ICIC Express Letters
Part B: Applications
Volume **12**, Number **1**, January **2021**

 ISSN 2185-2766
pp. **59–66**

A STUDY ON DETECTING VIOLENCE USING IMAGE PROCESSING TECHNOLOGY

Shuhei Misawa and Thi Thi Zin

Graduate School of Engineering
University of Miyazaki
1-1 Gakuen Kibanadai-nishi, Miyazaki City 889-2192, Japan
hl15044@student.miyazaki-u.ac.jp; thithi@cc.miyazaki-u.ac.jp

Received April 2020; accepted July 2020

Abstract. *In recent years, many security cameras have been installed for crime prevention in downtown areas and public facilities. These cameras have greatly contributed to crime prevention and criminal identification. However, the large number of installed cameras is problematic due to difficulties in manually monitoring and detecting violence and crime in real time, as well as in finding specific video footage recording the incidents. This paper describes the use of the background difference method in extracting human regions from data obtained using security cameras. In addition, the paper describes a method of detecting violence using features such as speed and moving distance after contact. Using video footage from seven data sets, these methods have been experimentally evaluated, confirming a high detection rate for incidents involving two people side by side.*

Keywords: Violence detection, Surveillance camera, Background subtraction method, Shadow removal, Feature extraction

1. **Introduction.** In recent years, the number of installed security cameras has been increasing due to improvements in security awareness, with the number of security cameras in Japan now exceeding 3 million [1]. As one justification for this increase, security cameras are very useful for police investigations. For example, when a violent incident occurs, security cameras in the area can confirm the incident, help identify the criminal, and contribute to solving the crime. However, of the many security cameras installed, few function to prevent or report an incident. At present, most surveillance systems involve human beings manually monitoring images. With such manual processes, oversight and labor costs are an issue, not to mention detection failures due to the fatigue resulting from monitoring over long periods of time [2]. A possible solution is to automate the process of detecting and reporting suspicious behavior and actual incidents of violence or crime to police and management. In addition, it is possible to reduce the number of crimes by sounding an alarm when detecting suspicious activity or criminal incidents. In this study, we aim to develop a system that detects violence and crime in real time, and extracts only video scenes in which incidents have occurred. This will greatly assist in pursuing criminals and solving crimes.

The proposed method specifies a human region in each video frame obtained by an RGB camera, and detects a violent act by using feature quantities extracted from the region. After that, we extract video images from before and after the occurrence of violence. The proposed method does not use machine learning. Though machine learning is often used in research on still images, the amount of data required for training is huge, especially for moving images. Moreover, the way that violence is perpetrated varies widely between individuals, and the technology for machine learning is inadequate for detecting such

DOI: 10.24507/icicelb.12.01.59

individual differences. In addition, as RGB cameras are the most commonly used for security, our study used RGB cameras instead of 3D cameras.

This paper is composed of five sections. Section 2 describes the proposed method. Section 3 describes the content of the experiment and the evaluation method. Section 4 discusses experimental results and considerations. Finally, Section 5 provides conclusions and future prospects.

2. **Proposed Method.** A flowchart for detecting violence is shown in Figure 1. First, human regions are extracted based on the input data. Next, the violent behavior is detected from the extracted result. This chapter describes the extraction of human regions and of features according to the flow shown in Figure 1.

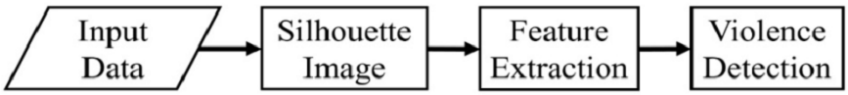

FIGURE 1. Flowchart for proposed method

2.1. **Human region extraction.** In order to detect violence or recognize human activity, it is necessary to extract human regions in the acquired frame. Accordingly, human regions are extracted from the input image by removing background, shadows, and noise. In this section, we describe a method for extracting human regions.

2.1.1. *Sharpening.* Since the read frame is resized, shape information is reduced. Therefore, it is possible to recover information lost in the resizing process by emphasizing changes in luminance values. The process of obtaining a clear image by emphasizing such changes in luminance values is called sharpening [3]. This process is applied to both the background image and the input image.

2.1.2. *Background subtraction method.* The background subtraction method involves extracting the human region (foreground) from the image by comparing areas in which the shape of a person is captured (input image) with areas in which no such shape exists (background) [4]. Furthermore, multiple color spaces are used in image processing, and the background subtraction method uses two color spaces in a two-step process to extract an accurate human region.

