International Journal of Innovative Computing, Information and Control
Volume **8**, Number **5(B)**, May **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 3565–3584

WAVELET MACH FILTER FOR OMNIDIRECTIONAL HUMAN ACTIVITY RECOGNITION

Tyzzkae Ang[1], Alan Weechiat Tan[1], Chukiong Loo[2] and Waikit Wong[1]

[1]Faculty of Engineering and Technology
Multimedia University
Jalan Ayer Keroh Lama, 75450, Ayer Keroh, Malaysia
{ tkang; wctan; wkwong }@mmu.edu.my

[2]Faculty of Computer Science and Information Technology
University of Malaya
Lembah Pantai, 50603, Kuala Lumpur, Malaysia
ckloo.um@um.edu.my

Received March 2011; revised October 2011

Abstract. *Action recognition is important in the field of intelligent security and surveillance. However, most surveillance cameras can only capture in one direction with limited viewing angle. This paper proposes an edge enhancement template-based method of omnidirectional action recognition that is able to detect specific actions at a 360 degree of view. A MACH filter captures intra-class variability by synthesizing a single action MACH filter for a given action class. The proposed method, based on the wavelet MACH filter, provides additional flexibility of an adaptive choice of wavelet scale factors and, in doing so, enables the selection of the size and orientation of the smoothing function in edge enhancement to optimize the performance of the MACH filter. Moreover, the use of wavelet transform improves the performance of the MACH filter by enhancing the cross-correlation peak intensity in the recognition process. The unwarping of an omnidirectional image into a panoramic image further enables action recognition in 360 degree wide angle of view.*

Keywords: Omnidirectional vision, Log-polar transformation, Action recognition, Maximum average correlation height, Mexican-hat wavelet, 3D normalized cross-correlation

1. Introduction.

Identifying moving objects from a video sequence is a fundamental and critical task in many computer-vision applications. In computer vision research, motion has played an important role for the past thirty years. A major goal of current computer vision research is to recognize and understand human motion [1]. Detection of moving objects in video can be difficult for several reasons. We need to account for possible motion of the camera, changes in illumination of a scene, background objects such as waving trees, or vehicles in motion, etc. Once the moving objects have been identified, tracking them through the video sequence can also be difficult, especially when the objects being tracked are occluded by buildings, or move in and out of the frame due to the motion of the camera [2]. Currently, the main detection algorithms include frame difference method, background subtraction method, optical flow method and statistical learning method [3-7]. Each method has its own advantages and disadvantages. For example, the frame difference and background subtraction methods are simple and easy to implement but are sensitive to illumination changes and changes in the background geometry. Optical flow and statistical learning methods provide higher accuracy in moving object detection but involve complex computation. Zhan [8] introduced an improved moving object detection based on frame difference and edge detection. Combining frame difference with edge detection reduces the sensitivity to the change of light, and thus, improves overall detection accuracy. However, the edge detector is incredibly sensitive to noise in pictures and highlights them as edges. This drawback can be resolved using wavelet filter. In recent study, M. A. R. Ahad [9] proposed the directional motion templates method to solve the motion self-occlusion problem. However, their study is limited to conventional camera that is only able to capture at a certain angle of image view. For omnidirectional viewing, a common method is to use multiple cameras. This method is easy to implement but requires at least two cameras, and each is not without its blind spot. To obtain a 360 degree viewing angle without blind spots, a single camera can also be used [10]. In this paper, the proposed wide FOV surveillance based on an Internet Protocol (IP) camera on a rotational platform is proposed to automatically detect abandoned and removed objects. As the camera rotates and captures continuously at the center, the images obtained are matched and concatenated to form a panoramic image. However, since it is necessary to rotate a camera at a full circle in order to obtain a single panoramic image, it is therefore impossible to generate video rate panoramic images. The third method does not need a full rotation to generate panoramic image but requires multiple cameras installed in a circle facing the surrounding. All captured images are then matched and concatenated as before. Nevertheless, working with a perspective camera, collecting such a large number of images is tedious and time consuming [11]. Yet another method is to use an omnidirectional camera. This camera mounts a convex mirror with a black needle at the apex of the mirror to avoid internal reflections on the glass cylinder [12]. The advantage of such a device provides the observer with a complete view of its surrounding in a single image which can be acquired at video rate. Omnidirectional visualization has significant application potential in areas such as security where it is important to have a simultaneous view of an entire area, or robotic navigation, where the advantage lies in the use of just one camera to obtain a complete visualization of the surrounding environment [13]. By applying log-polar transformation, an omnidirectional image can be converted into a panoramic image and, correspondingly, it becomes possible to generate panoramic videos.

Template matching by means of cross-correlation is a common practice in target detection and recognition problems. Object recognition is performed by correlating an input image with a synthesized template where high correlation points to the presence of the object of interest. For image detection, matched spatial filter (MSF) maximizes the output signal-to-noise ratio for the detection of a known image. However, MSF is very sensitive to deformations of the pattern in the reference image and the broad and unsharp peaks it produces are significant drawbacks [14]. These problems motivate the use of distortion invariant composite filters such as the Minimum Average Correlation Energy (MACE) filter and the Maximum Average Correlation Height (MACH) filter. These filters generalize from a collection of examples to create a single template that captures the intra-class variability of an action. Even though the MACE filter is a popular correlation filter for pattern recognition due to its sharp peak correlation, the MACH filter has a much higher peak correlation which is more suited to our work. Finally, wavelet filter have been introduced to the MACH filter. Unlike, Minimum Variance Synthetic Discriminant Function (MVSDF) [15], it has taken into consideration additive input noise. MACH filter does not have any special mechanism to consider input noise. Noise tolerance is greatly concerned in log-polar transformed video sequence. Therefore, wavelet filter can be used to remove this drawback.

Wavelet transform has been attracting increasing attention in the optical pattern recognition community because of its attractive multiresolution, denoising, and feature extraction capabilities [16,17]. Due to the edge-enhancement property, the wavelet-matched filter produces a sharper correlation and, thus, improves the discrimination capability of the conventional matched spatial filter [18]. Goyal et al. [19] proposed the waveMACH filter as a combination of the wavelet transform and the MACH filter for recognition of 0-360 degree out-of-plane rotated targets employing hybrid digital-optical correlator architecture. Based on the same idea, we exploit the characteristics of the wavelet transform to improve the conventional MACH filter in our study.

In this paper, we incorporate logarithmic-polar mapping in the wavelet MACH filter for omnidirectional human activity recognition. In doing this, the proposed method is capable of capturing intra-class variability by synthesizing a single action wavelet MACH filter for a given action class for wide angle images using omnidirectional cameras. The paper is organized as follows. In Section 2, the basic principles of log-polar mapping and the proposed MACH algorithms are detailed. All experiments and results are discussed in Section 3 followed by the conclusion in Section 4.

1.1. Motivation.

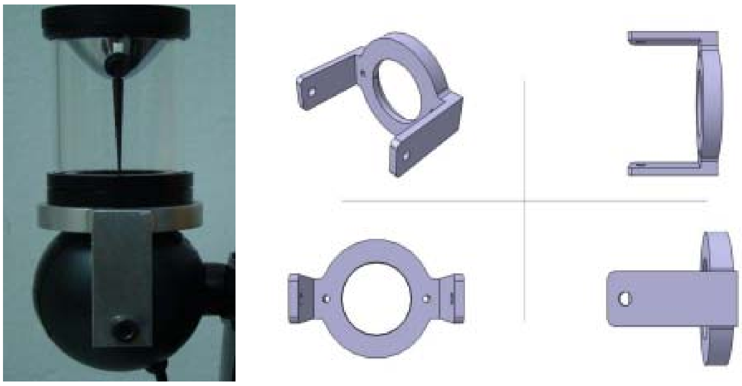
Most of the surveillance systems are only used in recording and monitoring purposes and it is very time consuming and impractical to monitor every moving objects in a surveillance area or playback when there is an incident occurring. Therefore, there is a need to develop a system which is able to track and recognize specific action for security purposes.

