**TELKOMNIKA**, Vol.16, No.3, June 2018, pp. 955~964
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
**DOI:** 10.12928/TELKOMNIKA.v16i3.9033

# Stereo Vision Human Motion Detection and Tracking in Uncontrolled Environment

**Bunseng Chan*[1], King Hann Lim[2], Lenin Gopal[3], Alpha Agape Gopalai[4], Darwin Gouwanda[5]**
[1,2,3] Department of Electrical and Computer Engineering, Faculty of Engineering and Science, Curtin University Malaysia
[4,5] Mechatronic Engineering Department, School of Engineering, Monash University Malaysia
*Corresponding author, e-mail: bunsengchan@ieee.org

***Abstract***

*Stereo vision in detecting human motion is an emerging research for automation, robotics, and sports science field due to the advancement of imaging sensors and information technology. The difficulty of human motion detection and tracking is relatively complex when it is applied to uncontrolled environment. In this paper, a hybrid filter approach is proposed to detect human motion in the stereo vision. The hybrid filter approach integrates Gaussian filter and median filter to reduce the coverage of shadow and sudden change of illumination. In addition, sequential thinning and thickening morphological method is used to construct the skeleton model. The proposed hybrid approach is compared with the normalized filter. As a result, the proposed approach produces better skeleton model with less influential effect on shadow and illumination. The output results of the proposed approach can show up to 86% of average accuracy matched with skeleton model. In addition, obtains approximately 94% of sensitivity measurement in the stereo vision. The proposed approach using hybrid filter and sequential morphology could improve the performance of the detection in the uncontrolled environment.*

***Keywords**: Stereo vision, human motion detection and tracking, sequential morphology, skeleton model*

## 1. Introduction

The technology of stereo vision has grown exponentially nowadays as a sensory tool to capture visual information. Inspired by human eyes, stereo vision uses two cameras displaced horizontally on a calibrated platform to reconstruct three-dimensional (3D) perception of an object based on the visual properties such as shape, color distribution, and illumination [1]-[2]. With the emerging of the imaging sensors and computer vision technology, the application of the stereo vision is extended to various industrial fields including virtual reality, gaming, military training filmmaking, and sports activities [3]-[4].

Stereo vision can be divided into two categories, i.e., active stereo-vision and passive stereo-vision. Active stereo vision uses direct light source radiance on a subject, and the reflected light is captured via a camera [5]. However, the cost of the active stereo vision is relatively high, and its visual output is affected by surrounding illumination [6]. On the other hand, active stereo vision incorporated with the depth sensors including ranging camera, flash LIDAR, time-of-flight, and RGB-D camera to improve its performance [7]. Nevertheless, the field of view and resolution for this camera system is restricted by the specification of depth sensors [8].

Conversely, passive stereo vision constructs the images based on the interreflection of the light source to the subject [5]. Passive stereo vision is commonly used in 3D image reconstruction due to its simplicity. It is a user-friendly system, and the system cost is low [9]. Ganapathi *et al*. [10] *used* the graphics processing unit (GPU) with single time-of-flight (ToF) camera to capture the object. The ToF camera measures the depth of the object with the assist of the depth sensors unit to produce the 3D information. This system increases the robustness of the system, but it requires high dimensionality of the memory space for the processing.

Human detection and segmentation methods are the initial step of the human features extraction before proceed to the gait analysis. Background subtraction is one of the commonly used background removals to detect the object in image sequences. However, background

*Received December 23, 2017; Revised April 9, 2018; Accepted April 27, 2018*

subtraction method has limitation in extracting fine human object from an uncontrolled environment due to illumination [11]. This limitation was led to the failure of full human feature extraction for further analysis. The human feature is also known as human silhouette, which is important for the human skeleton model construction in motion analysis [12]. Hence, image enhancement is important to improve the human feature extraction.

Image filtering is one of the image enhancing techniques integrating into background subtraction to enhance the image quality. Commonly, image filter is used to emphasize the image features or smoothen the noises of the image. However, the selection of the filter must be appropriated to achieve an optimum output for analysis purpose. The sudden illumination change is still one of the challenges needed to be overcome in image processing. Various type of filters including mean filter, Gaussian filter, and median filter has been applied by the researchers to improve the sudden illumination change issue [13]. Nevertheless, it does not achieve an ideal output required in the uncontrolled environment.

