Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
14/1 (2021), 17-25. DOI: http://dx.doi.org/10.21609/jiki.v14i1.900

Automatic Image Self-Enhancement for Multi-Scale Spectral Residual on Low-Resolution Video

[1]Arwin Halim, [2]Sunaryo Winardi, and [3]Erlina Halim

[1,2,3]Informatics Engineering Department, STMIK Mikroskil, Jl. Thamrin No 124, Medan, 20211, Indonesia

E-mail: arwin@mikroskil.ac.id, sunaryo.winardi@mikroskil.ac.id, erlina.halim@mikroskil.ac.id

Abstract

Multi-Scale Spectral Residual technique is used to reduce the search area in an image. However, this technique relies on image salience from the capture device. The aim of this study is to obtain a better search area with image enhancement to detect human objects on low resolution video. Enhanced image uses only pixels in each frame of the video using the Exposure Fusion Framework. The dataset is an artificial video obtained from a room with low resolution CCTV. This study compares the detection results before and after applying image enhancement on MSR. We are adopting Linear-SVM based human detection with Histogram of Gradient (HOG) features as a test case. Human detection was evaluated using precision, recall, f-score rate and validated by leave-one-out cross validation. The results show that enhanced images can improve overall performance by 64.46% compared to the original video in human detection on low resolution video, with an increase in recall of 3.21%

Keywords: *Exposure Fusion Framework, Human Detection, Multi-Scale Spectral Residual*

1. Introduction

Multi-scale spectral residue (MSR) is a method of reducing image search space that focuses on more distinctive image regions [1]. MSR can improve the object detection process in searching for the Histogram of Oriented Gradients (HOG) features using Support Vector Machine (SVM) classifier. MSR is able to detect objects three to five times faster with the same detector. However, implementing MSR on low resolution video is a challenge. Winardi et al [2] show that MSR in human detection was affected by brightness of the image. When the brightness is lower, human detection is less effective. To improve results, it is necessary to improve image quality for the better.

Various techniques have been proposed to enhance image quality. One of the techniques is contrast enhancements. There are many contrast enhancement techniques such as Histogram Equalization and its modified methods [3]. Improved contrast can reveal less visible regional information in the image. However, the concept of a good results remains unclear in contrast enhancement. Guo [4] tries to produce lighting maps from images using real-world images using Low-light IMage Enhancement (LIME). However, the obstacle is that we do not necessarily have references to low-light enhancement algorithms for finding high or low contrast areas. Different images in lighting can be used as a reference for the contrast enhancement algorithm. Ying et al. [5] attempts to fuse images with enhanced lighting with the Exposure Fusion Framework (EFF). Some areas that are less light become well exposed. It keeps the area with good lighting unchanged and increase the area with less light. Also, the enhanced contrast area is quite consistent with the initial reference image.

This study tries to detect humans in low resolution video by proposing to enhance the quality of each frame in the video. Enhanced image will be preprocessed with MSR to reduce the search area of human detection. MSR has tried to improve image quality with Contrast Limited Adaptive Histogram Equalization (CLAHE) normalization but the results are not optimal for video [2]. CLAHE works on the minor regions in the image rather than the whole image [3]. In this case, we combine it with EFF which focuses on resolving the problem of enhancement in the whole image. With less contrast and light distortion, EFF can produce results [5]. There are many techniques for detecting humans, such as neural network classifiers [6], SVM classifiers [7], template matching [8] which uses templates to model the human body, combination of features, classifier and steps for pedestrian detection [9], etc. Every human detection technique has its advantages. In this study, we detected humans using the HOG feature with the Linear-SVM classifier as a test case.

2. Literatur Review

A. Multi-Scale Spectral Residual

Multi-scale Spectral Residual (MSR) is a method for decreasing image search space by arranging regions based on visual importance [1]. The visual image on the MSR is very reliant on the identification of saliency. Saliency is used in the MSR to determine the need for further examination in certain image regions. MSR is used to evaluate parts of an image that do not require further detection/processing (such as detecting objects in an image), so they can be removed to speed up the search. The whole MSR process can be illustrated in Figure 1.