A specific action at a surrounding can be for safety and surveillance purposes, for example, a fall detection of old folks. Since some of the elderly are more likely to experience fall accidents than other age groups, we need to give special attention to them. In the case of elderly people living on their own, there is a particular need for monitoring their behavior. Many elders fall and are unable to get back up or call for assistance, while they may lie on the ground hurt until someone comes along and discovers them. Nurses or other qualified personnel are the best candidates to provide the care. However, manpower shortage is a common problem everywhere. In this case, surveillance systems can be chosen as an option. An effective fall detection algorithm can be designed to solve the manpower shortage problem. Therefore, the idea is good to implement into centers such as old folks' home, hospital, and handicapped center. This method also can be used as intruder detection. Traditionally we may need at least two cameras to monitor every corner of a room. By applying omnidirectional method, we are able to monitor a 360 degree surrounding view with single omnidirectional camera.

2. Approach and Methods.

2.1. Omnidirectional camera.

Most cameras only capture in one direction with limited viewing angle. Although there are cameras that can capture at wider angles, they are not able to capture in all directions in one view. Greiffenhagen et al. proposed a surveillance system by combining a CCD camera and a hyperbolic mirror for capturing panoramic scenes [20]. In this paper, an E3500 PLUS QUICKCAM costing around US$25 is used in the surveillance system. The webcam is able to capture high-quality VGA (640×480) video and 1.3-megapixel (software-enhanced) images. It is small in size and cheap in comparison with digital and CCTV cameras with fine output resolution. The digital control is also accomplished through the USB port connected to a laptop or PC via plug-and-play on Windows OS. The omnidirectional camera used in the research consists of a webcam and is attached with a hyperbolic mirror. The specifically designed hyperbolic optical mirror used in omnidirectional surveillance system is a small-sized wide-view type, with outer diameter 40mm and angle of view 30 degree above the horizontal plane. The mirror can reflect a 360 degree view of the surrounding and the omnidirectional image acquired is sent to a computer in a monitoring room to be processed for surveillance purposes. The advantage of this omnidirectional camera is that it is cost efficient because webcam is cheaper than digital or CCTV camera and hyperbolic mirror is cheaper than fish-eye lens with almost the same reflective quality. Figure 1(a) shows the design of low cost omnidirectional camera and Figure 1(b) is custom-made bracket that is designed to attach hyperbolic mirror to the webcam. Figure 2 shows the respective captured omnidirectional image.



(a) (b)

Figure 1. (a) is the design of low cost omnidirectional camera and (b) is isometric, back and side view of the custom-made bracket

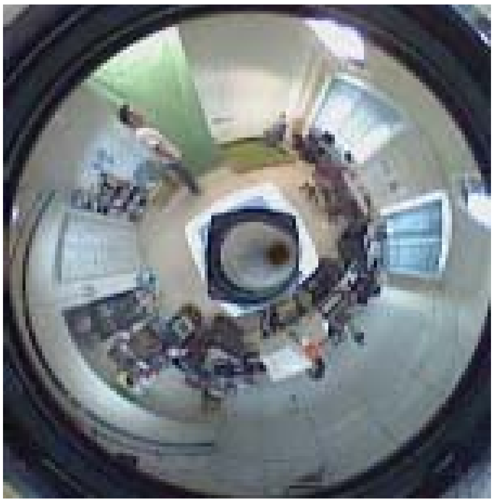


Figure 2. Image acquired by omnidirectional webcam

2.2. **Logarithmic-polar (log-polar) mapping.** An omnidirectional image is obtained using the omnidirectional camera in Figure 1. The geometry of the captured omnidirectional image is in Cartesian form (x_1, y_1). Log-polar sampling is used to sample the Cartesian omnidirectional images into log-polar form omnidirectional image. The polar coordinates (ρ, θ) correspond to radial distance from the center and angle, respectively [21]. In the log-polar domain, the coordinate $\log \rho$ corresponds to the logarithmic distance from the center of the image to a given point and θ corresponds to the angle of the point with the positive x_1-axis. Figure 3 illustrates how the regions of the rectangular grid in the log-polar domain in Cartesian coordinates as concentric rings whose width is proportional to their distance from the origin of the mapping. Each ring is then divided into a certain number of receptive fields. The receptive fields that lie on a radius are mapped on vertical lines on the (x_2, y_2) plane, while those lying on concentric circumferences centered on the origin are mapped on horizontal lines [22]. In this process, the omnidirectional image is unwarped into a panoramic image. Since the panoramic image is in Cartesian form, subsequent image processing task becomes much easier [31].

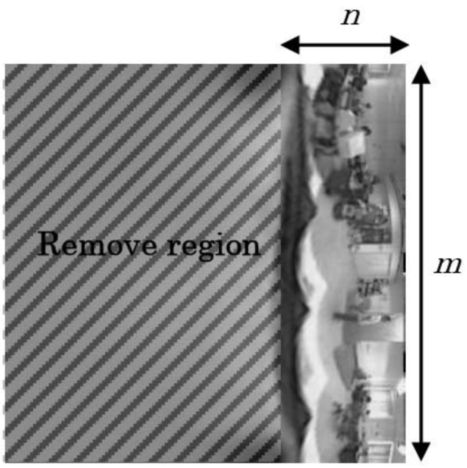


Figure 3. Log-polar transform (sampling and mapping)

The log-polar transformations are described by the pair of equations [23]: Specifically, (1) and (2) express the logarithmic and the polar form of the mapping.

$$\rho = \log_\lambda \frac{\sqrt{x_1^2 + y_1^2}}{r_o} \tag{1}$$

$$\theta = \frac{\Theta}{2\pi} \tan^{-1}\left(\frac{y_1}{x_1}\right) \tag{2}$$

The parameter:

$$\lambda = \frac{1 + \sin\dfrac{\pi}{\Theta}}{1 - \sin\dfrac{\pi}{\Theta}} \tag{3}$$

is the base of the logarithm, r_o is the scaling factor which defines the size of the $\rho = 0$ circle and Θ is the total number of pixels per ring. The centremost pixel in the log-polar mapping is given by [23]:

$$x_2 = \lambda^\rho r_o \cos\left(\frac{2\pi\theta}{\Theta}\right) \tag{4}$$

$$y_2 = \lambda^\rho r_o \sin\left(\frac{2\pi\theta}{\Theta}\right) \tag{5}$$

The resolution of the log-polar mapped image is $(m \times n)$ pixels, where n is the total number of concentric rings and $m = \lfloor \pi n \rfloor$ is the optimal circumference at half of the total number of n concentric rings. The log-polar transformed image of Figure 2 is shown in Figure 4. Figure 2 is the image captured by QUICKCAM through the hyperbolic optical mirror. Rotation and reflection transformations have been applied so that the generation of panoramic images allows using existing image processing. It also allows a human to see familiar panorama images instead of an unfamiliar omnidirectional input image deformed by the hyperbolic mirror.

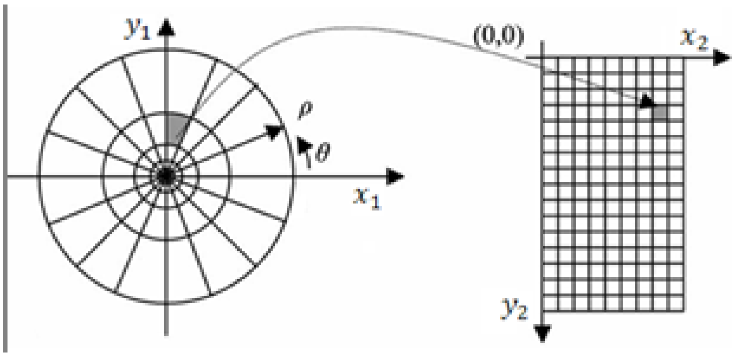


Figure 4. Resultant image from Figure 2 after the log-polar transformation

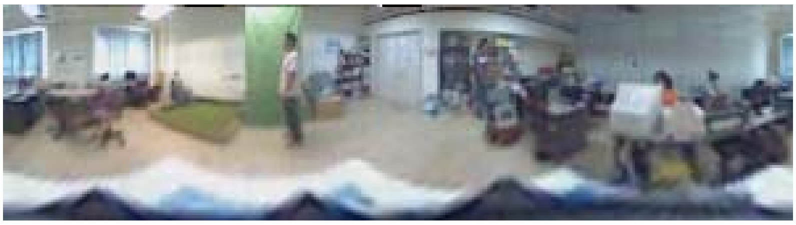


Figure 5. Final panoramic image

2.3. **Edge detection.** In the pre-processing stage, steps are taken to improve the quality of the panoramic image. The Gaussian smoothing operator is a 2D convolution operator that is used to blur images and remove detail and noise. Gaussian smoothing has the similarity of a mean filter, but uses a different function to calculate the pixel value. In 2D, an isotropic Gaussian has the form [24]:

$$G(x,y) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6}$$

where σ is the standard deviation of the distribution. The filter is used to remove small scale texture and noise for a given spatial extent in the image.