In this paper, the use of a hybrid filter is proposed for an accurate human detection in stereo vision under an uncontrolled environment. The paper is outlined into three sections. Sections 2 is the overview of human detection and tracking in stereo vision; and Section 3 shows the experimental setup, experimental results and discussion.

## 2. Human Detection and Tracking in Stereo Vision

This work aims to establish human detection and tracking in an uncontrolled environment using stereo cameras. Human motion being detected from stereo vision can be used as gait recognition and analysis. Hence, human detection is the important step to recognize the dynamic object in a video. Once human is detected, human tracking is applied to localize an human object continuously in a video to ensure the computational reduction for real-time application [14].

Human skeleton extraction is one of the tracking methods employed in human object. It basically uses the mathematical morphology algorithms to enhance the human model and display the relevant connection of the human joints [15]. Human detection and tracking is relatively complex when it is performed in the uncontrolled due to the unpredicted illumination changes and complex background [16].

An overview of the proposed algorithm is shown in Figure 1. It consists of two major sections, i.e., human detection and segmentation, and human tracking with skeleton model. The first part assists to capture, detect the human, and enhance the motion segmentation. Meanwhile, the second part serves as continuous tracking of the detected object from the previous segment and transforms into skeleton model.

Stereo camera emulates human vision using two passive cameras to initiate 3D capability of human eyes in a digital world [17]. The output of the stereo vision provided depth information which provided more detail information about the object. In addition, it improves the robustness of the object information. There are many applications in entertainment and training system due to its advantages.

In a color image, it consists of three color channels, i.e., Red, Green, and Blue (RGB). However, direct image processing using three channels largely increases the computational cost of system. Therefore, a pre-processing module was included to transform the color image obtained from the stereo camera into a grayscale image. By combining of the normalized RGB value, it produces the neutral white color (R+G+B=1). The grayscale image can be obtained by combining the 29.5% of Red value, 55.4% of the Green value, and 14.8% of the Blue value [18]. The combination of this value produces the grayscale image.The conversion of the color scale to grayscale image is based on the equation shown as follows:

$$f_{gray} = (0.295 \times R) + (0.554 \times G) + (0.148 \times B) \quad (1)$$

A two frame differentiation method is employed to remove the background and obtain the human motion from the frame. The frame differentiation formula defines as:

$$f_{d(k,k+1)} = \begin{cases} f_{k+1} - f_k, if\ (f_{k+1} - f_k) > 0 \\ 0, else \end{cases} \quad (2)$$

$$f_{d(k,k+1)} = \left| f_{k+1} - f_k \right| \tag{3}$$

where $f_{d(k,k+1)}$ is the two-frame differencing, $f_{k+1}$ is the frame at k+1 in the image sequences, and $f_k$ is the current frame or frame k in the image sequences.

Human motion detected using the two-frame differencing faces illumination issue when the system is implemented at the outdoor environment. Therefore, post-processing is required to remove the illumination problem. Generally, the median filter is used to reduce the noise of an image and remain the edges [19]. Meanwhile, the Gaussian filter uses Gaussian function to smooth the little-unwanted noises. In the post-processing module, a hybrid filter integrating Gaussian filter and median filter is proposed to remove the effect of shadow. The hybrid filter equation is shown as follows:

$$f_o(x,y) = G(x,y) + M(x,y) \tag{4}$$

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5}$$

$$M(x,y) = med\{f(x-i, y-j), i,j \in W\} \tag{6}$$

where $f_o(x,y)$ is the output image, G(x,y) and M(x,y) is the Gaussian filter and median filter respectively, $\sigma$ is the standard deviation of the distribution, and $W$ is the $n \times n$ mask.