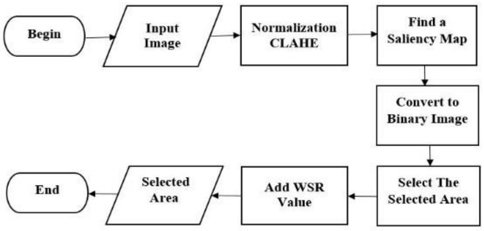


Figure 1. Multi-scale spectral residual process

The MSR process begins with normalizing the contrast of the input image so that the object is more clearly defined. Contrast adjustment is done by Contrast Limited Adaptive Histogram Equalization (CLAHE). After normalizing the input image contrast is done, then the MSR process can be done with the steps:

a. First step: Look for saliency maps. This process is used to get the most striking areas in the image.
   The average intensity of the HSV image is used to form the salience map. Equation 1 is used to convert the HSV image into a grayscale image.

$$Sals(I_k) = \frac{(I_{h\,mean} - I_h)^2 + (I_{s\,mean} - I_s)^2 + (I_{v\,mean} - I_v)^2}{MaxIntensity} \quad (1)$$

   MSR that uses a saliency map assumes that the object will always be in the brightest / most striking area. This process results if the human object, has a low light intensity (wearing black clothes), or is in a dark area, then there is a possibility that the object to be detected becomes smaller.

b. Second step: Change the grayscale image to a black-and-white/binary image in equation 2. The binary image is obtained through the process of separating pixels based on their gray degree. Pixels with a gray degree lower than the average grey level are given a value of 0, else a value of 1, or known as the Thresholding.

$$g(x,y) = \begin{cases} 1, & f(x,y) > T \\ 0, & f(x,y) < T \end{cases} \quad (2)$$

c. The third step: is to determine the selected area based on pixel value 0 and pixel value 1. This process involves searching for contours. Contour can be explained as a condition caused by changes in the intensity of neighboring pixels. Contour representation can be an edge list which is an ordered set of edge pixels. This representation is simplified by calculating the chain code. In the edge array, the chain code shows the position of each edge pixel. The directions used are 8 cardinal directions as shown in Figure 2.

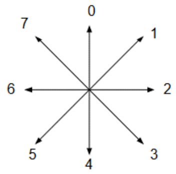


Figure 2. Eight cardinal directions

Starting with an edge pixel and clockwise, the direction of each edge pixel that forms the boundary of the object is encoded with one of the eight chain codes. The chain code represents the boundary of the object with the first edge pixel coordinates then followed by the chain code list. Edge Detection used because the MSR process only requires outer regions of the white pixel image. Edge detection used to speed up the contour search. A contour search will found all areas containing white pixels in black pixels. The results is the Minimum Bounding Rectangle (MBR), which is to determine the smallest square with white pixels. The square can be used to calculate the middle value between the top, bottom, left, and right of the square.

d. Step four:
Add the MSR square results with a Window Selection Rate (WSR). WSR indicates the windows to be processed. The wider the search area, the longer the search time will be. Too small search areas can also cause objects not to be in the search area. Figure 3 shows the difference in WSR value of 15% and 30%.

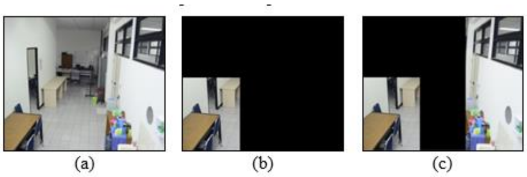


Figure 3. (a) Original Image; (b) MSR Process Results for WSR15%; (c) MSR Process Results for WSR 30%.

e. Step five: Filtrate the image. Image filtration is done to get rid of MSR selection areas that do not need to be processed in the next processing. The area of MSR results that will be removed is if the selected area are smaller than 64x128. Besides, the filter process is used to eliminate the same area (redundancy) so that the selected area is sufficiently processed once.

The results of the MSR process are rectangular areas in the image that represent the most striking areas of the image. In this study, a room with two cameras is used for capturing video. The camera cannot catch all areas. The information from two cameras could enhance the effectiveness and precise location of human beings. There are two stage to filter the information.

First stage, group the two nearest points with a threshold into one group. Second stage, the point nearest to the control point will be used as the outcome of object detection.

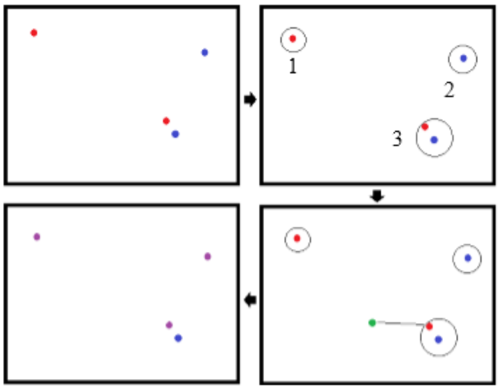


Figure 4. The result of a detection is seen in two cameras with a purple point and a green point as a control

This paper used object detection to determine human. MSR pre-processing can enhance and decrease errors the human detection using SVM classifier and HOG features. Human detection with MSR is highly affected by the level of brightness. By integrating human detection from 2 different camera, object mapping performance improves by an average of 87.07 percent with an accuracy of between 0.02-2.2 meters. The disparity in accuracy depends on the outcome of the SVM classification, which cannot accurately decide the position of the foot. If the distance between the two objects is near, then it can cause an error in the identification objects.