Edge detection is a problem of fundamental importance in image analysis. In typical images, edges characterize object boundaries and are therefore useful for segmentation, registration, and identification of objects in a scene. There are many methods of detecting edges. The majority of different methods may be grouped into these two categories, sobel and laplacian [25]. The sobel method detects the edges by looking for the maximum and minimum in the first derivative of the image. The laplacian method searches for zero crossings in the second derivative of the image for the edges. In our work, we have chosen the Sobel edge detector. The implementation of the Sobel edge detector follows as

$$N(x,y) = \sum_{k=-1}^{1}\sum_{j=-1}^{1} K(j,k)p(x-j,y-k) \tag{7}$$

where K is a pair of 3×3 convolution masks and p is the pixel of image. The gradient magnitude and direction are, respectively:

$$|g| = \sqrt{g_x^2 + g_y^2} \tag{8}$$

$$\theta = \tan^{-1}\left(\frac{g_y}{g_x}\right) \tag{9}$$

Prior to actual processing, we have used the Gaussian filter to smoothen the image and the 3 × 3 sobel edge detector on the smoothen image.

2.4. **Motion detection.** Background disturbance may cause template-based action detection to be inaccurate. Specifically, this usually occurs when different training actions are trained in different background. To overcome this problem, the frame difference method is applied so that background scenes can be neglected. Frame difference calculates the difference between two frames at every pixel position and stores the absolute difference. It is used to visualize the moving objects as a sequence of frames [24]. Frame difference, however, is susceptible to changes in the background brightness. To overcome that, we proposed to combine frame difference with edge detection. This approach not only retains the small impregnability of light from edge detection method, but also improves in noise removal. The approach, thus, has better noise suppression capability and higher detection accuracy [8]. The combination of frame difference and edge detection produces a smooth boundary of a moving object. The detection process is shown in Figure 6.

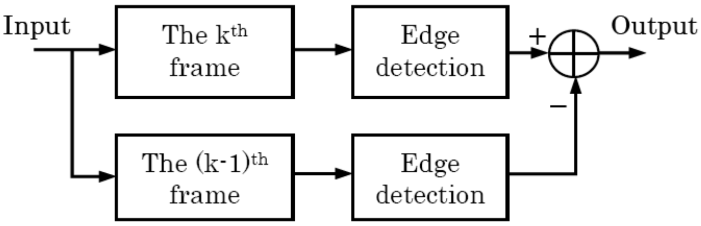


FIGURE 6. Block diagram of moving object detection

2.5. **Maximum average correlation height (MACH) filter.** The MACH filter is a method of creating invariance to distortions in the target object through introducing the expected distortions in the construction of the filter. The MACH filter is derived by maximizing a performance metric called the Average Correlation Height (ACH). Three other performance measures, Average Correlation Energy (ACE), Average Similarity Measure (ASM) and Output Noise Variance (ONV), have to be balanced to better suit different application scenarios. The following energy function is formed and minimized with respect to one criterion, holding the others constant [26]:

$$\begin{aligned} E(h) &= \alpha(ONV) + \beta(ACE) + \gamma(ASM) - \delta(ACH) \\ &= \alpha h^T C h + \beta h^T D_x h + \gamma h^T S_x h - \delta|h^T m_x| \end{aligned} \tag{10}$$

where α, β and γ are non-negative optimal trade-off parameters and m_x is the average of the training image vector $x_1, x_2, \ldots, x_N$ (in the frequency-domain). The matrix C is the diagonal power spectral density matrix of additive input noise, usually set as the white noise covariance matrix, i.e., $C = \sigma^2 I$, where σ is the standard deviation parameter and I is an identity matrix. The matrix D_x is the diagonal average power spectral density of the training images.

$$D_x = \frac{1}{N_e}\sum_{i=1}^{N_e} X_i^* X_i \tag{11}$$

where N_e is the total number of sample actions in the training dataset, X_i is diagonal matrix of the i^{th} training image and $*$ represents the conjugate operation. Lastly, the matrix S_x denotes the similarity matrix of the training images:

$$S_x = \frac{1}{N_e}\sum_{i=1}^{N_e}(X_i - M_x)^*(X_i - M_x) \tag{12}$$

where M_x is the average of X_i. The MACH filter can be synthesized in the frequency-domain as [27]:

$$h = (\alpha C + \beta D_x + \gamma S_x)^{-1} m_x \tag{13}$$

The different values of α, β and γ control the MACH filter's behavior to match different application requirements [28].

2.6. **Wavelet transform.** The wavelets $h_{a,b}(x)$, which are the basis functions of the wavelet transforms, are generated from the mother wavelet $h(x)$ by dilations and translations [18]:

$$h_{a,b}(x) = \frac{1}{\sqrt{a}} h\left(\frac{x-b}{a}\right) \tag{14}$$

where a is the dilation factor and b is the translation factor. The mother wavelet must satisfy the admissibility conditions in that it must be oscillatory, decays quickly to zero, and integrates to zero. The wavelet transform is defined as an inner product between the signal $f(x)$ and a set of wavelets $h_{a,b}(x)$, and is given as:

$$W_f(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} h^*\left(\frac{x-b}{a}\right) f(x)dx \tag{15}$$

where $*$ denotes the complex conjugate and $W_f(a,b)$ can be considered as a function of spatial shift b for each fixed scale a that displays the information $f(x)$ at various resolution levels. In the frequency-domain the wavelets are expressed as:

$$\begin{aligned} H_{s,b}(f) &= \int_{-\infty}^{\infty} \exp(-i2\pi f x) h_{a,b}(x) dx \\ &= \sqrt{a}\exp(-i2\pi f b) H(af) \end{aligned} \tag{16}$$

where $H(f)$ is the Fourier transform of the mother wavelet $h(x)$. Various types of wavelet functions have been defined [18]. The Mexican hat wavelet and the Morlet wavelet are the most widely used wavelets because their Fourier spectra are real valued and symmetrical. Mexican hat wavelet is well known as the Laplacian operator and, in this study, a 3D Mexican-hat wavelet is used. The 3D Mexican-hat wavelet is the second derivative of the Gaussian function and is given as:

$$\begin{aligned} h(x,y,t) &= \frac{1}{s^2}\left[\left(\frac{x^2+y^2+t^2}{s^2} - 2\right)\right] \\ &\quad \times \exp\left(-\frac{x^2+y^2+t^2}{2s^2}\right) \end{aligned} \tag{17}$$

The corresponding Fourier transform of the 3D Mexican-hat wavelet is

$$\begin{aligned} H(u,v,w) &= 4\pi^2 s^2(u^2+v^2+w^2) \\ &\quad \times \exp[-2s^2(u^2+v^2+w^2)] \end{aligned} \tag{18}$$

The performance of the wavelet MACH filter depends on the scale s. At smaller value of s, the high-frequency components of the image is retained and the resulting wavelet MACH filter has good discrimination capability but may be more sensitive to noise. With larger values of s, the wavelet MACH filter retains the low-frequency components of the image and suffers from a poor discrimination capability but is less sensitive to noise.

2.7. **Wavelet MACH filter.** The training process involves the creation of a series of spatio-temporal volumes of test action sequences by concatenating the frames of a complete cycle of an action. The temporal derivative of each pixel is subsequently computed and the result is a volume for each training sequence. Following the construction of the spatio-temporal volumes for each action in the training set, we proceed to represent each volume in the frequency-domain by performing a 3D FFT operation given by [27]:

$$F(u,v,w)=\sum_{t=0}^{N-1}\sum_{y=0}^{M-1}\sum_{t=0}^{L-1}f(x,y,t)\exp\left(-j2\pi\left(\frac{ux}{L}+\frac{vy}{M}+\frac{wt}{N}\right)\right) \tag{19}$$

where $f(x,y,t)$ is the volume corresponding to the temporal derivative of the input sequence, $F(u,v,w)$ is the resulting volume in the frequency-domain, L is the number of columns, M is the number of rows and N is the number of frames. Following that, the frequency-domain volume is converted into a column vector by concatenating all the columns of the 3D matrix. Once the column vectors are obtained, the MACH filter can then be synthesized in the frequency-domain as in Equation (18). The wavelet transform is then used to convert the MACH filter into a wavelet MACH filter. Specifically, the wavelet MACH filter is expressed as $(\alpha C+\beta D_x+\gamma S_x)^{-1}m_x|H(u,v,w)|$. Subsequently, 3D IFFT is performed to convert the wavelet MACH filter to time-domain. The resulting matrix constitutes the wavelet MACH filter for the particular action. The flow chart of this design procedure is shown in Figure 7.