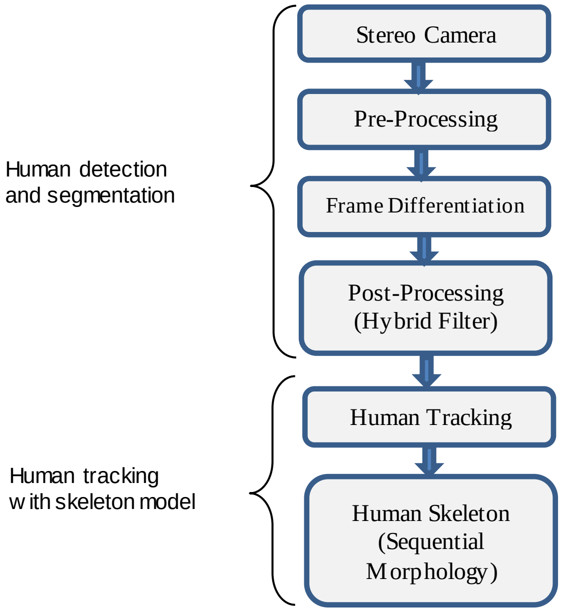

Stereo Camera
Pre-Processing
Frame Differentiation
Post-Processing (Hybrid Filter)
Human Tracking
Human Skeleton (Sequential Morphology)
Human detection and segmentation
Human tracking with skeleton model

Figure 1. Overview of the proposed approach

The human detected is then enhanced its features after the noise filtering. The enhancement involving mathematical morphology to removes the cavities, holes, or opens up holes of the unwanted region. The mathematical morphology involves in this process including the opening and closing morphology. The mathematical morphology for opening and closing of an image $f(x,y)$ by a structuring element $e(x,y)$ and its defined accordingly as:

$$f \circ e = (f \ominus e) \oplus e \tag{7}$$

$$f \bullet e = (f \oplus e) \ominus e \tag{8}$$

The detected human shape recovered its features then is used for continuous tracking of its movement and transform into skeleton model. The skeleton model is a stick diagram of an object which is used for to representing its basic shape. Basically, the skeleton construction can be done in two different techniques, i.e., maximal balls technique, and thinning and thickening technique [15]. Maximal balls technique used multiple constant radius balls concept in a Euclidean space to identify the center of the structure to construct the skeleton [20]. However, maximal balls technique produced disconnected skeleton model during its construction and did not provide an efficient extraction [20]. Hence, it is not suitable for the human skeleton construction.

Therefore, the human skeleton construction in this module will focus only on thinning and thickening technique. Basically, the thinning and thickening technique is using the sequential method as structuring elements [21]. The basic mathematical morphology for thinning and thickening of an image $f(x,y)$ is composite with a structuring element $e(x,y)$ as defined accordingly as:

$$f \oslash e = f \setminus (f \otimes e) \tag{9}$$

$$f \odot e = f \cup (f \otimes e) \tag{10}$$

Sequential structuring element will be employed to construct the skeleton model using the thinning and thickening equation respectively as shown:

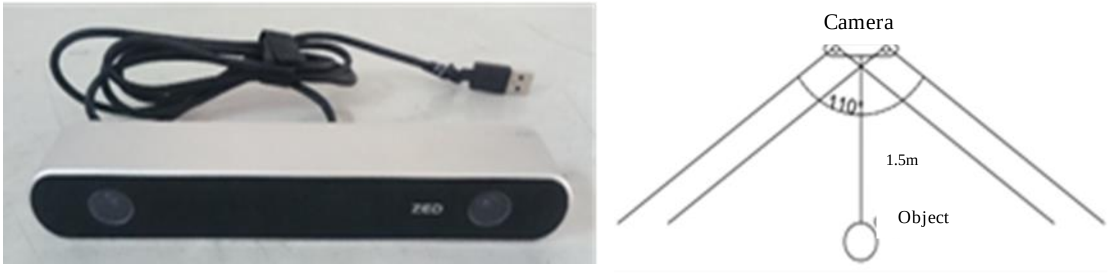
Let $e = \{e_1, e_2, e_3, \ldots, e_n\}$

$$f \oslash e = \left( \left( (f \oslash e_1) \oslash e_2 \right) \ldots \oslash e_n \right) \tag{11}$$

$$f \odot e = \left( \left( (f \odot e_1) \odot e_2 \right) \ldots \odot e_n \right) \tag{12}$$

## 3. Experimental Results and Discussion

An experimental setup was implemented using a set of high resolution and high frame rate stereo camera as shown in Figure 2(a). The camera module with depth range between 700mm to 20000mm and the maximum viewing angle provided is 110$^o$. The frame size in this experiment is 1344 × 376 pixels, and the frame rate is 60 FPS. The proposed distance between camera and subject in this study is 1.5 meter, and details of setup are depicted in Figure 2(b). The distance of the setup in Figure 2(b) is based on the human capture requirement in gold standard human motion capture system. The proposed camera height from the floor is 0.5m depicted in Figure 3.