B. Exposure Fusion Framework

Exposure Fusion Framework is a technique of improving the contrast of low-quality images by increasing lighting in dark areas and maintaining the lighting in bright areas [5]. It calculates lighting weight values to produce synthesis images. The synthesis image will be obtained based on the lighting ratio and combined with the original image to get an image with an excessive contrast increase. General description of how this algorithm works can be seen in Figure 5.

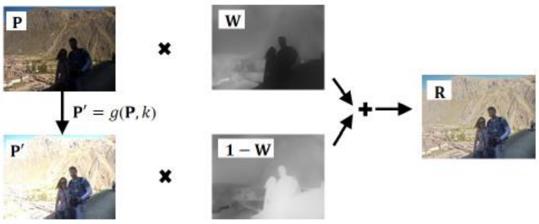


Figure 5. Exposure Fusion Framework works

If the contrast increase is done thoroughly, then an image with sufficient lighting will get excessive light so that it can ruin the good picture. This was corrected by the Exposure Fusion Framework by fusing images using equation 3.

$$R^c = \sum_{i=1}^{N} W_i \circ P_i^c \quad (3)$$

where N is the number of images, Pi is the i picture on the exposure set, Wi is the weight map for the i picture, c is the color channel index and R is the result of image enhancement. Parts of the image with good lighting will have a high weight value while parts of the picture with less lighting will have a low weight value. The weight value will be normalized so that it meets equation 4.

$$\sum_{i=1}^{N} W_i = 1 \quad (4)$$

To get the exposure set on Pi, use the camera response model [10]. The camera response model can produce a collection of images that are mutually associated and derived from an input image. This model uses the mapping function between the light images, Brightness Transform Function (BTF).

Input images will be processed with the BTF function to obtain various images based on the lighting ratio according to equation 5.

$$P_i = g(P, k_i) \qquad (5)$$

Where ki is the lighting ratio and g is a BTF function. Exposure Fusion Framework will combine the original image with an image of the lighting ratio according to equation 6.

$$R^c = W \circ P^c + (1 - W) \circ g(P^c, k) \qquad (6)$$

C. Weight Matrix Estimation

The value of w is the key in this improvement algorithm. This value will help refine low contrast areas and maintain good contrast areas. The value of w will be large on pixels with good contrast and vice versa low on pixels with poor contrast. The weight matrix will be calculated using equation 7.

$$\mathrm{W} = \mathrm{T}\mu \qquad (7)$$

where T is the scene illumination map and μ is a parameter controlling the enhance degree. When μ = 0, there are no enhancement. When μ = 1, pixels are enhanced. The lightness factor can be used to estimate the scene illumination map. The initial illumination for each pixel *x* is calculated using equation 8.

$$L(x) = \max_{c \in \{R,G,B\}} P_c(x) \qquad (8)$$

Ideal illumination for the regions with identical structures should have local consistency. As in [4], **T** is refined by solving the optimization in equation 9.

$$\min_{\mathbf{T}} \|\mathbf{T} - \mathbf{L}\|_2^2 + \lambda \|\mathrm{M} \mathrm{o} \nabla \mathrm{T}\|_1 \qquad (9)$$

Where ∇ contains $\nabla_{\mathrm{h}}\mathbf{T}$ (horizontal) and $\nabla_{\mathrm{h}}\mathbf{T}$ (vertical). M is the weight matrix and λ is the coefficient.

The design of **M** is important for the illumination map refinement. As in [4], weight matrix design as:

$$\mathrm{M_d}(\mathrm{x}) = \frac{1}{\left|\sum_{y \in w(x)} \nabla_{\mathrm{d}} \mathrm{L}(\mathrm{y})\right| + \varepsilon}, \quad d \in \{\mathrm{h}, \mathrm{v}\} \qquad (10)$$

where *w(x)* is the local window centered at the pixel x and ε is a very small constant to avoid the zero denominators.