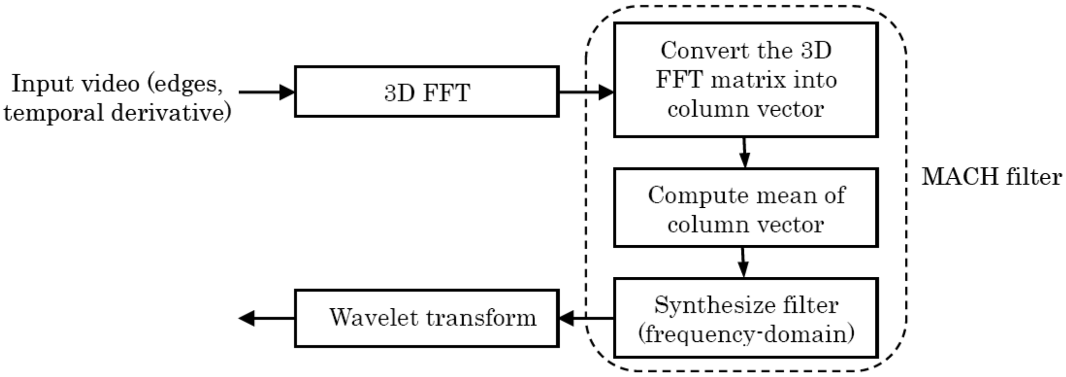

Input video (edges, temporal derivative)
3D FFT
Convert the 3D FFT matrix into column vector
Compute mean of column vector
Synthesize filter (frequency-domain)
Wavelet transform
MACH filter

FIGURE 7. Flowchart of the wavelet MACH filter

2.8. **3D fast normalized cross-correlation.** Once the wavelet MACH filter is synthesized, it can be used to detect similar actions by applying the 3D Fast Normalized Cross-Correlation [29,32]. Let I be the $M_x\times M_y\times M_t$ input volume and T be the $N_x\times N_y\times N_t$ template volume where $N_x<M_x$, $N_y<M_y$ and $N_t<M_t$. The 3D Fast Normalized Cross-Correlation is given as:

$$C(u,v,w)=\frac{\sum_{x,y,t}(I(u+x,v+y,w+t)-\overline{I_{u,v,w}})\hat{T}(x,y,t)}{\sqrt{\sum_{x,y,t}(I(u+x,v+y,w+t)-\overline{I_{u,v,w}})^2\sum_{x,y,t}\hat{T}(x,y,t)^2}} \tag{20}$$

where $x\in\{0,\dots,M_x-1\}$, $y\in\{0,\dots,M_y-1\}$, $t\in\{0,\dots,M_t-1\}$, $u\in\{0,1,\dots,M_x-N_x\}$, $v\in\{0,1,\dots,M_y-N_y\}$, $w\in\{0,1,\dots,M_t-N_t\}$, $\overline{I_{u,v,w}}$ is the local image mean at location (u,v,w), and $\hat{T}(x,y,t)=T(x,y,t)-\bar{T}$ is the mean-subtracted template.

As the result of this operation, we obtain a response, C of size $(L-P+1)\times(M-Q+1)\times(N-R+1)$. The response of the normalized correlation lies within the interval $[-1,1]$. The peak value in the correlation is then compared with a threshold τ. If this value is

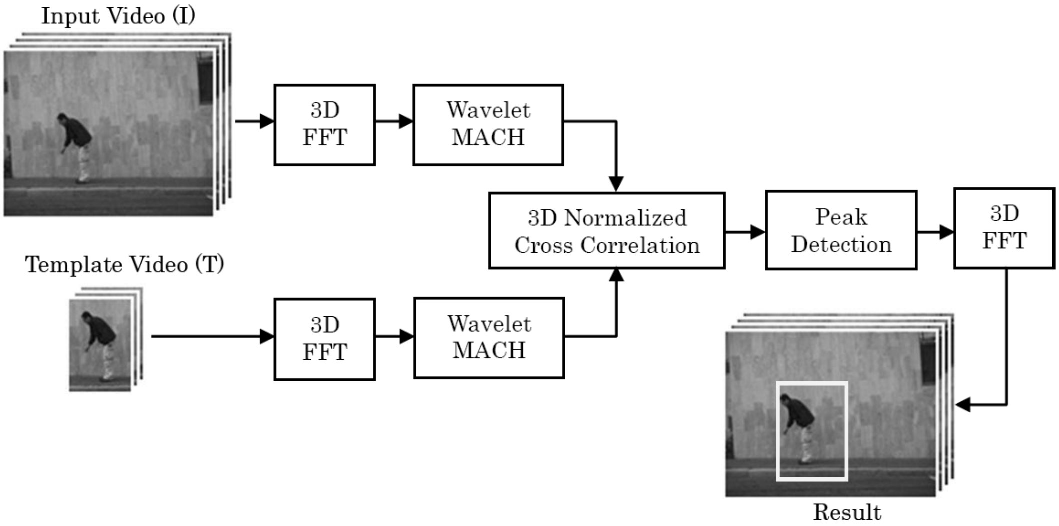


Figure 8. Flowchart of peak detection

greater than τ, the action is identified and a rectangular bounding box is plotted in the test video. In the experiments, the threshold is set as $\tau = \varepsilon * \min(p_1, p_2, \cdots, p_N)$, where p_i is the peak value obtained from the correlation response of the i^{th} training video, ε is a constant parameter, and N is the number of training videos [27]. The peak detection procedure is illustrated in Figure 8.

3. **Experiments and Results.**

An extensive set of experiments is performed to evaluate the performance of the proposed method on a publicly available dataset and a collection of actions acquired by an omnidirectional camera. In the action dataset acquisition, it is assumed that the experiment is made in a controlled environment and that the actions made by the actors are consistent. The results from these experiments are detailed in the following. The performance of the wavelet MACH filter very much depends on the scale s. Roberge [18] reported that with a small value of scale s the wavelet transform has good discrimination capability but is vulnerable to noise. When the scale s increases, the wavelet transform becomes more robust to noise but loses the discrimination capability. False detections may also be observed when the scale s is set too large. The scale is chosen, after several experimental trials, to be 1.3 for the wavelet filter to yield the optimal result.

3.1. **Weizmann action dataset.**

In the first simulation, the Weizmann action dataset [30] is used. The dataset contains nine people performing five different actions, namely, “bending”, “jumping-jack”, “waving-one-hand”, “waving-two-hands”, and “jumping-at-same-point”. These actions are recorded as (180 × 144 pixels, 25 fps) video sequences. Each video sequence contains one actor repeatedly performing an action. The five actions are illustrated in Figure 9.

The data from this collection was partitioned for training and testing. Three sets are selected for training and the remaining six sets are selected for testing. The performance of the MACH filter and wavelet MACH filter were tested between the bending and waving-one-hand action. In the experiments, the action to be detected is a bending action, whereas the action to be rejected is the waving-one-hand action which has a very similar silhouette to the bending action.