(a) (b)

Figure 2. (a) High frame rate stereo camera module, (b) Proposed camera setup

We ran the proposed setup using a Jetson TX1. Jetson TX1 module possessed NVIDIA Maxwell$^{TM}$ GPU with 256 NVIDIA$^{®}$ CUDA$^{®}$ Cores and Quad-core ARM$^{®}$ Cortex$^{®}$-A57 MPCore Processor with 4GB LPDDR4 memory.The proposed stereo vision human tracking was tested with controlled and uncontrolled environments. Two videos were recorded in this experiment for quantitative and performance analysis. Video #1 contains human walking under the natural light at the outdoor environment as shown in Figure 4(a). There are some illumination changes throughout the video recording for Video #1. On the other hand, video #2 contains human walking in the controlled environment with fluorescent light as shown in Figure 4(b). The video recorded is processed according to the proposed approach as written in the pseudocode as shown in Figure 5.

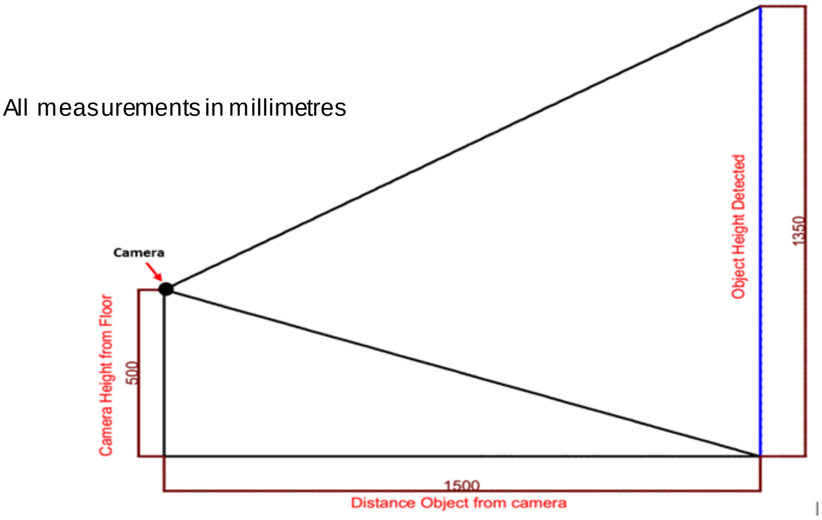


Figure 3. The details camera setup of this study

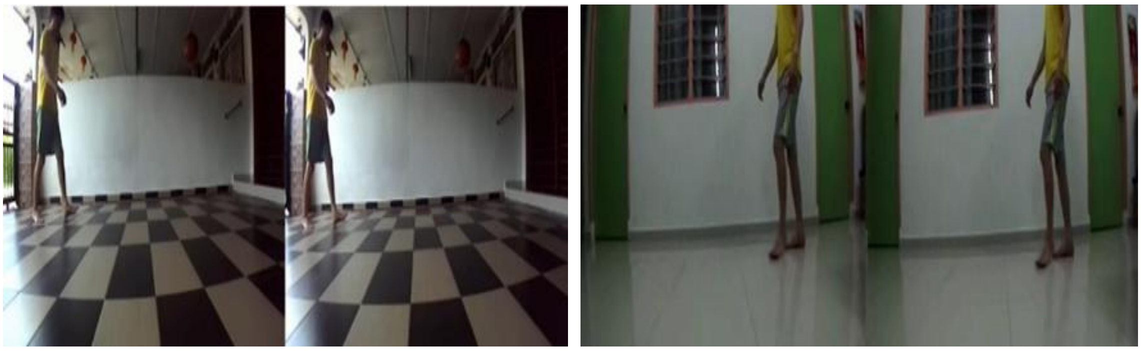

(a) (b)