Equation 11 shows a reduction in the complexity of equation 9.

$$\min_{\mathbf{T}} \sum_{\mathrm{x}}((\mathbf{T}(x) - \mathbf{L}(x)^2 + \lambda \sum_{\mathrm{d} \in \{\mathrm{h,v}\}} \frac{\mathbf{M_d}(x)(\nabla_{\mathrm{d}}\mathbf{T}(x))^2}{|\nabla_{\mathrm{d}}\mathbf{L}(x)| + \varepsilon}) \qquad (11)$$

Let $\mathbf{M}_\mathrm{d}$, **L**, **T** and $\nabla_{\mathrm{d}}\mathbf{L}$ denote the vectorized version of $\mathbf{M}_\mathrm{d}$, **L**, **T** and $\nabla_{\mathrm{d}}\mathbf{L}$ respectively. Then, by solving the following linear function in equation 12, the solution can be achieved.

$$(\mathbf{I} + \lambda \sum_{\mathrm{d} \in \{\mathrm{h,v}\}} \mathrm{D}_{\mathrm{d}}^{\mathrm{T}} Diag(M_d \emptyset (|\nabla_{\mathrm{d}}\mathbf{L}| + \in))\mathbf{D_d})\mathbf{t} = \mathbf{I} \qquad (12)$$

where ∅ is the element-wise division, **I** is the unit matrix, the operator Diag(**v**) is to construct a diagonal matrix using vector **v**, and $\mathbf{D}_\mathrm{d}$ is the Toeplitz matrices from the discrete gradient operators with a forward difference.

D. Camera Response Model

Camera Response model [10] or also referred to as Beta-Gamma Correction uses the brightness transform function (BTF). BTF is a mapping function between two pictures taken in the same scene with different lighting. The BTF model on the Exposure Fusion Framework can be shown in equation 13.

$$g\ (P, k) = \beta P^{\gamma} = e^{b(1-k^a)} P^{(k^a)} \qquad (13)$$

where β and γ are the two parameters in the model that can be calculated from the camera parameters a, b, and the lighting ratio k. This model assumes that there is no information about the camera provided and it uses the parameters of a fixed camera (a = −0.3293, b = 1.1258) that can fit most cameras [5].

E. Exposure Ratio Determination

Exposure Ratio Determination is used to determine the best ratio of the synthetic image in the regions where the original image is under-exposed. First, the well-exposed pixels are discarded and an under-exposed pixels are obtained. Second, extract the low illuminated pixels as:

$$\boldsymbol{Q} = \{\mathbf{P}(x) | \mathrm{T}(x) < 0.5 \qquad (14)$$

where **Q** contains only the under-exposed pixels. Although the colour is the same, the brightness of the images under different exposures obviously varies. Third, consider the brightness factor. The brightness factor B is defined as the three-channel geometric mean:

$$\boldsymbol{B} = \sqrt[3]{\boldsymbol{Q}_r o \boldsymbol{Q}_g o \boldsymbol{Q}_b} \qquad (15)$$

Qr, Qg, and Qb are the red, green, and blue channels of the input image Q. The geometric mean use since it has the same BTF model parameters (β and γ) with all three color channels, as shown in Equation 16.

$$\begin{aligned}\boldsymbol{B} &= \sqrt[3]{\boldsymbol{Q'}_r o \boldsymbol{Q'}_g o \boldsymbol{Q'}_b} \\ &= \sqrt[3]{(\beta\boldsymbol{Q}_r^\gamma) o (\beta\boldsymbol{Q}_g^\gamma) o (\beta\boldsymbol{Q}_b^\gamma)} \\ &= \sqrt[3]{\beta(\boldsymbol{Q}_r o \boldsymbol{Q}_g o \boldsymbol{Q}_b)^\gamma} \\ &= \beta\mathbf{B}^\gamma \end{aligned} \quad (16)$$

Images with higher exposure can provide more detailed information for humans. Thus, the optimal k should give the largest information. The entropy of the image was defined as Equation 17 to measure the amount of information:

$$H(\mathbf{B}) = -\sum_{i=1}^{N} p_i . log_2 p_i \quad (17)$$

p_i is the i-th bin of the histogram of $\mathbf{B}$ counts the number of data valued in $(\frac{i}{N}, \frac{i+1}{N})$ and N is the number of bins (N is often set to be 256). Finally, the optimal k calculated by maximizing the image entropy of the enhancement brightness as Equation 18.

$$k = \underset{k}{argmax}\, H(g(\boldsymbol{B}, k)) \quad (18)$$

The optimized k can be solved by a one-dimensional minimizer. Resize the input image to 50×50 when optimizing k can improve the calculation effciency.

EFF provide a good contrast enhancement. The framework use illumination to get weight matrix, using camera response for multi-exposure and determine effective exposure ratio. Finally, it combine the input and synthetic image according to the weight matrix. In hundreds of low-light images from five public datasets, the experimental showed the advanced solution compared with several state-of-the-art alternatives [5].

