Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey of Action Recognition Methods for Human Activity Recognition

Dr. S. Ravimaran[1], R. Anuradha[2]
[1] Principal, M.A.M College of Engineering, Trichy, Tamilnadu, India
[2] Department of CSE, M.A.M College of Engineering, Trichy, Tamilnadu, India

Abstract— This paper aims to recognize human activity using various deep and shallow learning techniques. Activity recognition locates a moving person with respect to time. Its uses are in surveillance, animation and robotics and even associated with the video tracking. We combine image sequences to recognize the human activity. While recognizing action, due to motion over lapping, Self-Occlusion problem is caused and is very complex to solve. Many recognition methods make them complex to solve or skip them.This paper compares the performance of Binary Motion Image concept with the Motion History Image[1], Directional Motion History representation[1], Gait Energy Image[3] and Motion Energy Image[1],[2]to solve the self-occlusion problem by feeding them into the Convolutional Neural network[1].

Keywords: BMI, MHI, GEI, CNN

I. INTRODUCTION

An emerging active area of research in computer vision with wide scale of applications in video surveillance, virtual reality, computer human interfaces (robotic interaction with humans), sports video analysis etc. is Human activity recognition.This method first takes three-dimensional model of a person and then recognizes the motion using representation and recognition theory that decomposed motion-based recognition describing where there is motion (the spatial pattern)and then describing how the motion is moving.

Self – Occlusion due to motion overlapping makes the task daunting for motion recognition methodologies address to recognize and understand varieties of human activities. These methods either bypass this problem or solve this problem in complex manner. There are various approaches for activity recognition such as (i) Spatio-temporal (ii) Frequency based (iii) local Descriptors (iv) Shape Based and (v) Appearance based. In this paper, we concentrate on motion self-occlusion problem due to motion overlapping in various complex activities for recognition.

Our paper uses various approach to recognize the human activity, where one of them combines the image sequences into a single image called Binary Motion Image (BMI).Another is MHI [1] and DMHI [1]. One can express the motion flow or sequence by using Appearance-based template matching paradigms. It uses the intensity of every pixel in temporal manner.Motion Energy Image (MEI) used in this paperdescribes the motionshape and spatial distribution of motion [1], [2]. We have chosen a new spatio-temporal gait representation like Gait Energy Image (GEI) [3], for recognition movements like walking. Unlike other gait representations which consider gait as a sequence of poses, GEI represents human motion sequence in a single image while preserving some temporal information. We feed all these into Convolutional Neural Network for action recognition and a comparison of their performances is made.

The paper is organized as follows: Section II gives an overview of all related works. Section III describes Overview of methods which describes the process of feeding in MHI, MEI, GEI and BMI to CNN as input for activity recognition. Section IV presents the comparative results and analysis using recognition rates and confusion matrices for various inputs given in Section III. Then a comparative study of activity recognition methods is made in section V. Finally, Section VI concludes the paper.

II. RELATED WORKS

This paper presents methods along with the MHI method, which is very widely employed method for various applications. For example, using the MHI method, Davis et al. has developed a virtual aerobics trainer that watches and responds to a user as he/she performs a workout. An interactive art demonstration can be constructed from the motion templates. An interactive and narrative play space for children, called Kids Room was developed using the MHI method successfully. Yau et al. has developed a method for visual speech recognition employing the MHI method. The video data of the speaker's mouth is represented using grayscale images named as motion history image. Automatically localizing and tracking moving person or vehicle for an automatic visual surveillance system was demonstrated in by employing the MHI method before employing an extended mean shift approach.In recent years, various approaches have been proposed for human recognition by gait. These approaches can be divided into two categories: model-based approaches and model-free approaches.

Model-based gait recognition approaches focus on recoveringa structural model of human motion, and the gait patterns are then generated from the model parameters for recognition. Niyogi and Adelson [14] make an initial attempt in a spatiotemporal (XYT) volume. They first find the bounding contours of the walker, and then fit a simplified stick model on them. A characteristic gait pattern in XYT is generated from the modelparameters for recognition. Yoo et al.

[19] estimate hip and knee angles from the body contour by linear regression analysis. Then trigonometric-polynomial interpolant functions are fitted to the angle sequences, and the parameters so-obtained are used for recognition. In Lee and Grimson's work [11], human silhouette is divided into local regions corresponding to different human body parts, and ellipses are fitted to each region to represent the human structure. Spatial and spectral features are extractedfrom these local regions for recognition and classification. Bhanu and Han [5] propose a kinematic-based approach to recognize individuals by gait. The 3D human walking parameters are estimated by performing a least squares fit of the 3D kinematic model to the 2D silhouette extracted from a monocular image sequence. Human gait signatures are generated by selecting features from the estimated parameters. In these model-based approaches, the accuracy of human model reconstruction strongly depends on the quality of the extracted human silhouette. In the presence of noise, the estimated parameters may not be reliable. To obtain more reliable estimates, Tanawongsuwan and Bobick [17] reconstruct the human structure by tracking 3D sensors attached on fixed joint positions. However, their approach needs lots of human interaction which is not applicable in most surveillance applications.