Figure 4. (a) Stereo image of Video #1, (b) Stereo image of Video #2

Implementation of background subtraction without the filter has been included as shown in Figure 6. It can be observed that the illumination problem is given an indistinct output of the human silhouette as well as the shadow. Whereas, this vague image is unable to reconstruct the human skeleton model for analysis. Normalized filter and the proposed hybrid filter was included to investigate the feasibility of human skeleton model construction. The sequences images obtained from the Video #1 and Video #2 are processed using the linear normalized filtering method by using the normalize equation as shown:

$$I_N = (I_O - I_{Max})\frac{I_{NewMax} - I_{NewMin}}{I_{Max} - I_{Min}} + I_{NewMin} \quad (13)$$

where $I_O$ is the intensity of the original image, $I_{Min}$ and $I_{Max}$ is the minimum and maximum intensity of the original image respectively, and $I_{New_{Min}}$ and $I_{New_{Max}}$ is the new minimum and maximum intensity value respectively determine by the user.

```
Start:
Define the input frame as $f_k(x,y)$,  Morphology structuring element as em(x,y),
and Filter structuring element as ef(x,y)

**while** $f_k(x,y)$!=0
{
    **if** k<300
              $f_{k<300}(x,y)$  is ignored due to initialization
    **else**
              Frame captured from stereo camera copy into $f_k(x,y)$
              Convert $f_k(x,y)$  image into grey level image $f_{kgrey}(x,y)$
              Frame differencing $f_{d(k,k+1)}$= $f_{kgrey+1}(x,y)$  - $f_{kgrey}(x,y)$
    **do** {
              **for** each (x,y) of the $f_{d(k,k+1)}$
              Filter the noise while remaining the edges using the median filter}
    **do** {
              **for** each (x,y) of the $f_{dmed(k,k+1)}$
              Smoothing the $f_{dmed(k,k+1)}$  using the Gaussian filter and ef(x,y) }
    **do** {
              **for** each (x,y) of the $f_{dmg(k,k+1)}$
              open the very small region using the $f \circ e = (f \ominus e) \oplus e$
              close the holes using the $f \bullet e = (f \oplus e) \ominus e$
        }
    **while** pixel !=1
    {
              Construct the skeleton model using the sequential thinning and thickening
              method
              $em = \{em_1, em_2, em_3, \ldots, em_n\}$
              $f \oslash em = \left(\left((f \oslash em_1) \oslash em_2\right) \ldots \oslash em_n\right)$
              $f \odot em = \left(\left((f \odot em_1) \odot em_2\right) \ldots \odot em_n\right)$
    }
}
End
```

Figure 5. Pseudocode of the program

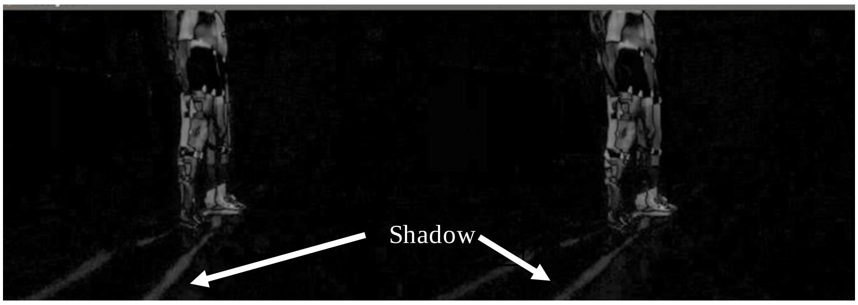


Figure 6. State-of-the-art background subtraction output image

The output of the filtered image and skeleton model as shown in Figure 7(a) and Figure 7(b) respectively. It can be observed that it is unable to detect the object fully and the shadow cannot be filtered. Therefore, the skeleton model constructed is not accurate when merging with the model. The purpose of filtering is important to remove the unwanted noise. However, normalize filtering can remove the noise but it remains the shadow. It enhances the image contrast while normalizing the pixels and directly improve the contrast of the shadow. Meanwhile, the same video has been processed using the approach proposed. The output result of the filtered image from the outdoor and indoor environment has been shown in Figure 8 and Figure 9 respectively.

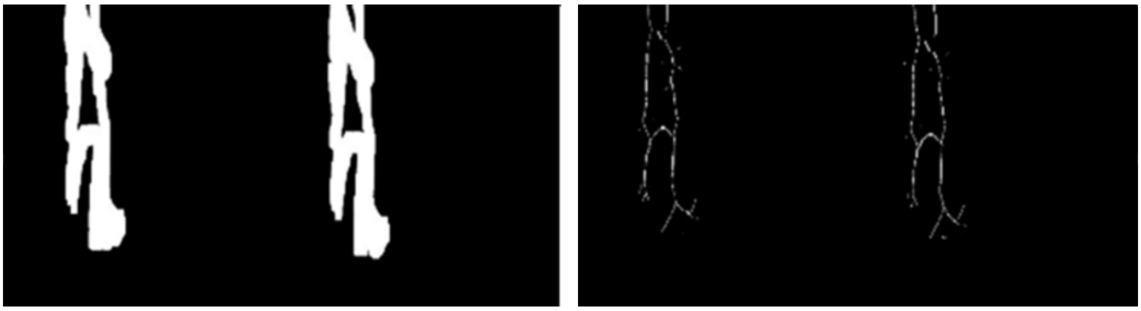

(a) (b)