3. Methodology

A. Data Collection

In this section, we use a video dataset made by Winardi et al [2]. The video comes from CCTV that is mounted on a room placed in a corner of the room at a height of 2.47 meters as shown in Figure 6. Each camera uses a resolution of 640 X 360, 30 fps and takes about one minute. The video is taken during the day with sunlight in the room where one side of the room is a wide window.

The video consists of 6 (six) pieces. Video details consist of:

1. Video 1: video with 2 (two) persons detected and the second person will disappear and reappear in the middle of the video.
2. Video 2: video with 1 (one) person detected.
3. Video 3-6: video with 2 (two) persons detected and the second person will only enter and exit once.

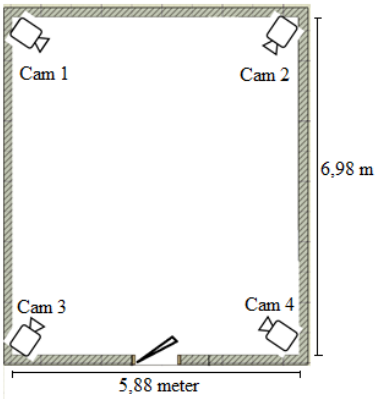
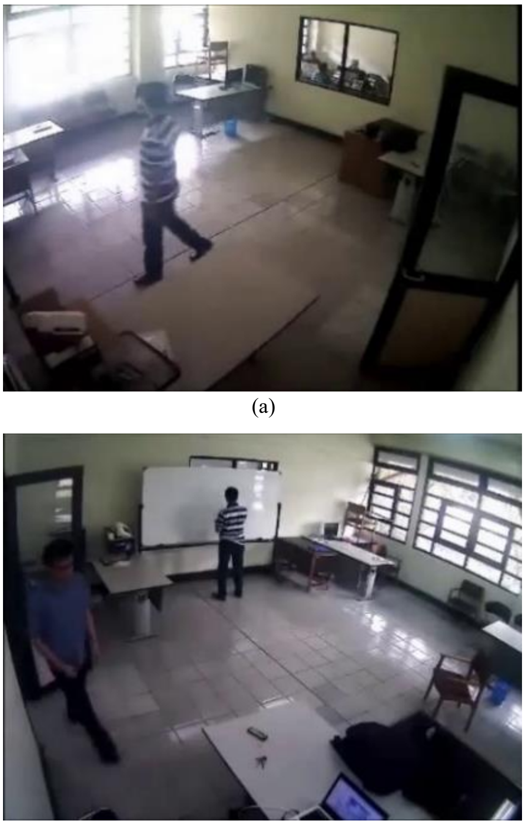


Figure 6. Camera shooting location

Sample frame on a dataset can be seen in Figure 7.


(a)

(b)

Figure 7. Sample frame: (a) One person detected; (b) Two person detected.

The criteria of object validity in the frame are as follows:

1. An object is considered valid if the object is a person that has been previously recognized.
2. The object is considered valid if the object has entered the video capture area.
3. The object is considered valid if the object is fully appeared on the video (not a part)
4. The object is considered valid even though the object is small in size (far from CCTV video observations) but still in the video capture area and fully.
5. If a frame consists of several objects at once, then the number of objects will be calculated based on the total object that appears in the frame on the video.

In this study, the number of human appearances on video shows the number of human objects that were detected manually. It will be calculated according to direct observation based on the frame for each video. The criteria of human appearances are as follows:

1. The detected person will be counted as one occurrence in one frame
2. The detected person will be counted for every frame in the video without skip
3. The detected person that intersect with other objects in the frame is still be counted as one occurrence as long as it is still visually visible based on observation.
4. The object of the person is still be counted until the person is out of CCTV surveillance.

Table 1 shows the description of the artificial video dataset.

Table 1. Description of the artificial video dataset

Data-set	FPS	Dura-tion	Number of Frame	Number of Person	The number of human appearances on video
V-1	30	1:05	1946	2	1541
V-2	30	0:57	1575	1	1244
V-3	30	1:03	1880	2	1900
V-4	30	0:57	1699	2	1877
V-5	30	0:55	1659	2	1767
V-6	30	0:54	1605	2	1486

B. Our Approach

In this section, we elaborate our approach in Figure 8.