Son et al. has calculated the MHI and then combined with background model to detect candidate road image. Gait History Image and Gait Energy Image are created for gait analysis based on the concept of the MHI has a threat assessment method for automated visual surveillance with the aid of the MHI. A PDA-based recognition system based on the MHI method is developed by Petras et al. have devised a flexible test-bed for unusual behavior detection and automatic event analysis using the MHI. A recent method by Kellokumpu et al. [20] have proposed a recognition method at the top of the MHI method using texture-based description of the movements by employing local binary pattern (LBP) operator. Later they have used HMM for recognition.

Yuxiao Hu, Liang liangCao,FengjunLv, Shuicheng Yan , Yihong Gong and Thomas S. Huang (2006), "Action Detection in Complex Scenes with Spatial and Temporal Ambiguities", ShugaoMaJianming Zhang NazliIkizler-CinbisStanSclaroff (2013), Action Recognition and Localization by Hierarchical Space-Time Segments", these works employ multi-instance learning (MIL) based Support Vector Machine (SVM) to handle these ambiguities in both spatial and temporal domains. This multi-instance method provides a way to not only recognize the action of interest, but also locate the exact position and time period of the action. The paper called the proposed algorithm as Simulated annealing Multiple Instance Learning (SMILE). The action detection in complex scenes with cluttered backgrounds or partially occluded crowds, it is very difficult to locate human body precisely. In addition, ambiguities may also exist in temporal domain.

Xinxiao Wu Dong XuLixinDuan (2011), Action Recognition using Context and Appearance Distribution Features, this paper first proposes a new spatio-temporal context distribution feature of interest points for human action recognition. Each action video is expressed as a set of relative XYT coordinates between pairwise interest points in a local region. This paper first propose a new visual feature by using multiple GMMs to characterize the spatio-temporal context distributions about the relative coordinates between pairwise interest points over multiple space-time scales. Specifically, for each local region (i.e., sub-volume) in a video, the relative coordinates between a pair of interest points in XYT space is considered as the spatio-temporal context feature. Then each action is represented by a set of context features extracted from all pairs of interest points over all the local regions in a video volume. Gaussian Mixture Model (GMM) is adopted to model the distribution of context features for each video. However, the context features from one video may not contain sufficient information to robustly estimate the parameters of GMM.

III. OVERVIEW OF METHODS

Binary Motion image is built by combining the action sequences into a single image. This is actually a 2D representation of the image retained by View based methodologies where the entire raw image as a single image in HD space. Activity is recognized with the help of this. The outstanding performance of this method compared to others is presented in the section below. This method can be easily scaled to 3D depth maps.

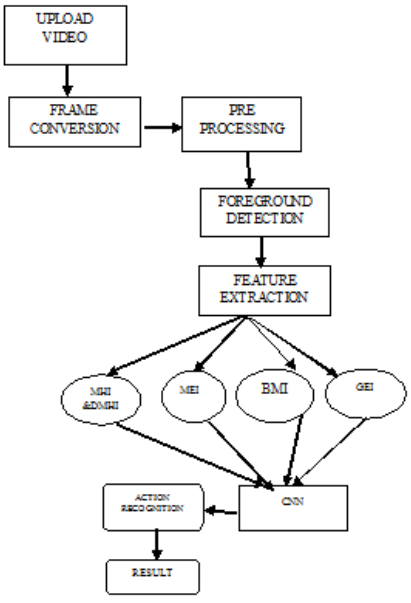

Fig.1 Overview of methods for analysis.

Our paper has various modules and they are described as follows:

- Frame conversion & Preprocessing
- Foreground detection
- Feature extraction
- Feeding MHI,DMHI,MEI and GEI as inputs
- Convolutional neural network
- Action recognition

A. Frame Conversion and Preprocessing

Frame conversion is conversion of video to image so that there is a separate value to know about the value of element and the pixel value localization results and compare the image process and implement the function. Prior to analyzing main data and extracting of information, Preprocessing is required. This includes operations on images at lowest level of abstraction. Here the input and output are intensity images. This suppresses unwanted distortions and enhances image features that can be given for further processing. The video is uploaded, image values are segmented and ultimately noise is reduced here.