Figure 7. (a) Output image filtered by normalized method, (b) Skeleton model constructed from Figure 7(a)

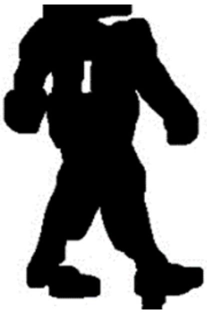


Figure 8. Filtered image using proposed approach at the outdoor environment

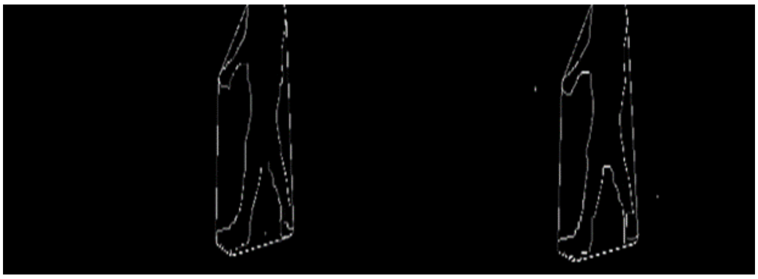


Figure 9. Human shape discovered for the indoor environment using proposed approach

From the proposed approach, it can produce complete silhouette information compares to Figure 7. From the Figure 8, the shadow and the noise of the image also successful in reducing by using the mixture of Gaussian and median filter. From the filtered image, it is processed using the mathematical morphological to obtain enhanced silhouette image as well as the skeleton model. The results of the morphological operations are shown in Figure 9. In Figure 8, the human silhouette has been enhanced to remove the cavities or holes of the interest region. A 3×3 structuring element and the closing morphology (refer to the equation 8) is used to close the unwanted holes. Output result produced in Figure 10 has complete object information as compared to Figure 7. Hence, the result from Figure 10 can construct the complete human skeleton model.

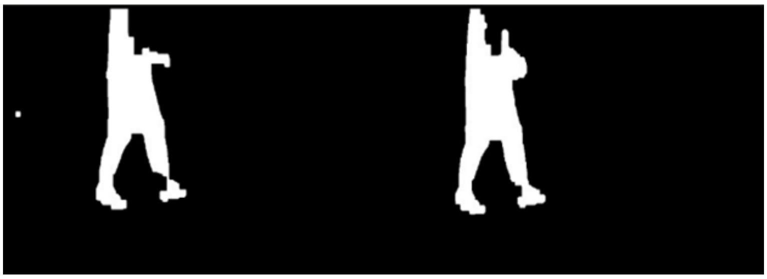


Figure 10. Mathematical morphology output

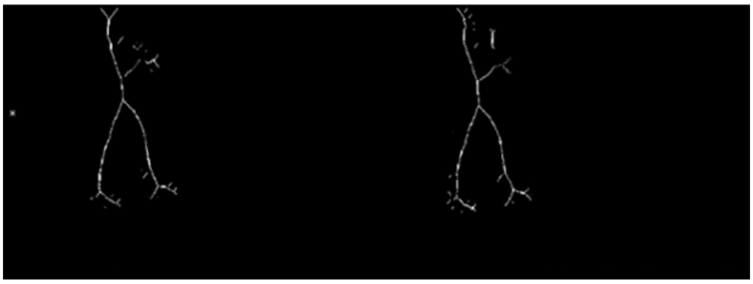


Figure 11. Skeleton model constructed

Sequential thinning and thickening morphological technique is applied to construct human skeleton model. Figure 11 shows the result of the skeleton model by using the output image from the proposed hybrid filter approach. From this approach, the image is continuing the process of thinning and thickening method until the image pixel obtained is equal to 1. It shows that the hybrid filter approach implemented to synthesize the images in Figure 8 has constructed better human skeleton model compare to the normalized algorithm as shown in Figure 7(b).