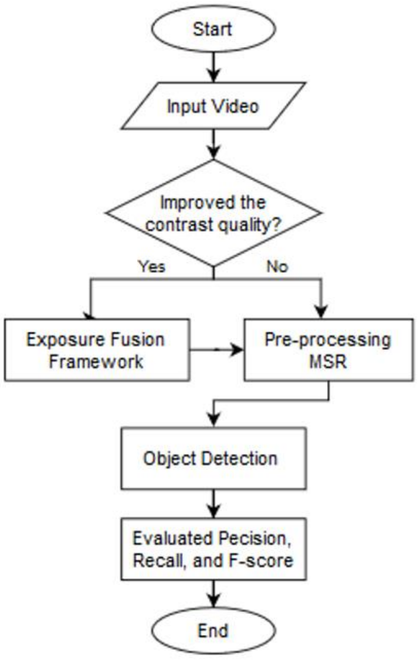

Start
Input Video
Improved the contrast quality?
Yes
No
Exposure Fusion Framework
Pre-processing MSR
Object Detection
Evaluated Pecision, Recall, and F-score
End

Figure 8. The flowchart for human detection

Our approach consists of several steps.

1. Our dataset has been described previously as input. The videos are converted into frame sequences. We extract the people manually from each frame as the number of human appearances on video.
2. Enhanced image contast for each frame using Exposure Fusion Framework (EFF). This technique can enhance the image without any reference. EFF parameters remain in all experiments, $\lambda = 1$, $\varepsilon = 0.001$, $\mu = 0.5$ [5]. λ is the coefficient of weight matrix. When $\lambda = 1$, it means the weight of the matrix does not change. ε is a very small constant to avoid the zero denominator in weight matrix. Small values are suitable for this parameter. When $\mu = 0.5$, it means that the enhanced degree for both the under-exposed pixels and well-exposed pixels is equal.
3. Preprocessing using MSR. Normalizing images using CLAHE and forming salience maps. The block size of MSR is 64x128. The WSR value is 30% [1].
4. Human detection using the HOG features with the Linear-SVM classifier. This stage is processed with the EmguCV library.
5. To evaluate the performance, we can calculate the precision, recall and f-score as in equations 19, 20 and 21. For validation the results, leave-one-out cross validation is used. Human detection is done for the entire frame of the video and for all frames without skip. The process will be repeated and get the same results.

$$precision = \frac{True\ Positive}{True\ Positive+False\ Positive} \quad (19)$$

$$recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \quad (20)$$

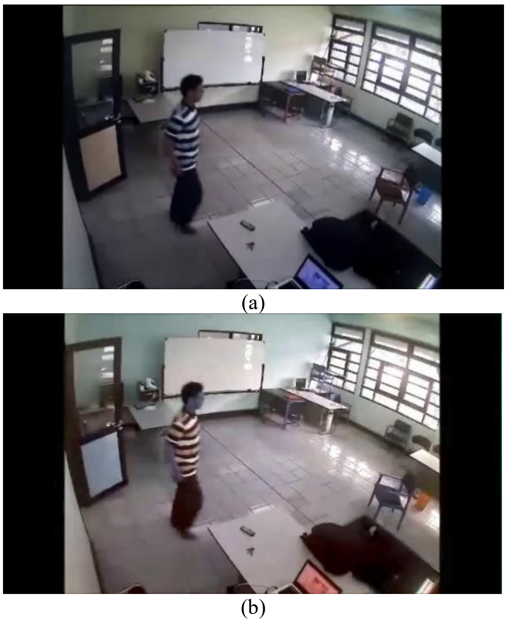
$$fscore = 2\ x\ \frac{precision\ x\ recall}{percision+recall} \quad (21)$$

which

- True Positive (TP): system detected human(s) correctly.
- False Negative (FN): system has not detected humans
- False Positive (FP): system detected humans, but there are no humans in the image.

4. Experimental Result

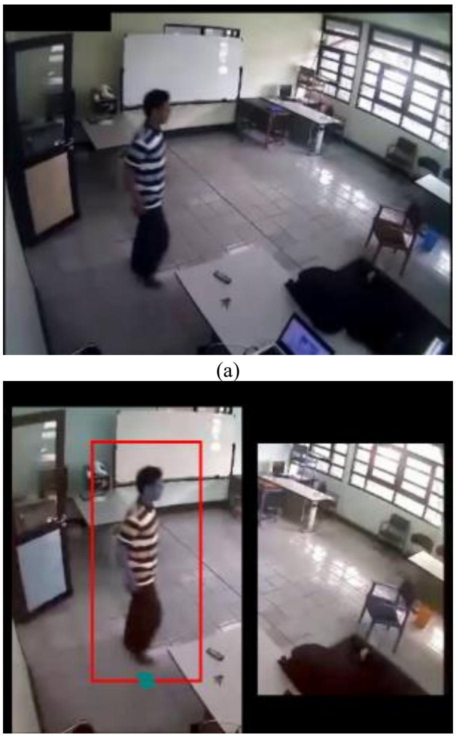
In experiments, all frame sequences are extracted from each video. Figure 9 shows the sample frame from artificial video. Figure 9(a) is original frame. This image is from CCTV without any enhancement. Figure 9(b) is an enhanced image from the original frame. We perform contrast enhancement for each frame using EFF and combining frames to make a better video quality.