A gray scale image is used as the first color space. A gray scale image represents a color image using 256 steps of luminance values from 0 to 255. If the difference in luminance value exceeds a threshold value at the same coordinates in the two images, it is output as a foreground area. An RGB image is used as the second color space. The RGB image also has 256 steps of luminance values from 0 to 255 for each RGB color (Red, Green, Blue) and is capable of expressing 256^3 colors. The difference between the Euclidean distance of the input image and the background image in the RGB color space is determined, and when the value exceeds a threshold, it is output as a foreground area. Then, a logical sum is calculated with respect to the results obtained by the two steps of the background subtraction method, and the result is set as the foreground.

2.1.3. *Shadow processing.* In processing up to this point, both the human region and the accompanying shadow area are extracted. The shadow area is noise and must be deleted to extract the human region. The luminance value of the input image is calculated from the luminance value of the background image. At this time, the shadow area has a smaller amount of change than the human region. The shadow area is detected by using this feature and Equation (1). Assuming that the pixel value of the input image is $\alpha(x, y)$, and the pixel value of the background image is $\beta(x, y)$, the change rate $\gamma(x, y)$ is calculated. All these values are only for the luminance.

$$\gamma(x,y) = \frac{\beta(x,y) - \alpha(x,y)}{\beta(x,y)} \tag{1}$$

Two major types of shadows exist: light shadows and dark shadows. If shadows are deleted using only one threshold for the change rate, the human region might be mistaken as part of the shadow area. Therefore, shadows are deleted in stages, each using a different range of threshold values for the change rate: $0.05 \leq \gamma < 0.20$, $0.20 \leq \gamma < 0.275$ and $0.275 \leq \gamma < 0.325$. This three-stage process removes shadows from the image.

First, as the weakest shadow range, only shadows for which γ falls within the range of $0.05 \leq \gamma < 0.20$ are removed. However, since this is the widest of the three ranges, it includes and outputs human regions that are not shadow areas. Therefore, inadvertent removal of human regions is prevented by defining noise with an area smaller than the threshold as noise which is not a shadow area. In this image, the difference is calculated with respect to the image obtained by the background subtraction method, and because of the possibility of removing more of the human region, the shadow in the range of $0.05 \leq \gamma < 0.20$ is removed through morphological processing.

Next, the shadow in the range of $0.20 \leq \gamma < 0.275$ is removed. This range is not as wide as that used for removing the first shadow, so the extracted shadow area is smaller. Therefore, the shadows on the individual ranges are not subjected to noise removal in this area. The image obtained by removing shadow in this range is subtracted from the image previously obtained by removing the shadow in the range of $0.05 \leq \gamma < 0.20$. The human regions which have been removed by this shadow removing process can be recovered using morphological operations. The next step removes shadows in the range of $0.05 \leq \gamma < 0.275$. Finally, shadows in the range of $0.275 \leq \gamma < 0.325$ are removed. This step involves the same processing as that used for the second range. Therefore, shadow areas in the range of $0.05 \leq \gamma < 0.325$ can be removed. Moreover, although shadows with a value more than this exist, since there is a high possibility that a human region has been erroneously detected as a shadow, only shadow processing is performed for the ranges discussed up to this point. Along with that, removing areas as noise is performed where the label is smaller than that of the human area.

2.1.4. *Noise removal.* Much noise remains in the image after background subtraction and shadow processing. After areas are calculated by the labeling process for all objects including noise, objects below the threshold are eliminated as noise [5]. Set this threshold slightly smaller than that used for the human region.

2.1.5. *Morphological processing.* When performing noise processing, a human region might be removed as noise by mistake. Morphological processing is used to recover a lost human region using its complement. Morphological processing mainly refers to simple processing that acts on figures in binary images. Examples of basic morphological processing include contraction processing and expansion processing. Two types of morphological processing are used in the proposed system, majority processing and closing processing [6]. The two processes are described below.

First, majority processing involves setting the central pixel to 1 when five or more pixels in the neighboring three rows columns are 1. Majority processing is good for noise removal.

In closing processing, expansion processing is performed on white pixels, followed by contraction processing. As an advantage of this process, it can be complemented when holes form in the human region by noise removal or the like, and thin portions disappear, such as arms and necks. The image obtained as a result of using the background subtraction method with shadows removed is defined as a human region. Features are extracted based on this image.