The performance of the wavelet MACH filter very much depends on the scale s. Roberge et al. reported that with a small value of scale s the wavelet transform has a good discrimination capability but is vulnerable to noise. When the scale s increases, the wavelet

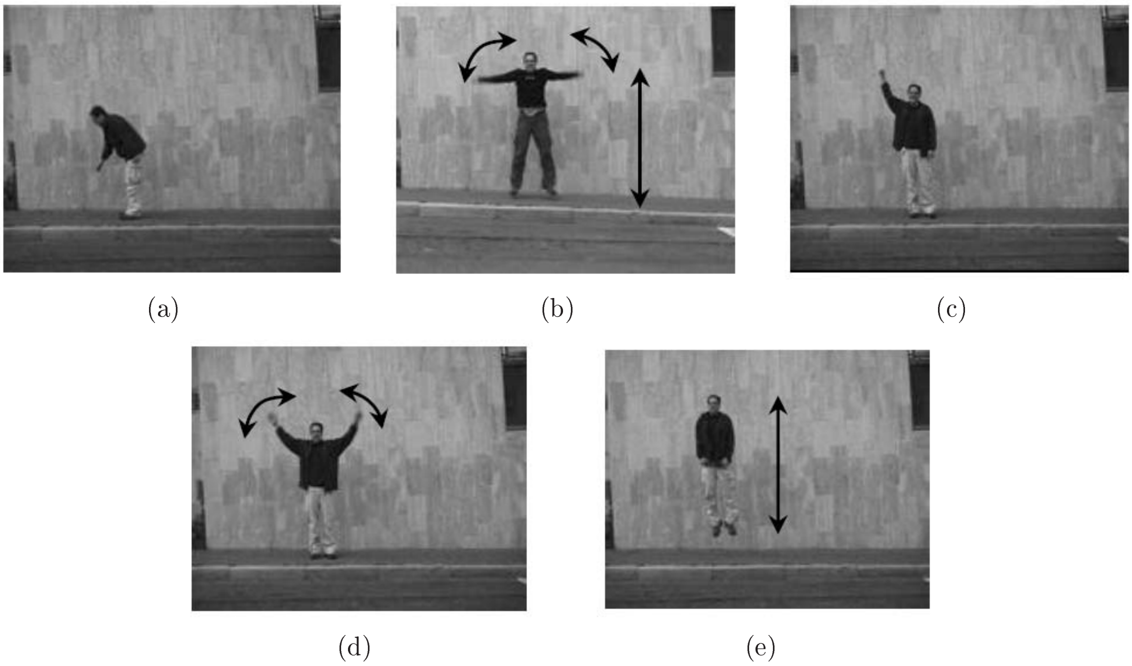


Figure 9. Weizmann action dataset: a) bending, b) jumping-jack, c) waving-one-hand, d) waving-two-hands, and e) jumping-at-same-point

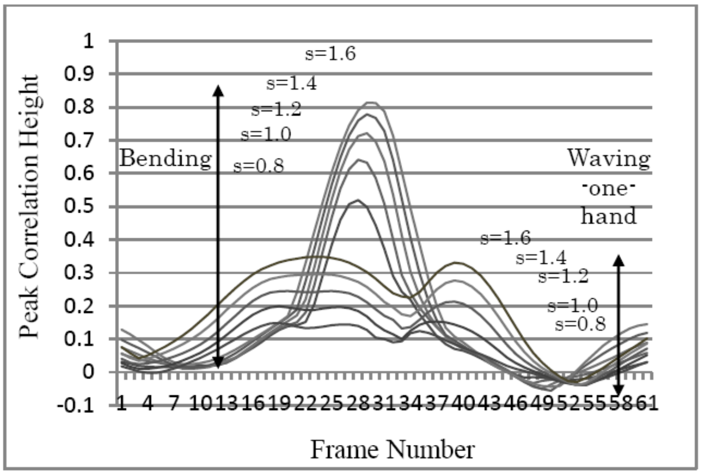


Figure 10. The comparison PCH of bending and waving-one-hand detection with different scale s

transform becomes more robust to noise but loses the discrimination capability. False detections may also be observed when the scale s is set too large. The first part of the experiment, therefore, aims to investigate the most suitable value of s in human action recognition. The bending and waving-one-hand actions are used in the test. The Peak Correlation Height (PCH) at different scale s is investigated. As shown in Figure 10, when scale s increases, the PCH of bending and waving-one-hand action increases respectively.

In general, the performance of the wavelet MACH filter increases with the increase of scale s. When scale s is small, the PCH is low and thus affects the region of detection, but when s is too big, it will cause false detection. In order to find the optimal scale, the Peak Correlation Height Difference (PCHD) of bending and waving-one-hand actions is studied. In Figure 11, the PCHD begins to decrease after scale 1.3. After this point,

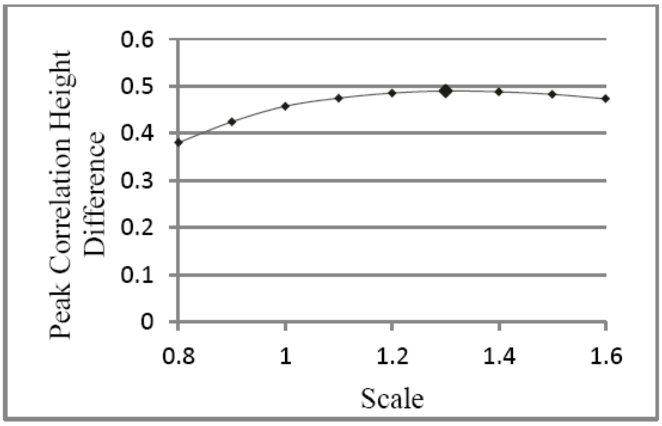


Figure 11. PCHD between bending and waving-one-hand action

the wavelet MACH filter starts to lose its discrimination capability. Thus, the best scale that produces optimal recognition in this case is $s = 1.3$. As a conclusion, when scale s is smaller than the optimal scale, the PCH and PCHD is decreases but when the scale s is greater than the optimal scale, the PCH increases but the PCHD is decreases.

3.2. **Comparison performance analysis.** Among the five actions in Figure 9, bending action is used to assess the performance of the filter. In the simulations, the performance is measured by observing the peak correlation height (PCH). Figure 12 shows the PCH of the bending detection using the MACH filter and Figure 13 shows the PCH of the bending detection using the wavelet MACH filter. The PCH of the MACH filter is at 0.4986 while the PCH of the wavelet MACH filter is at 0.7159, which is much higher than the conventional MACH filter which is used in [27].

As shown in the above figures, the peak correlation height difference (PCHD) for the wavelet MACH filter is much higher than that of the MACH filter. Confusion matrix tables below are additionally used to evaluate the accuracy of the filter. For this purpose, in each of the actions, 90 different combinations of training and testing action sample are assessed. For example, in case of 90 jumping-jack actions in Table 1, 9 samples were having false detection on MACH filter while 100% accuracy on wavelet MACH filter. Table 2 shows that the proposed wavelet MACH filter has better discriminability in accept/reject actions, thus, reduces false detection. The geometry mean for MACH filter is 88.40%

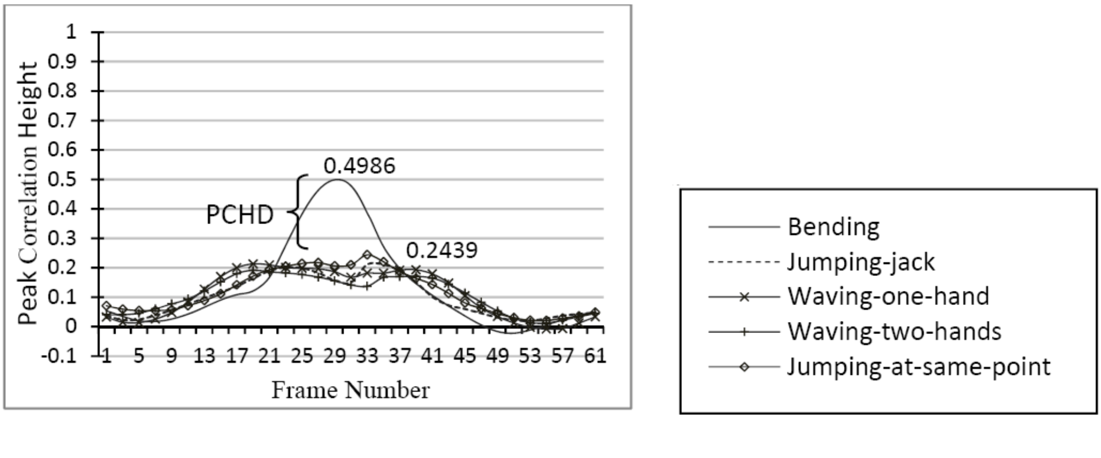


Figure 12. PCH of bending detection with MACH filter

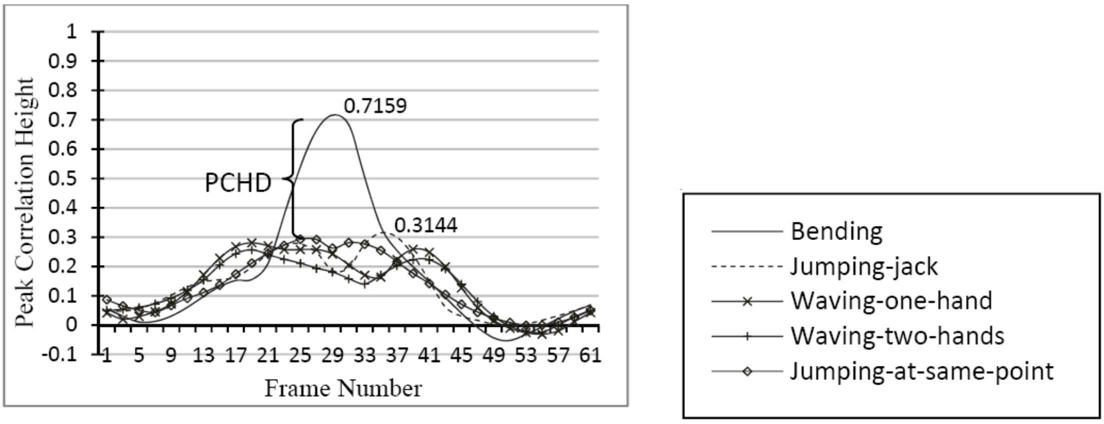


Figure 13. PCH of bending detection with wavelet MACH filter

whereas, the mean accuracy for wavelet MACH filter is 96.03%, that is, an improvement of 7.63%. These results affirm the overall superiority of the wavelet MACH filter over the conventional MACH filter.