B. Foreground Detection

The Foreground Detection aims in detecting changes in image sequences. This separates the changes taking place in background. This detects changes when background is set. So the background model should be developed first. In spite of shapes shadows and moving objects, it should be robust to lighting changes, repetitive movements and long term changes. System object compares a color or gray scale video frame to a background model to determine whether individual pixels are part of the background or the foreground. The color is first compared and segmentation is done after element detection. The output of this step is given as input to feature extraction.

C. Feature Extraction

Feature extraction aims for a transformation of large redundant set of input data into a reduced set of features. These features contain relevant information from input data. This makes an individual to perform the desired action on this feature that is informative and leading to better human interpretations. This is useful for image matching and retrieval. Here we give MEI, MHI, DMHI, BMI and GEI as inputs. The process here is to represent image parts, which is then matched and retrieved as shown in the figure.

D. Convolutional Neural Network

A general architecture of CNN is composed of input map such as image, a number of hidden feature maps and output processing layer. To feature map layer is obtained when convolution with a trainable kernel is done. The Gabor like filters obtain edges along different orientations. This is followed by an activation function. This constitutes the first step. The second step is sub-sampling, which involves averaging or max-pooling sub-region and obtaining a spatially down-sampled map. It consists of a single trainable weight and additive bias. This is done to reduce the size of maps and also helps in imparting a small degree of shift and distortion invariance. After a series of convolution and sub sampling layers, a convolution map is developed by randomly selecting a number of trainable weights and obtaining a single matrix. This helps in exploring different features during training. Finally, a linear transformation is applied to obtain an output layer to tell which class is classified. Similar to ANNs, CNNs are trained using Feed Forward and Back Propagation algorithms while keeping in mind the concept of shared weights.

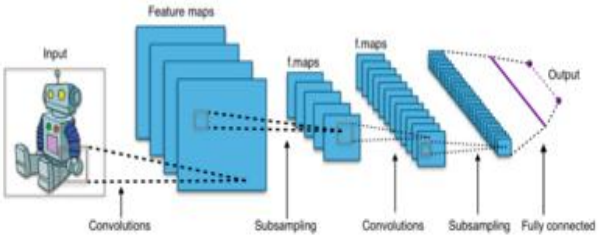

Fig.2 CNN Architechture

The above figure shows most common form of a CNN architecture stacing a few layers, following them with POOL layers, and repeating this pattern until the input has been merged spatially to a small size.Convolutional neural networks (CNNs) consist of multiple layers of small neuron collections which process portions of the input image, called receptive fields. Convolutional networks may include local or global pooling layers, which combine the outputs of neuron clusters at some point, it is common to transition to fully connected layers. The last fully connected layer holds the output.

All the above mentioned DMHI, MEI, GEI and BMI are fed into this CNN and performance of all these are calculated and compared.

E. Action Recognition:

The action recognition will be performed by comparing the outputs from CNN as mentioned below in the upcoming section.

IV. COMPARITIVE STUDY OF ACTION RECOGNITION METHODS

A. Motion History Image Template and CNN:

MHI attempts to develop a view-based approach for representation and recognition of movements designed to support the direct recognition of the motion. Focus is done on accumulating and recognizing holistic patterns of motiondespite trajectories of structural features.MHI can be considered as a component version of a temporal template, a vector-valued image where the component of each pixel is some function of the motion at that pixel location. This is a scalar-valued image where more recently moving pixels are brighter (i.e., the presence of high intensity pixels), and vice versa.Zerointensity denotes no motion at that specific point.

MHI $H_\tau(x,y,t)$ can be computed from an update function $\Psi(x,y,t)$ as described in the equation below.

$$H_\tau(x,y,t) = \begin{cases} \tau & ; \text{ if } \psi(x,y,t) = 1; \\ \max(0, H_\tau(x,y,t-1) - \delta); & \text{otherwise} \end{cases}$$

$$\Psi(x,y,t) = \begin{cases} 1 & ; \text{if } D(x,y,t) \geq \xi \\ 0 & ; \text{otherwise} \end{cases}$$

Here, (x,y) and t show the position and the time respectively; $\Psi(x,y,t)$ signals object presence (or motion) in the current video image; τ decides the temporal duration of the MHI (e.g., in terms of frames); and δ is the decay parameter. For every new video frame analyzed in the sequence, this update function is called. Possible techniques for defining this update function are background subtraction, image differencing, optical flow, etc. The MHI is generated from an image, obtained from frame subtraction, using a threshold ξ.