(a)

(b)

Figure 9. Sample frame: (a) original; (b) enhaced using EFF.

The next step is preprocessing the video using MSR before detect the human object. The preprocessing stage using MSR will reduce the video search area. In this experiments, we use HOG features with Linear-SVM classifier for human detection. Figure 10(a) shows person in a sample frame that was not detected successfully. The image on this frame comes from original video without enhancements. The MSR and Linear-SVM have difficulty detect human in this video. Figure 10(b) shows a human being successfully detected in an enhanced sample frame. MSR is able to reduce the search area on a video compared to the original video.


(a)

(b)

Figure 10. Human Detection in the sample frame using MSR and Linear-SVM: (a) original (b) enhanced using EFF.

For evaluation, we will compare our approach with an approach without contrast enhancement. An approach without contrast enhancement uses original (ori) videos. Our approach uses enhanced (our) videos. Both approaches run the MSR preprocessing and human detection stages using the same configuration parameters. The evaluation of human detection on the original (ori) and enhanced (our) videos are shown in Table 2.

Table 2. Evaluation of human detection

Dataset	TP		FP		FN	
	ori	our	ori	our	ori	our
V-1	45	**86**	0	0	1496	1455
V-2	64	**114**	0	0	1180	1130
V-3	112	**133**	**1**	6	1787	1761
V-4	46	**121**	0	0	1831	1756
V-5	105	**186**	0	0	1662	1581
V-6	80	**123**	**0**	1	1406	1362

In Table 2, humans who were detected correctly did improve with our approach. This is indicated by increasing True Positive (TP) value in our approach. For all videos, there are 311 additional human detections correctly from the original videos, which means an increase of 68.81%. For each video, the results showed that on average 51 additional objects were detected as humans correctly and an average increase of 80.32% compared to the original video. The details are shown in Table 3.

Table 3. The addition of human detected for each video in our approach

Dataset	TP-ori	Additional detected-our	Additional-our (percentage)
V-1	45	41	91.11%
V-2	64	50	78.13%
V-3	112	21	18.75%
V-4	46	75	163.04%
V-5	105	81	77.14%
V-6	80	43	53.75%
Average		51.8	80.32%

However, the number of objects that were detected correctly as a human is relatively low compared to the number of human appearances on video.

In Table 2, humans who were detected wrong are also increased with our approach. This is indicated by increasing the value of False Positive (FP) in our approach to V-3 and V-6. The results show that on average 1 additional object was detected as human incorrectly. Figure 11 shows the sample of incorrect human detection. In video 3 and video 6, incorrect detection is caused by thick and dark glass door border or the human part (feet) passing through the door.

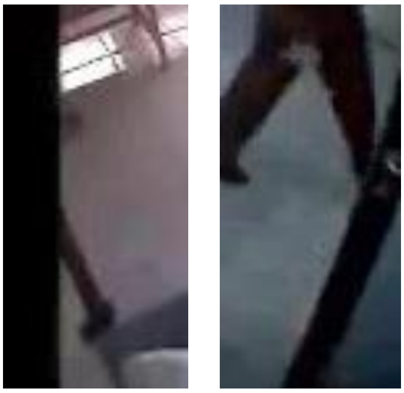


Figure 11. Incorrect human detection in our approach

In addition, a high False Negative value indicates there are still many humans on the video that have not yet been detected. Many aspects affect this result such as video quality, MSR capabilities, feature selection, classifier capabilities, and so on. In this experiment, no object has been detected in the bright area around the window with sunlight. This causes the overall performance to be low. Enhanced videos have improve MSR to determine the search area, but there is still space for improvement.

For performance analysis, we will compute the precision, recall and f-score rate. The precision and recall of human detection on the original (ori) and enhanced (our) videos are shown in Table 4.

Table 4. Performance analysis of precision and recall

dataset	precision		recall	
	ori	our	ori	our
V-1	100.00%	100.00%	2.92%	5.58%
V-2	100.00%	100.00%	5.14%	9.16%
V-3	99.12%	95.68%	5.89%	7.00%
V-4	100.00%	100.00%	2.45%	6.45%
V-5	100.00%	100.00%	5.94%	10.53%
V-6	100.00%	99.19%	5.38%	8.28%
Average	99.85%	99.15%	4.62%	7.83%

In Table 4, the precision rate of original and our approach are almost perfect. A high average precision rate means the object detected really is human. Otherwise, the recall rate of original and our approach are still low. It means that this approach has not been able to detect all humans on video.