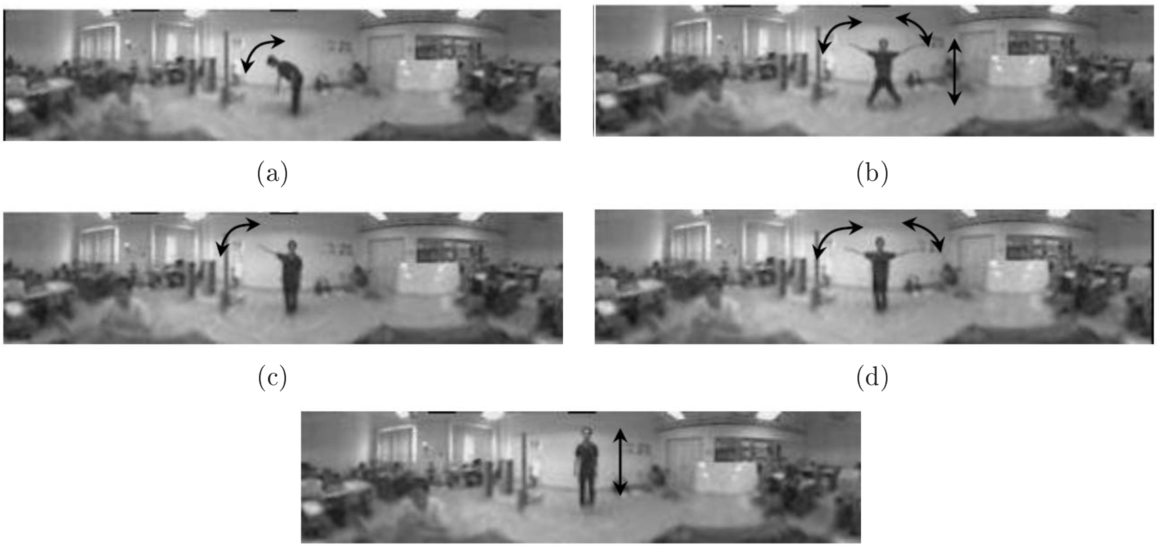
Tables 3 and 4 shows the PCH of bending detection on MACH filter and wavelet MACH filter respectively. The difference between MACH filter and wavelet MACH filter is that the MACH filter has higher PCH. A higher PCH indicates that there is more similarity between the template and the test action, and this gives rise to an increased PCHD, and so the system can better discriminate between the accepted and the rejected actions. This is proof that the wavelet MACH filter is more accurate as compared with the MACH filter.

3.3. **Panoramic action dataset.** In the next simulation, the same five action classes in Section 3.1 are collected by the omnidirectional camera. After performing log-polar mapping, these actions are shown in Figure 14. These video sequences will be used to assess the performance of the proposed filter in panoramic log-polar mapped videos. In each of the actions, 40 different combinations of training and testing action samples are assessed.

As in the previous simulation, the PCH of the wavelet MACH filter is also higher than the MACH filter in log-polar transformed panoramic videos (see Figure 16). The PCH of the MACH filter is 0.5854 while the PCH of the wavelet MACH filter is 0.8642. Comparing the confusion matrices in Tables 5 and 6, the overall recognition rate of the wavelet MACH filter is also better with lower false detection rate than the MACH filter.

Based on the results in Table 3, the detection of the "bending" and "jumping-jack" actions are perfect, whereas high false detection are observed for the other actions. This may be due to the fact that action images acquired in panoramic video format are, in general, small, and so making the motion sequences of dissimilar actions appear alike. This drawback can be somewhat resolved by applying the wavelet filter (compare Tables 3 and 4).

Based on Table 6, it is clear that the detection of the "bending", "jumping-jack, "waving-two-hands" and "jumping-at-same-point" actions is good. However, high false detection is observed for the "waving-one-hand" action in which all the 8 "waving-two-hands" action samples were falsely detected as the "waving-one-hand" action. Referring to Figure 17, it is apparent that both of these actions are actually rather similar in that the "waving-one-hand" action is one half of the "waving-two-hands" action. As a result, the correlation filter returned a high peak correlation. The reverse is however not true, i.e.,



(a)

(b)

(c)

(d)

(e)

Figure 14. Panoramic action dataset: a) bending, b) jumping-jack, c) waving-one-hand, d) waving-two-hands, and e) jumping-at-same-point

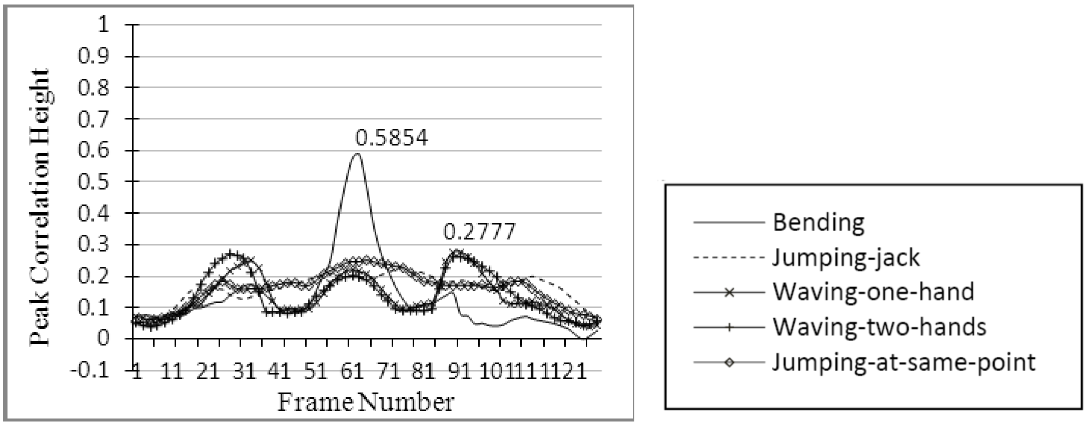


Figure 15. PCH of bending detection with MACH filter on panoramic video

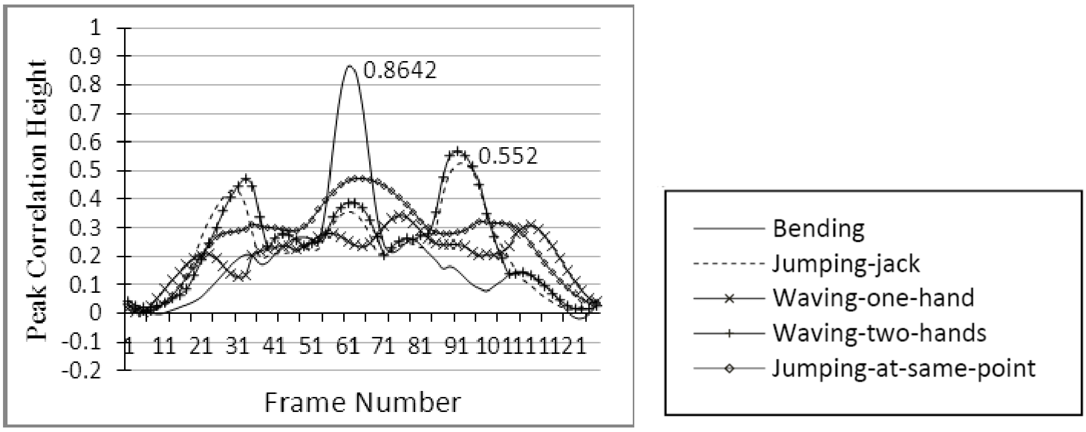


Figure 16. PCH of bending detection with wavelet MACH filter on panoramic video

Table 1. 2-class confusion matrix for the MACH filter for the Weizmann action dataset. Geometric mean is 88.40%.