The human skeleton constructed was verified with the grayscale video to ensure the performance. The verification results are classified into two categories, i.e.: (1) Detected frame- more than 90% of the skeleton structure fall into the human silhouette region, (2) Non-detected frame- skeleton model did not match the first category. The skeleton model merged with the grayscale image as shown in Figure 12. It is proved that the skeleton model using the proposed approach do not affect by the shadow and the unexpected changes of outdoor illumination. The average of matching accuracy is calculated based on the equation:

$$T_{avg} = \frac{f_D}{f_T} \tag{14}$$

where $f_D$ and $f_T$ are the detected frame and the total frames respectively.

In this experiment, two different environment videos were analyzed. The model produced by proposed approach gives up to 86.19% of the average matching for the videos including the sudden changes of the illumination and shadow at any environment. Meanwhile, 70.95% of the average matching accuracy has been detected by using the normalized filter for the same environment. The matching accuracy for the proposed approach and normalized filter as shown in Table 1.

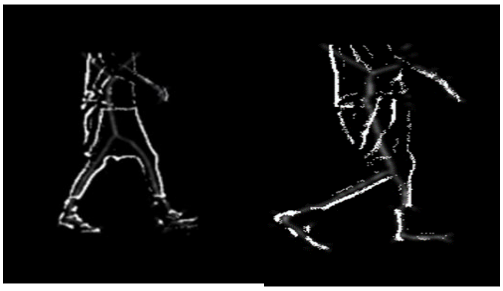


Figure 12. Stitching the skeleton model and the grayscale image

Table 1. Matching accuracy of proposed approach

| Method | | $f_D$ | $f_T$ | Accuracy |
|---|---|---|---|---|
| Proposed approach | Video #1 | 600 | 700 | 85.71% |
| | Video #2 | 780 | 900 | 86.66% |
| Normalized Filter [22] | Video #1 | 480 | 700 | 68.57% |
| | Video #2 | 660 | 900 | 73.33% |

In Table 1, it can be observed that the matching accuracy for the proposed approach can achieve 86% of accuracy, which is higher than normalized filter. Meanwhile, the sensitivity of the proposed approach is measured as the number of detected human in video frames without the influence of the sudden illumination change and shadow [23]. The sensitivity measurement is shown as follows [24]:

$$S = \frac{TP}{TP+FN} \quad (15)$$

where TP is the number of frames detected as a human object when human exists in the frame, and FN is the number of frames detected as a non-human object when human exists in the frame.

Comparison of the sensitivity measurement for proposed approach and normalized filter in two videos as shown in Table 2. It can be seen that the sensitivity of the proposed approach in human detection is much higher than the normalized filter in uncontrolled environment. Therefore, this proposed approach can improve the sudden illumination changing and shadow issues.

Table 2. Sensitivity analysis of proposed approach

| Method | | TP | FN | S |
|---|---|---|---|---|
| Proposed approach | Video #1 | 420 | 30 | 93.3% |
| | Video #2 | 590 | 40 | 93.7% |
| Normalized Filter | Video #1 | 360 | 120 | 75.0% |
| | Video #2 | 490 | 180 | 73.1% |

## 4. Conclusion

A hybrid filter approach consisting of a Gaussian and Median filter is proposed in this paper to detect human motion in an uncontrolled environment. Problems of sudden changes of the illumination at the outdoor environment and the shadow produced during human movement are removed using the hybrid filter. Subsequently, a sequential morphological method is used to construct the skeleton model. With the proposed approach, the shadow of an image is reduced, and the skeleton built are matched with the stereo vision. It has obtained approximately 86% of the average matching accuracy and achieved nearly 94% of the detection sensitivity. In future work, the focus on the occlusion issue will be addressed by three-dimensional human construction and human model tracking with the proposed approach. In addition, multiple stereo cameras are used to increase the distance of the recording to capture the full object information for the tracking purpose.

**Acknowledgement**

The authors would like to acknowledge the financial support from E-Science-Fund (03-02-07-SF0035) of MOSTI Malaysia. We gratefully acknowledge the support of NVIDIA Corporation with the contribution of the TITAN X used for this research.

