-K(i,j)]}{FX} \quad (9)$$

$$PSNR\ (L,K) = 10.\log_{10}\left(\frac{max^2}{MSE\ (L,K)}\right) \quad (10)$$

In the equation above, F refers to the ground truth image, and K refers to the segmentation image with size F x X. *max* is the maximum value of the pixel of the image that is possible. Figures 4 and 5, respectively show the MSE and PSNR from the first video (playing flute).

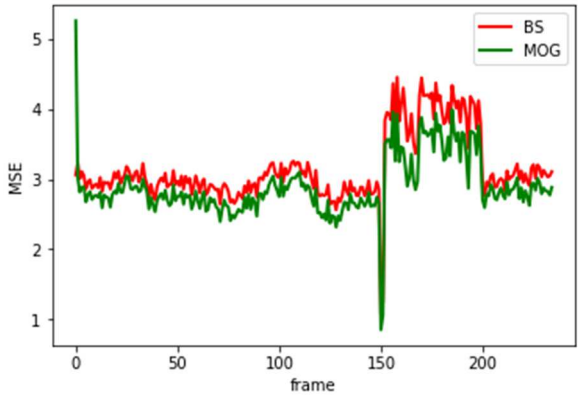


Fig. 4. MSE of standard background subtraction in red line and MSE of background subtraction with Mixture of Gaussian in green line for dataset 1 (playing flute)

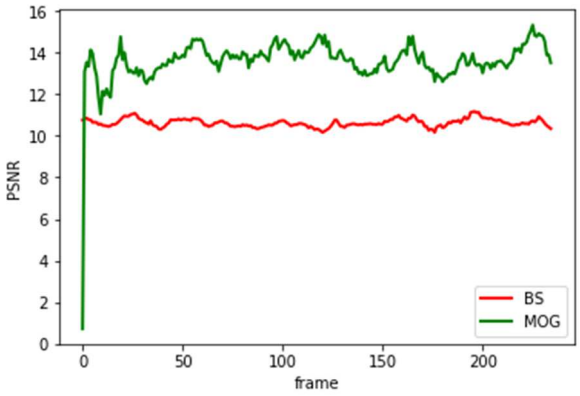


Fig. 5. PSNR of standard background subtraction in red line and PSNR of background subtraction with Mixture of Gaussian in green line for dataset 1 (playing flute)

Figures 4-5, respectively show the MSE and PSNR of the first video, playing flute activity. The graphs in figure 4 show the best performance of background subtraction based on GMM at frame 150. In this frame, MSE= 0,85 and PSNR= 13.56 dB. Averagely, the MSE of GMM-based background subtraction for the first video is 2,89 and the MSE of standard background subtraction is 3,15. The PSNR average of the GMM-based background subtraction and the standard background subtraction respectively is 13,62 dB and 10,63 dB.

Figures 6-7, respectively show the MSE and PSNR of the second video, playing violin activity. The graphs in figure 6 show the best performance of GMM-based background subtraction at frame 87. In this frame, MSE= 7,52 and PSNR= 10.44 dB. Averagely, the MSE of GMM-based background subtraction for the second video is 3,90 and the MSE of standard background subtraction is 4,80. The PSNR average of the GMM-based background subtraction and the standard background subtraction respectively is 9,97 dB and 6,44 dB.

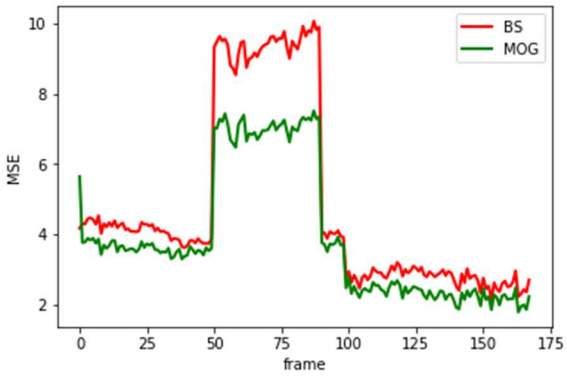


Fig. 6. MSE of standard background subtraction in red line and MSE of background subtraction with Mixture of Gaussian in green line for dataset 2 (playing violin)