(a) Bending

		Predicted	
		Real	False
Actual	Real	15	0
	False	3	72

(b) Jumping-jack

		Predicted	
		Real	False
Actual	Real	18	9
	False	0	63

(c) Waving-one-hand

		Predicted	
		Real	False
Actual	Real	18	2
	False	0	70

(d) Waving-two-hands

		Predicted	
		Real	False
Actual	Real	16	0
	False	2	72

(e) Jumping-at-same-point

		Predicted	
		Real	False
Actual	Real	13	0
	False	5	70

Table 2. 2-class confusion matrix for the wavelet MACH filter for the Weizmann action dataset. Geometric mean is 96.03%.

(a) Bending

		Predicted	
		Real	False
Actual	Real	16	0
	False	2	72

(b) Jumping-jack

		Predicted	
		Real	False
Actual	Real	18	0
	False	0	72

(c) Waving-one-hand

		Predicted	
		Real	False
Actual	Real	17	0
	False	1	72

(d) Waving-two-hands

		Predicted	
		Real	False
Actual	Real	16	0
	False	2	72

(e) Jumping-at-same-point

		Predicted	
		Real	False
Actual	Real	16	0
	False	2	72

Table 3. PCH of bending detection on MACH filter (Weizmanna action dataset)

Train	Test	Ido	Ira	Lena	Lyova	Moshe	Shahar
Bend	**Bend**	**0.606**	**0.6036**	**0.5861**	**0.6072**	**0.5752**	**0.5544**
	Jack	0.3142	0.2598	0.293	0.2762	0.2995	0.2667
	Wave1	0.3631	0.3631	0.3595	0.3321	0.3117	0.3373
	Wave2	0.2994	0.2746	0.266	0.2918	0.3067	0.3142
	pJump	0.3135	0.3038	0.2705	0.301	0.3171	0.3584

"waving-two-hand" action will not be falsely detected as the "waving-one-hand" action. The geometric mean for MACH filter is 80.98% whereas, the geometric mean for wavelet MACH filter is 90.35%, or an improvement of 9.37%. Table 7 shows the improvement of proposed wavelet MACH filter against the conventional MACH filter which used in [27] on Weizmann action dataset. Table 8 compares MACH filter and wavelet MACH filter on panoramic action dataset.

TABLE 4. PCH of bending detection on wavelet MACH filter (Weizmann action dataset)

Train	Test	Ido	Ira	Lena	Lyova	Moshe	Shahar
Bend	**Bend**	**0.727**	**0.809**	**0.7441**	**0.8028**	**0.6879**	**0.7157**
	Jack	0.3187	0.3063	0.3611	0.3122	0.3149	0.3202
	Wave1	0.3662	0.3158	0.3361	0.2922	0.3067	0.3377
	Wave2	0.3408	0.3172	0.3113	0.3086	0.3364	0.3212
	pJump	0.3195	0.3282	0.3212	0.3184	0.3293	0.3989

TABLE 5. 2-class confusion matrices using conventional MACH filter for the panoramic action dataset. Mean accuracy is 80.98%.

(a) Bending

		Predicted	
		Real	False
Actual	Real	8	0
	False	0	32

(b) Jumping-jack

		Predicted	
		Real	False
Actual	Real	8	0
	False	0	32

(c) Waving-one-hand

		Predicted	
		Real	False
Actual	Real	8	11
	False	0	21

(d) Waving-two-hands

		Predicted	
		Real	False
Actual	Real	8	5
	False	0	27

(e) Jumping-at-same-point

		Predicted	
		Real	False
Actual	Real	8	5
	False	0	27

TABLE 6. 2-class confusion matrices using wavelet MACH filter for the panoramic action dataset. Mean accuracy is 90.35%.

(a) Bending

		Predicted	
		Real	False
Actual	Real	8	0
	False	0	32

(b) Jumping-jack

		Predicted	
		Real	False
Actual	Real	8	0
	False	0	32

(c) Waving-one-hand

		Predicted	
		Real	False
Actual	Real	8	8
	False	0	24

(d) Waving-two-hands

		Predicted	
		Real	False
Actual	Real	8	0
	False	0	32

(e) Jumping-at-same-point

		Predicted	
		Real	False
Actual	Real	8	1
	False	0	31

TABLE 7. Geometric mean on Weizmann action dataset

MACH filter	Wavelet MACH filter	% Improvement
88.40	96.03	7.63

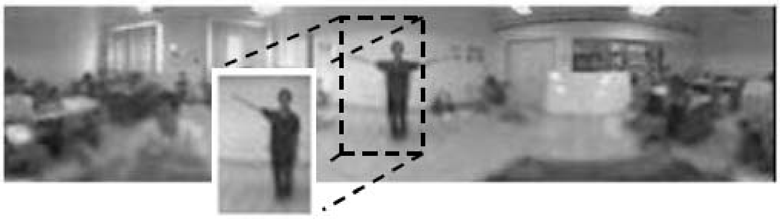


Figure 17. Similarity of waving-one-hand and waving-two-hands

Table 8. Geometric mean on panoramic action dataset

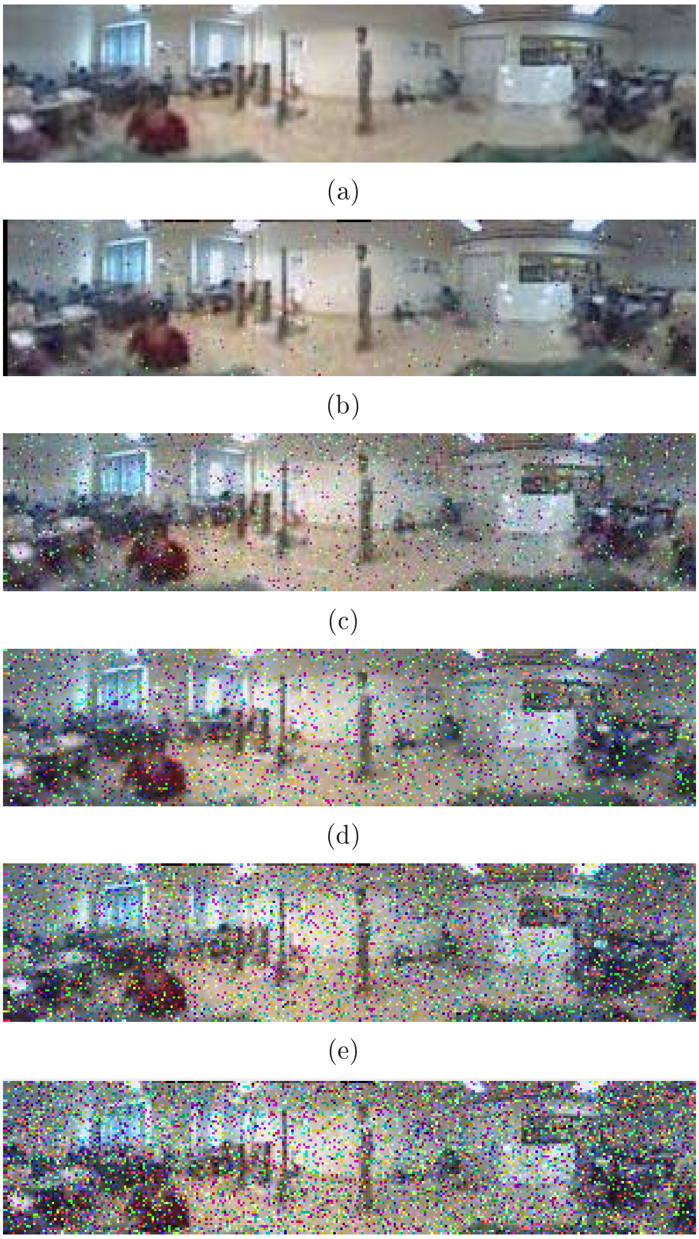
MACH filter	Wavelet MACH filter	% Improvement
80.98	90.35	9.37

Overall, the proposed filter is successful in detecting the desired action and further improves the performance of the conventional MACH filter. The result also proves that the proposed method works just as well for log-polar transformed panoramic videos.