2.2. **Background update and background estimation.** When using the background subtraction method in real time, changes in illumination due to weather or the like reduce the accuracy of extracting human regions. Therefore, it is necessary to update the background image sequentially. In addition, background estimation is also required to suppress noise caused by camera failures or sudden changes in illumination.

In this case, errors such as noise are minimized in background estimation by adopting the median value of each coordinate for an image for one second (15 sheets at a frame rate of 15 fps). Next, background updating is performed using the acquired background estimation image. However, since updating may be performed with the background estimation image as the background, the background image may include the human region. Therefore, the non-human region is updated by using Equation (2). I is an input image, B is a background image, and S is a silhouette image used for background updating. Update the background based on these [7]. The silhouette image used is an image obtained by performing expansion processing on the human region extracted in Section 2.1 because the human region might be extracted as a background.

$$B_t(x,y) = \begin{cases} I_t(x,y) & \text{if } S(x,y) = 1 \\ B_t(x,y) & \text{if } S(x,y) = 0 \end{cases} \tag{2}$$

2.3. **Feature extraction.** From the human region extracted in Section 2.1, the kind of action the person in the frame takes is identified.

2.3.1. *Increase or decrease in area of the human region.* Basically, violence is not identified unless both a perpetrator and a victim are present. In other words, when the area in the frame is less than the area required for two or more persons, it is not judged to be violence.

2.3.2. *Increase or decrease of the label of the human region.* It is important to know whether people are in contact with each other to judge whether violence has occurred. For example, when persons are not yet in contact in image processing, the number of labels is 2. However, when two persons come in contact, they are treated as one object, and the number of labels decreases to one. Using this feature, it is determined whether or not people have touched each other. When a person leaves the frame of the camera, the number of labels also decreases, but as described in Section 2.3.1, this is not judged to be violence without the presence of two or more people.

2.3.3. *Contact speed.* Violence should be differentiated from nonviolence, such as a handshake. If contact is made quickly, violence is probable. Also, when human beings are shaken or kicked, arms and legs flail out from the body. Detecting this occurrence involves removing the torso from the human region extracted as discussed in Section 3.

First of all, the number of white pixels in each column of the human region is calculated. If the sum is equal to or less than the threshold, deletion is performed to obtain a rough image of a human torso. However, depending on the movement of the body, the center of the torso may be deleted as well, so the closing process is performed to compensate. A subjected to such processing is defined as a torso image.

Since the human regions and torso images are well defined, it is possible to define parts protruding from the body. Next, the amount of movement across the width of the bounding box in that portion between frames is taken as the amount of movement. By multiplying this value by 15, which is the frame rate used, it is possible to measure how quickly the part protruding from the body has moved in one second. When this value exceeds the threshold, violent behavior is a possibility.

2.3.4. *Extraction of human center of gravity.* A judgment of violence based only on speed of contact increases the possibility of false detection. However, it is possible to reduce the number of false positives by recording what happens several frames after the suspected incident. This is done by measuring how far the center of gravity moves in 1 second after

the incident. For example, consider the case in which a handshake is mis-detected as violence. In a few cases, movement occurs during or after shaking hands, but this will not be considered violence if the distance moved per second is less than a threshold.

Next, the method of extracting the center of gravity will be described. When two human regions overlap, the number of labels in the frame does not match the number of human regions, and the centers of gravity cannot be accurately extracted. In this case, first use the torso image obtained as described in Section 2.3.3. Then use the height of the average center of gravity of adult men, which is a distance from the floor of about 56% of total height [8]. The waist image comprises several pixels above this height. The center of the extracted waist image is the center of gravity of the human region.

2.4. **Violence detection.** Violence detection is performed based on the quantity of features extracted as described in Section 2.3. A flowchart for violence detection is shown in Figure 2.