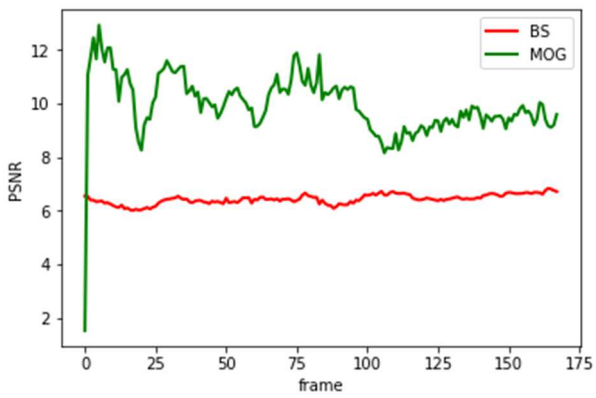


Fig. 7. PSNR of standard background subtraction in red line and PSNR of background subtraction with Mixture of Gaussian in green line for dataset 2 (playing violin)

In this experiment we use several datasets, to test the proposed method. From each dataset used for testing, its performance is measured using MSE and PSNR. So it is necessary to know the average of each MSE and PSNR of the overall datasets. Table 1-2 shows the averages of the MSE and PSNR respectively in this experiment.

Table 1 shows the MSE average value of each dataset and the MSE average value of overall data used for each method. The MSE average value of overall datasets using standard background subtraction is 3,98 and the MSE average value of overall datasets using GMM-based background subtraction is 3,40.

Table 2 shows the PSNR average value of each dataset and the PSNR average value of overall data used for each method. The PSNR average value of overall datasets using standard background subtraction is 8,54 db and the PSNR average value of overall datasets using GMM-based background subtraction is 11,80.

From the result of the experiment, shows that the PSNR of GMM-based background subtraction is higher than the PSNR of standard background subtraction, it means that the result of the detection will be better [20].

TABLE I. AVERAGE PERFORMANCE EVALUATION WITH MSE

Method	MSE Average		
	Dataset 1	***Dataset 2***	***Overall***
Standard Background Subtraction	3,15	4,80	3,98
Background Subtraction based on GMM	2,89	3,90	3,40

TABLE II. AVERAGE PERFORMANCE EVALUATION WITH PSNR (DB)

Method	PSNR Average		
	Dataset 1	***Dataset 2***	***Overall***
Standard Background Subtraction	10,63	6,44	8,54
Background Subtraction based on GMM	13,62	9,97	11,80

V. CONCLUSION

Getting the detection result of moving objects accurately on recognition plays important role. Various techniques and methods were developed to obtain the best results to get accurate detection. The different technique of foreground extraction for moving object detection especially human detection by using background subtraction was presented in this paper. Different techniques of morphology, erosion and dilation used in this process.

The real practice of this experiment can be applied to human detection for action recognition, using video from CCTV surveillance cameras. The results of human detection through surveillance cameras can then be further analyzed for various applications.

The MSE average value of standard background subtraction is 3,98 and the MSE average value of GMM-based background subtraction is 3,40. The PSNR average value of standard background subtraction is 8,54 db and the PSNR average value of GMM-based background subtraction is 11,80 db. The result of the experiment shows the performance of the algorithm to detect human objects based on GMM is better than the standard Background Subtraction. The proposed method produced higher PSNR and lower MSE compared to the standard background subtraction. In our next work, we would perform different method in threshold selection technique, such as using the deep learning method.

REFERENCES

[1] M. A. Soeleman, M. Hariadi, and M. H. Purnomo, "Adaptive threshold for background subtraction in moving object detection using Fuzzy C-Means clustering," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, pp. 1–5, 2012, doi: 10.1109/TENCON.2012.6412265.

[2] L. Alawneh, T. Alsarhan, M. Al-Zinati, M. Al-Ayyoub, Y. Jararweh, and H. Lu, "Enhancing human activity recognition using deep learning and time series augmented data," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 12, pp. 10565–10580, 2021, doi: 10.1007/s12652-020-02865-4.