The precision rate of our approach is 0.7% lower compared to the original video. However, this decrease was slightly improved by an increase in recall of 3.21%. To see the overall performance, we will compute the f-score rate.

Table 5 shows the overall performance obtained from the f-score of human detection on the original (ori) and enhanced (our) videos.

Table 5. Overall performance analysis

Dataset	F-Score		
	ori	our	Δ
V-1	5.67%	10.57%	4.90%
V-2	9.79%	16.79%	7.00%
V-3	11.13%	13.05%	1.92%
V-4	4.78%	12.11%	7.33%
V-5	11.22%	19.05%	7.83%
V-6	10.22%	15.28%	5.06%
Average	8.80%	14.48%	5.67%

The highest performance of our approach was on V-5 with 19.05%, followed by V-2 and V-6 with 16.79% and 15.28%. These results are better than the best performance of the original approach.

The average f-score on the original video is 8.8%. It shows the human detection performance in overall is relatively low. In this result, we also see that the average f-score of our approach is still quite low. This is due to the high false negative rate. In our datasets, room conditions consist of a fairly dark area and very bright area (sunlight). The same objects will walk around in both regions. The objects was distorted by surrounding conditions and difficult to detect.

However, the overall performance of our approach is better than the original video in Table 5. There is an increase in performance for each video. Overall performance rose 5.67% which means an increase of 64.46% from the original video.

5. Conclusion

In this paper, the image self-enhancement to obtain a better search area for low resolution video detection has been carried out. The Exposure Fusion Framework is applied. The proposed approach is evaluated using human detection experiments on six artificial CCTV videos, HOG features and Linear-SVM as test case. The results are computed in terms of the average precision, recall and f-score rate and improved performance is compared with the results of the original video. From the experiment results, proposed approach shows an increase in human detection when image enhancement applied. It can improve performance in recall of 3.21% and f-score of 5.67% in human detection on low resolution video. The precision rate has decreased slightly by 0.7%, but it is still acceptable compared to other performance improvements. For all videos, there is a 68.81% increase in the number of humans detected correctly and an average increase of 80.32% in the number of humans detected for each video. For overall performance, there is an increase of 64.46% compared to the original video. However, it is not showing the best overall performance. Experiments also suggest that the proposed approach is still difficult to detect humans in bright areas, objects become distorted by surrounding conditions, and there is still space for improvement to determine the search area and other object detection algorithm such as deep learning techniques.

Acknowledgement

This research is supported by STMIK Mikroskil under grant Nr. 1918/ST.45/01/PN/2019

References

[1] G. Silva, L. Schnitman, & L. Oliveira, "Constraining Image Object Search by Multi-Scale Spectral Residue Analysis," *Pattern Recognition Letters*, vol. 39, pp. 31-38. 2014.

[2] S. Winardi, S. Akbar, & Y.D. Wardhana. "Human Localization with Multi-Camera Using Detection and Tracking Object," *In Proceeding of the IEEE ICIC 2019*, pp. 1-6, 2019.

[3] R.K. Hanspal, & K. Sahoo, "A Survey of Image Enhancement Techniques," *International Journal of Science and Research*, vol. 6, no. 5, pp. 2467-2471. 2017

[4] X. Guo, "LIME: A Method for Low-Light Image Enhancement," *In Proceedings of the 24th ACM International Conference on Multimedia*, pp. 87-91. 2016.

[5] Z Ying, G Li, Y Ren, R Wang, W Wang, "A New Image Contrast Enhancement Algorithm Using Exposure Fusion Framework," *In Proceeding of the Springer International Conference on Computer Analysis of Images and Patterns*, pp. 36-46, 2017.

[6] J. Kwon & N. Kwak, "Human detection by Neural Networks using a low-cost short-range Doppler radar sensor," *In Proceedings of the IEEE Radar Conference (RadarConf)*, pp. 0755-0760. 2017.

[7] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection, “ *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 886–893. 2005.

[8] M. Y. Liu, O. Tuzel, A. Veeraraghavan, R. Chellappa, “Fast directional chamfer matching, “ *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1696–1703. 2010.

[9] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao and X. Li, “Detection of Sudden Pedestrian Crossings for Driving Assistance Systems, “ *In Proceedings of the IEEE Transactions on Systems, Man, and Cybernetics*, pp. 729-739. 2012.

[10] Z. Ying, G. Li, Y. Ren, R. Wang & W. Wang, "A New Low-Light Image Enhancement Algorithm Using Camera Response Model," *In Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3015-3022. 2017.