3.4. **Noise robustness on panoramic action dataset.** In image processing, natural defects such as noise of image are somehow unavoidable. Therefore, a high degree of robustness to noise is desired. In this experiment, the recognition result with different level of noise rate has been evaluated. Figures 18(b), 18(c), 18(d), 18(e) and 18(f) show the panoramic action dataset added with salt and pepper noise at 1%, 5%, 10%, 15% and 20% repectively.

The same five actions in Figure 14 are used to test the noise robustness. A system is said to have failed the test if any of the PCHD in the test actions is less than zero. The MACH filter failed the test when 1% of salt & pepper noise is added to the test video while, the wavelet MACH filter failed the test only when 20% of salt & pepper noise is added to the test video. As a conclusion, conventional MACH filter is very sensitive to noise, therefore, by adopting the wavelet filter, the detection system becomes much more robust against noise.

4. **Conclusions.** In this paper, we have integrated the wavelet MACH filter with log-polar mapping in an attempt to recognize human actions in videos acquired by an omnidirectional camera. The wavelet MACH filter has been proposed as a means to improve the performance of the conventional MACH filter. Comparison has been made between the conventional MACH filter and the wavelet MACH filter for both normal videos and log-polar mapped panoramic videos. Simulation results on the classical Weizmann action dataset and our own collection of log-polar mapped panoramic videos appear promising. Lastly, conventional MACH filter is very sensitive to additive noise. With the additional noise, it loses the discrimination capability. The wavelet MACH filter is more robust to noise. It can resist up to 20% of salt & pepper noise. Therefore, it can be concluded that the proposed method has better recognition rate and is more robust to noise. In future, thermal imaging apparatus can be embedded onto this detection system in outdoor environment where the lightning condition is natural varying, and also in night vision. For practical applications, FPGA boards may be used in place of computers.since FPGA board is much lighter, space utilization lesser, cheaper and faster processing speed. These topics will be addressed in future work.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 18. (a) is the original image and (b), (c), (d), (e), and (f) are images which added with 1%, 5%, 10%, 15% and 20% noise respectively

REFERENCES

[1] J. K. Aggarwal, Human activity recognition – A grand challenge, *Proc. of the Digital Image Computing: Techniques and Applications*, pp.1-1, 2005.

[2] *https://computation.llnl.gov/casc/sapphire/video/video.html*, 2005.

[3] M. Seki, H. Fujiwara and K. Sumi, A robust background subtraction method for changing background, *Proc. of the 5th IEEE Workshop on Applications of Computer Vision*, pp.207-213, 2000.

[4] Z. Tang, Z. J. Miao and Y. L. Wan, Background subtraction using running gaussian average and frame difference, *Lecturer Notes in Computer Science*, vol.4740, pp.411-414, 2007.

[5] A. Mittal and N. Paragios, Motion-based background subtraction using adaptive kernel density estimation, *Proc. of the 2004 Conference on Computer Vision and Pattern Recognition*, vol.2, no.2, pp.302-309, 2004.

[6] H. Wechsler, Z. Duric, F. Y. Li and V. Cherkassky, Motion estimation using statistical learning theory, *Pattern Analysis and Machine Intelligence*, vol.26, no.4, pp.466-478, 2004.

[7] W.-C. Hu and J.-F. Hsu, Foreground extraction-based video object segmentation using motion information and gradient compensation, *International Journal of Innovative Computing, Information and Control*, vol.7, no.8, pp.4849-4859, 2011.

[8] C. H. Zhan, X. H. Duan, S. Y. Xu and M. Luo, An improved moving object detection algorithm based on frame difference and edge detection, *Proc. of the 4th International Conference on Image and Graphics*, pp.519-523, 2007.

[9] M. A. R. Ahad, T. Ogata, J. K. Tan, H. S. Kim and S. Ishikawa, A complex motion recognition technique employing directional motion templates, *International Journal of Innovative Computing, Information and Control*, vol.4, no.8, pp.1943-1954, 2008.

[10] W.-C. Hu, D.-Y. Huang and W.-H. Chen, Adaptive wide field-of-view surveillance based on an IP camera on a rotational platform for automatic detection of abandoned and removed objects, *ICIC Express Letters, Part B: Applications*, vol.1, no.1, pp.45-50, 2010.

[11] E. Menegatti, T. Maeda and H. Ishiguro, Image-based memory for robot navigation using properties of omnidirectional images, *Robotics and Autonomous Systems*, vol.47, no.4, pp.251-267, 2004.

[12] H. Ishiguro, Development of low-cost compact omnidirectional vision sensors, *Panoramic Vision: Sensors, Theory, and Applications*, pp.23-28, 2001.

[13] J. S. Victor and A. Bernardino, Vision-based navigation, environmental representations and imaging geometries, *Proc. of the 10th International Symposium Robotics Research*, vol.6, pp.347-360, 2003.

[14] R. Brunelli, Template matching: Matched spatial filters and beyond, *Technical Report, DSpace@M IT, AIM-1549*, 1995.

[15] A. Mahalanobis, B. V. K. Vijaya Kumar, S. Song, S. R. F. Sims and J. F. Epperson, Unconstrained correlation filters, *Applied Optics*, vol.33, pp.3751-3759, 1994.

[16] M. Misiti, U. Misiti, G. Oppenheim and J. Poggi, *Wavelet Toolbox User's Guide*, The MathWorks, 1996.

[17] A. Sinha and K. Singh, The design of a composite wavelet matched filter for face recognition using breeder genetic algorithm, *Optics and Lasers in Engineering*, vol.43, no.12, pp.1277-1291, 2005.

[18] D. Roberge and Y. Sheng, Optical wavelet matched filter, *Applied Optics*, vol.33, no.23, pp.5287-5293, 1994.

[19] S. Goyal, N. K. Nishchal, V. K. Beri and A. K. Gupta, Wavelet-modified maximum average correlation height filter for out-of-plane rotation invariance, *International Journal for Light and Electron Optics*, vol.120, no.2, pp.62-67, 2009.

[20] M. Greiffenhagen, D. Comaniciu, H. Neimann and V. Ramesh, Design analysis, and engineering of video monitoring systems: An approach and a case study, *Proc. of IEEE*, vol.89, no.10, pp.1498-1517, 2001.

[21] W. K. Wong, C. W. Choo, C. K. Loo and J. P. Teh, FPGA implementation of log-polar mapping, *IEEE Proc. of the 15th International Conference on Mechatronics and Machine Vision in Practice*, pp.45-50, 2008.

[22] F. Berton, G. Sandini and G. Metta, Anthropomorphic visual sensors, *The Encyclopedia of Sensors*, vol.10, pp.1-16, 2005.

[23] F. Berton, A brief introduction to log-polar mapping, *Technical Report*, LIRA_Lab, University of Genova, 2006.

[24] P. Lokesh, Object tracking and velocity determination using TMS320C6416T DSK, *Internship Project Report*, 2008.

[25] O. R. Vincent and O. Folorunso, A descriptive algorithm for sobel image edge detection, *Proc. of Informing Science & IT Education Conference*, vol.70, pp.91-95, 2009.

[26] H. Lu, P. Bone, R. Young and C. Chris, A novel logarithmic mapping algorithm for the human iris recognition using a MACH filter, *Proc. of the 15th Signal Processing and Communications Applications*, pp.1-4, 2007.

[27] M. D. Rogdriguez, J. Ahmed and M. Shah, Action MACH: A spatio-temporal maximum average correlation height filter for action recognition, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2008.

[28] M. D. R. Sullivan and M. Shah, Visual surveillance in maritime port facilities, *Proc. of SPIE, the International Society for Optical Engineering*, vol.6978, pp.697811-8, 2008.

[29] A. J. H. Hii, C. E. Hann, J. G. Chase and E. E. W. Van Houten, Fast normalized cross-correlation for motion tracking using basis functions, *Computer Methods and Programs in Biomedicine*, vol.82, no.2, pp.144-156, 2006.

[30] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, Actions as space-time shapes, *Proc. of the 10th IEEE International Conference on Computer Vision*, vol.2, pp.1395-1402, 2005.

[31] *http://users.isr.ist.utl.pt/∼alex/Projects/TemplateTracking/logpolar.htm.*

[32] D. Eaton, *http://www.cs.ubc.ca/∼deaton/remarks_ncc.html*, 2006.