[3] M. Y. Ali, X. De Zhang, and M. Harun-Ar-rashid, "Student activity detection using deep learning with YOLOV3," *Indones. J. Electr. Eng. Informatics*, vol. 8, no. 4, pp. 757–769, 2020, doi: 10.11591/ijeei.v8i4.2585.

[4] P. Zhao and M. Wang, "Mobile behavior trusted certification based on multivariate behavior sequences," *Neurocomputing*, vol. 419, pp. 203–214, 2021, doi: 10.1016/j.neucom.2020.08.003.

[5] G. Cicirelli, R. Marani, A. Petitti, A. Milella, and T. D'orazio, "Ambient assisted living: A review of technologies, methodologies and future perspectives for healthy aging of population," *Sensors*, vol. 21, no. 10, 2021, doi: 10.3390/s21103549.

[6] J. Liu, "Convolutional Neural Network-Based Human Movement Recognition Algorithm in Sports Analysis," *Front. Psychol.*, vol. 12, no. June, pp. 1–7, 2021, doi: 10.3389/fpsyg.2021.663359.

[7] H. B. Zhang *et al.*, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, pp. 1–20, 2019, doi: 10.3390/s19051005.

[8] A. K. Langroodi, F. Vahdatikhaki, and A. Doree, "Activity recognition of construction equipment using fractional random

forest," *Autom. Constr.*, vol. 122, no. April 2020, p. 103465, 2021, doi: 10.1016/j.autcon.2020.103465.

[9] J. Wan, M. J. O'Grady, and G. M. P. O'Hare, "Dynamic sensor event segmentation for real-time activity recognition in a smart home context," *Pers. Ubiquitous Comput.*, vol. 19, no. 2, pp. 287–301, 2015, doi: 10.1007/s00779-014-0824-x.

[10] J. Rafferty, C. D. Nugent, J. Liu, and L. Chen, "From Activity Recognition to Intention Recognition for Assisted Living Within Smart Homes," *IEEE Trans. Human-Machine Syst.*, vol. 47, no. 3, pp. 368–379, 2017, doi: 10.1109/THMS.2016.2641388.

[11] Y. Zhang, W. Zheng, K. Leng, and H. Li, "Background Subtraction Using an Adaptive Local Median Texture Feature in Illumination Changes Urban Traffic Scenes," *IEEE Access*, vol. 8, pp. 130367–130378, 2020, doi: 10.1109/ACCESS.2020.3009104.

[12] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, p. 107561, 2020, doi: 10.1016/j.patcog.2020.107561.

[13] B. N. Subudhi, S. Ghosh, S. C. K. Shiu, and A. Ghosh, "Statistical feature bag based background subtraction for local change detection," *Inf. Sci. (Ny).*, vol. 366, pp. 31–47, 2016, doi: 10.1016/j.ins.2016.04.049.

[14] K. K. Ng and E. J. Delp, "Object tracking initialization using automatic moving object detection," *Vis. Inf. Process. Commun.*, vol. 7543, p. 75430M, 2010, doi: 10.1117/12.839126.

[15] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, 2018, doi: 10.1016/j.patcog.2017.09.040.

[16] M. A. Soeleman *et al.*, "Adaptive threshold for moving objects detection using gaussian mixture model," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 1122–1129, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14878.

[17] X. Song, J. Chen, C. He, and X. Zhou, "A robust moving objects detection based on improved Gaussian mixture model," *Proc. - Int. Conf. Artif. Intell. Comput. Intell. AICI 2010*, vol. 2, no. 3, pp. 54–58, 2010, doi: 10.1109/AICI.2010.134.

[18] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 246–252, 1999, doi: 10.1109/cvpr.1999.784637.

[19] W. Kurniawan, S. Ibrahim, and M. Sulistyo, "People detection and tracking methods for intelligent surveillance system," *AIP Conf. Proc.*, vol. 2217, no. April, 2020, doi: 10.1063/5.0001022.

[20] K. Bhoyar and O. Kakde, "Color Image Segmentation Based on Jnd Color Histogram," *Int. J. Image Process.*, vol. 3, no. 6, pp. 283–292, 2010.