## References

[1] RA Hamzah and H Ibrahim. Literature survey on stereo vision disparity map algorithms. *J. Sensors*. 2016; 2016.
[2] T Inoue and H Ohzu. Accommodative responses to stereoscopic three-dimensional display. *Appl. Opt.* 1997; 36(19): 4509–4515.
[3] R Khilar, S Chitrakala, SS Parvathy. *3D image reconstruction: Techniques, applications and challenges.* 2013 Int. Conf. Opt. Imaging Sens. Secur. 2013: 1–6.
[4] DL Fernández, FJM Cuevas, AC Poyato, RM Salinas, RM Carnicer. A new approach for multi-view gait recognition on unconstrained paths. *J. Vis. Commun. Image Represent.* 2016.
[5] G Bianco, A Gallo, F Bruno, M Muzzupappa. A Comparison Between Active and Passive Techniques for Underwater 3D Applications. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* 2012; XXXVIII-5: 357–363.
[6] A Leone, G Diraco, C Distante. A Stereo Vision Framework for 3-D Underwater Mosaicking. *Stereo Vis.* 2008: 372.
[7] T Deley. Low-Cost Depth Cameras to Emerge in 2010. *Hizook*. 2010.
[8] K Litomisky. Consumer rgb-d cameras and their applications. *Rapp. Tech.* 2012.
[9] N Uchida, T Shibahara, T Aoki, H Nakajima, K Kobayashi. *3D face recognition using passive stereo vision.* IEEE Int. Conf. Image Process. 2005: 2: 950–953.
[10] V Ganapathi, C Plagemann, D Koller, S Thrun. *Real time motion capture using a single time-of-flight camera.* 2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2010: 755–762.
[11] M Moussa, M Hmila, A Douik. Comparative study of statistical background modeling and subtraction. *Indonesiab Journal of Electical Engineering and Computer Science.* 2017; 8(2): 287–295.
[12] JK I and P Peer. Human Skeleton Model Based Dynamic Features for Walking Speed Invariant Gait Recognition. 2014.
[13] R Chandel and G Gupta. Image Filtering Algorithms and Techniques: A Review. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 2013; 3(10): 2277–128.
[14] L Wang, T Tan, S Member, H Ning, W Hu. Silhouette Analysis-Based Gait Recognition for Human Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* 2003; 25(12): 1505–1518.
[15] F Angella, O Lavialle, P Baylou. *Hierarchical Skeleton Extraction Based on A Deformable Particle System*. Signal Processing Conference (EUSIPCO 1998). 1998.
[16] X Yuan, X Hao, H Chen, X Wei. Background Modeling Method based on 3D Shape Reconstruction Technology. *Indonesian Jounal of Electrical Engineering and Computer Science.* 2013; 11(4): 2079–2083.
[17] R Dahyot, G Lacey, KD Howe, F Pitié, D Moloney. *3D Reconstruction of Reflective Spherical Surfaces from Multiple Images.* IRISH MACHINE VISION & IMAGE PROCESSING Conference proceedings 2015. 2014: 201.
[18] VA Gowri and AC Subhajini. A Flexible Algorithm for Conversion of RGB Image to Gray Image using MATLAB. *Int. J. Control Theory Appl.* 2017; 10(27): 153–161.
[19] EA Castro and DL Donoho. Does Median Filtering Truly Preserve Edges Better Than Linear Filtering. *Ann. Stat.* 2009; 37(3): 1172–1206.
[20] S Beucher. Critical Balls. *Int. Congr. Stereol.* 2013.
[21] G Sanniti and CE Caianiello. *2D Grey-level Skeleton Computation: A Discrete 3D Approach*. International Conference on Pattern Recognition. 2004: 2–5.
[22] W Wu, J Shao, W Guo. *Moving-object Detection Based on Shadow Removal and Prospect*. International Conference on Artificial Intelligence and Soft Computing. 2012; 12(2): 369–374.
[23] DK Singh. *Gaussian Elliptical Fitting based Skin Color Modeling for Human Detection*. in 2017 IEEE 8th Control and System Graduate Research Colloquium. 2017: 197–201.
[24] M Sundaram and A Mani. Face Recognition: Demystification of Multifarious Aspect in Evaluation Metrics. in *Face Recognition - Semisupervised Classification, Subspace Projection and Evaluation Methods*, S. Ramakrishnan, Ed. INTECH. 2016: 75–92.