This MHI is fed into CNN as input. One of the advantages of the MHI representations is that it encodes a range of times from frame to frame to several seconds in a single frame, and hence, this spans the time scale of human gesture. Evenin extremely low resolution with no information or strong features about the element of the scene,observer can easily recognize action. The representations for action are view-based template descriptions of the coarse image motion. So, we extend the evaluation with other representation methods that claim to be a solution such as Hierarchical Motion History Histogram (HMHH) representation and the Multi-Level Motion History Image (MMHI) method.

MHI cannot solve motion self-occlusion or overwrite problem, as previous motion is deleted or overwritten by the later motion information. So we choose a method, wherean action is separated into its directional elementscalled Directional Motion History Image (DMHI) method and we give this as input into CNN.

B. Directional Motion History Image Representation in CNN

For complex actions or activities, the basic MHI method cannot recognize properly due to its inherent nature of motion history calculation. We developed a robust method considering four compass directions (namely, up, down, left and right directions) of an action, and thereby we can simply separate an action into its directional elements. This method is called Directional Motion History Image (DMHI) method]. This method is based on the computation of optical flow. This is fed into the CNN for activity recognition. The DMHI method outperforms the basic MHI method can solve the motion-overwriting problem due to self-occlusion, which is inherent in the MHI method in human activities.

In this approach, instead of background or frame subtraction, gradient-based optical flow vector [32], $\Psi(x,y,t)$ is calculated between consecutive two frames and split it into four channels (as depicted in Fig.4.1), based on the concept of motion descriptors, based on smoothed and aggregated optical flow measurements in four channels [27]. The non-negative channels are,

$\Psi_x(x,y,t) = \psi_x^+(x,y,t) - \psi_x^-(x,y,t)$

$\Psi_y(x,y,t) = \psi_y^+(x,y,t) - \psi_y^-(x,y,t)$

The figure shows how Optical flow is split in the figure below.

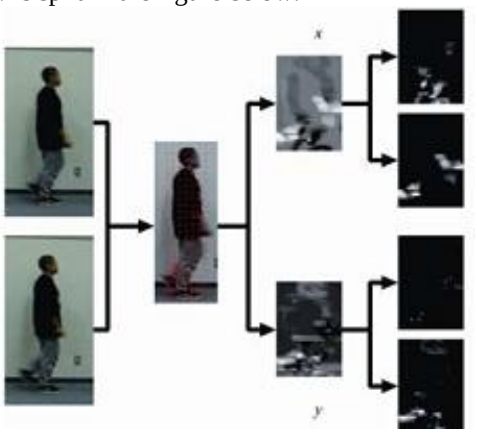

Figure.3 Optical flow is split into four different channels

.

In this way, we get four-directional motion templates for left, right, up and down directions based on a threshold ζ . We calculate the corresponding four history images as follows,

$$H_\tau^{-x}(x,y,t)=\begin{cases}\tau & ;\text{if } \psi_x^-(x,y,t)\geq\zeta;\\ \max(0, H_\tau^{-x}(x,y,t-1)-\delta) & ;\text{otherwise}\end{cases}$$

$$H_\tau^{+x}(x,y,t)=\begin{cases}\tau & ;\text{if } \psi_x^+(x,y,t)\geq\zeta;\\ \max(0, H_\tau^{+x}(x,y,t-1)-\delta) & ;\text{otherwise}\end{cases}$$

$$H_\tau^{-y}(x,y,t)=\begin{cases}\tau & ;\text{if } \psi_y^-(x,y,t)\geq\zeta;\\ \max(0, H_\tau^{-y}(x,y,t-1)-\delta) & ;\text{otherwise}\end{cases}$$

$$H_\tau^{+y}(x,y,t)=\begin{cases}\tau & ;\text{if } \psi_y^+(x,y,t)\geq\zeta;\\ \max(0, H_\tau^{+y}(x,y,t-1)-\delta) & ;\text{otherwise}\end{cases}$$

Where, $H_\tau(x,y,t)$ is MHI, $\psi(x,y,t)$ isupdate function and δ is the decay parameter.
Moreover, the corresponding energy images are calculated as,

$$E_\tau^{-x}(x,y,t)=\begin{cases}1 & ;\text{if } H_\tau^{-x}(x,y,t)\geq 0;\\ 0 & ;\text{otherwise}\end{cases}$$

$$E_\tau^{+x}(x,y,t)=\begin{cases}1 & ;\text{if } H_\tau^{+x}(x,y,t)\geq 0;\\ 0 & ;\text{otherwise}\end{cases}$$

$$E_\tau^{-y}(x,y,t)=\begin{cases}1 & ;\text{if } H_\tau^{-y}(x,y,t)\geq 0;\\ 0 & ;\text{otherwise}\end{cases}$$

$$E_\tau^{+y}(x,y,t)=\begin{cases}1 & ;\text{if } H_\tau^{+y}(x,y,t)\geq 0;\\ 0 & ;\text{otherwise}\end{cases}$$