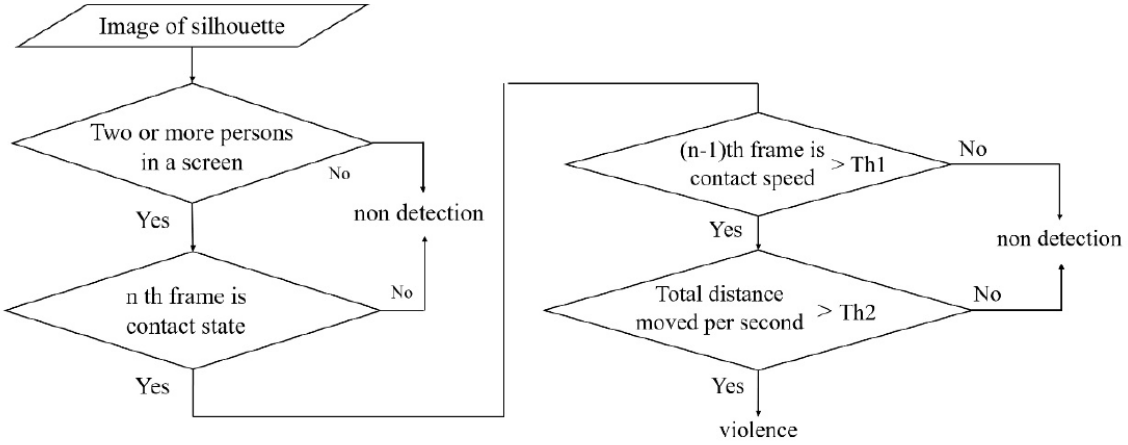

FIGURE 2. Flowchart for violence detection

Violence is detected when a recorded situation satisfies the following criteria. First, violence detection is only initiated when two or more human regions occur in a frame. If the number of labels decreases when two persons touch and when the speed of contact exceeds a threshold, then the incident is recorded as a candidate for violence. Next, the total distance the center of gravity moves in 15 frames is calculated based on the number of frames recorded as candidates for violence. If this value exceeds the threshold, violence is detected. Violence is still violence regardless of whether the movement involves kicks with a foot or strikes with a fist.

3. **Experimental Environment and Evaluation Method.** The data set used for this study was the "ICPR 2010 Contest on Semantic Description of Human Activities" [9]. This data set has 10 videos that include violent acts, but videos with a camera offset or lacking a background were not used for this study. Each video has several acts of violence and non-violence within a few minutes, involving several people of different heights in various clothing. For the lighting conditions, the weather was cloudy during recording sessions, but the brightness and the depth of the shade varied.

The purpose of this paper is to describe an automatic violence detection method that also extracts those portions of video recording relevant to the incident. This method was evaluated according to whether it could accurately detect violent acts in the data set.

4. **Experimental Results and Considerations.** This section describes experimental results using the proposed method, and provides considerations.

4.1. **Results.** The proposed method of detecting violence was performed on each of eight videos in the data set, and the accuracy of violence detection was calculated. In videos 1 through 4, only two individuals appear side by side with respect to the camera. In videos 5 through 7, two or more individuals move in multiple directions. Finally, video 8 did not include any violence, and the author checked whether the proposed method mis-detected violence in a scene without violence. The detection results are shown in Table 1. Here, the number of violent incidents indicates the number of times violence occurs in each video, and the number of correct answers indicates the number of times violence was correctly detected. The undetected number indicates the number of times violence occurred but was undetected, and the number of false detections indicates the number of times violence was mistakenly detected. Accuracy is calculated using total the number of violent incidents and the number of correct detections.

Table 1. Experimental results

Video	Number of violences	Number of correct answers	False positives	Not detected	Accuracy
1	3	3	0	0	100%
2	3	2	0	1	66.7%
3	2	2	0	0	100%
4	3	2	0	1	66.7%
5	2	1	2	1	50%
6	3	3	2	0	100%
7	6	3	1	3	50%
8	0	0	0		100%
Total	22	16	5	6	
				Average accuracy	79.18%

4.2. **Considerations.** From Table 1, false positives were seen in videos 5 through 7, in which two or more people are moving in multiple directions. In addition, detection failures occurred in videos 1 through 4, in which only two individuals appeared side by side with respect to the camera. From this, the cause of both non-detection and the false detection will be described.

First of all, it is conceivable that inaccurately measuring contact speed was the cause of non-detection. In fact, a violent act is performed, but the foot was not extracted as a portion protruding from the body, and the contact speed could not be measured well. As a result, the incident was incorrectly judged to be nonviolent.

Next, false detection is considered to be mainly caused by vertical contact. The contact speed exceeds the threshold as a person passes behind a person appearing in front. In this case, since the person is passing at a high speed, the moving distance after contact becomes a large value. This was mis-detected as violence. In addition, a scene occurs in which the human region is not accurately extracted because the shadow processing method was ineffective in a dark shadow. The failure to extract an accurate human region in this way is also a factor in false detection.

5. **Conclusion and Future Outlook.** In this paper, we evaluated the proposed method using multiple data sets for violence detection using an RGB camera. As a result, researchers confirmed its usefulness in detecting violence when the two individuals appear side by side in a frame. However, when two individuals appear vertically or when a plurality of persons appear in a frame, many false positives and non-detections occur. There are two major causes for this.

First of all, effective processing has not been developed for the case in which two people overlap vertically in a frame. If two or more persons overlap in a vertical direction, the label for the human region becomes one long vertical label, and the two persons cannot be evaluated separately. As a possible solution, when it is determined that two or more persons overlap vertically, they might be separated using color information from their clothes.

Second, effective processing has not been developed for dark shadows. Currently, shadow processing is only effective for relatively light shadows. Human regions extracted from dark shadows contain an excessive amount of noise. This problem cannot be solved by simply raising the threshold value used in the shadow processing shown in Equation (1). Therefore, it is necessary to formulate a new expression for shadow processing. Currently, it is considered that more robust shadow processing can be performed using Equation (3) [10].

$$SP_k(x,y) = \begin{cases} 1 & \text{if } \alpha \le \frac{I_k^V(x,y)}{B_k^V(x,y)} \le \beta \wedge \left(I_k^S(x,y) - B_k^S(x,y)\right) \\ & \quad \le \tau_S \wedge \left|I_k^H(x,y) - B_k^H(x,y)\right| \le \tau_H \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$SP_k(x,y)$ in Equation (3) indicates a shadow. The range of conditions on the right side is 1 and the others are 0, obtaining a binary image of only the shadow region. Important in this equation are the four parameters on the right side. Determining the most versatile of these four parameters is problematic, but the authors consider this an important question.

We hope to improve our method using a more suitable data set in the future. We think that there are the following three advantages by this.

The first point is that it is possible to consider what kind of characteristics violence has, leading to the discovery of new characteristic quantities.

Second, we think that it is necessary to improve the contents of the data set because the data set currently being used is a violent act that is a little distant from reality. Specifically, the perpetrator often kicks or beats slowly so as not to hurt the other party. Also, the victim role may not move at all even though he is scolded because he knows that the perpetrator role is acting. Therefore, there is a problem that it is difficult to obtain accurate figures and feature quantities as violence. We believe that increasing the data set can solve this problem.

The third point is that the versatility of the proposed violence detection algorithm can be confirmed, and new challenges can be known simultaneously. Although the problems still remain, we think that it will approach more accurate ones for practical use by solving one by one.

REFERENCES

[1] *https://www.nikkei.com/article/DGXNZO43502310X00C12A7EL1P00/*, Accessed on 23 Jan., 2019.

[2] T. Tatsuya, *A Study on Detection of Suspicious Persons for Intelligent Monitoring System*, Master Thesis, Department of Engineering in Energy Course, Graduate School of Engineering, University of Miyazaki, 2016.

[3] *http://www.wakayamau.ac.jp/∼chen/education/image/2002/filter.pdf*, Accessed on 23 Jan., 2019.

[4] J. Kurohane, *Feature Extraction for Detection of Violent Behavior between Two Parties*, Graduation Thesis, Faculty of Engineering Electrical System Engineering, University of Miyazaki, 2015.

[5] H. Tsushita, *A Study on Detection of Abnormal Behavior by a Surveillance Camera Image*, Master Thesis, Department of Engineering in Energy Course, Graduate School of Engineering, University of Miyazaki, 2016.

[6] K. Yamada, *Research on Tracking of Target Object and Automatic Determination of Human or Not*, Graduation Thesis, Faculty of Engineering Electrical System Engineering, University of Miyazaki, 2016.

[7] A. Kawano, *A Study on Violence Behavior Detection System between Two Persons*, Graduation Thesis, Faculty of Engineering Electrical System Engineering, University of Miyazaki, 2018.

[8] *https://www.hus.ac.jp/∼gisisougu/rikigakukiso002-jyusin.html*, Accessed on 23 Jan., 2019.
[9] M. S. Ryoo, C.-C. Chen, J. K. Aggarwal and A. Roy-Chowdhury, An overview of contest on semantic description of human activities (SDHA) 2010, *ICPR 2010: Recognizing Patterns in Signals, Speech, Images and Videos*, pp.270-285, 2010.
[10] R. Cucchiara, C. Grana, M. Piccardi et al., Improving shadow suppression in moving object detection with HSV color information, *2001 IEEE Intelligent Transportation Systems*, pp.334-339, 2001.