This DMHI approach can solve the motion overwriting problem significantly, and also performs well for repetitive or complex activities. These eight motion templates are required for further motion representations. Using only history images or energy images for motion representation are not worthy. For better motion representations, it has been proved that both templates are required. Therefore, to accommodate various activities, similar to the MHI method, the motion energy templates are utilized.

C. MEI and CNN:

A binary Motion Energy Image (MEI) is initially computed to act as an index into the action library. This coarsely describes the spatial distribution of motion energy for a given view of a given action. Any stored MEIs that matches the unknown input MEI are then tested for a coarse, motion history agreement with a known motion model of the action. This MEI can be calculated using the below technique and is given to CNN for action recognition.MEI is a binary image, generated by thresholding the MHI above zero.

$$E_t(x,y,t)=\begin{cases}1 & ;\text{if } H_t(x,y,t)\geq 1;\\ 0 & ;\text{otherwise}\end{cases}$$

Figure.4 (a) The MHI (left image); (b) The MEI (right) for an action.

The above figure shows the MHI and the MEI for a hand-waving action.

1) Computingmotion Energyestimation Using SAD: The Motion Energy estimation is based on the calculation ofthe **S**um of **A**bsolute **D**ifferences (SAD)" according to thefollowing equation:

$\sum_i \sum_j \left| I_k(i,j) - I_{k-1}(i,j) \right|$

Where:$I_k(i,j)$ – Current K^{th} frame;$I_{k-1}(i,j)$ – Current $(K-1)^{th}$ frame;

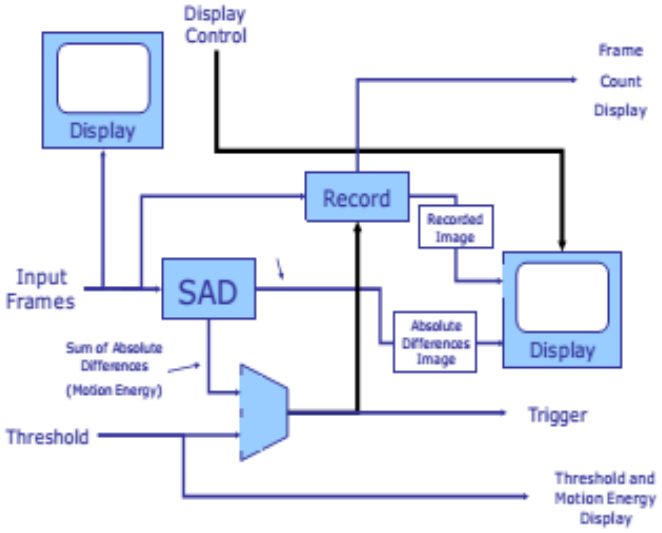

Fig.5 Computation of MEI using SAD

2) Sum of Absolute Differences (SAD): This is a widely used assimple algorithm for measuring the similarity between imageblocks. It works by taking the absolute difference betweeneach pixel in the original block and the corresponding pixel inthe block being used for comparison. These differences aresummed to create a simple metric of block similarity. Here we use this sum of absolute differences for person recognition.

To detect changes between two images, the images are compared pixel by pixel. For this purpose we form a difference image. Consider a sequence of image frames f(x, y, t1), f(x,y,t2) …….. f(x,y,tn) and let f(x,y,t1) be the reference image. An accumulative difference image ADI is formed by comparing this reference image with every subsequent image in the sequence. A counter for each pixel location in the accumulative image is incremented, every time a difference occurs at that pixel location between the reference and an image in the sequence. ADI corresponds to the two types of accumulative difference images: positive and negative.

After this MEI computation, it is fed to CNN for classification and activity recognition and later compared with other approaches for performance.

D. GEI and CNN

Gait analysis is the study of human motion, using the eye and the brain of observers. Gait analysis is used to assess, plan, and treat individuals with conditions affecting their ability to walk. It is also commonly used in sports biomechanics to help athletes. The study encompasses quantification, (i.e., introduction and analysis of measurable parameters of gaits), as well as interpretation, i.e., drawing various conclusions about the animal (health, age, size, weight, speed etc.) from its gait pattern. In this paper, we have taken a new spatio-temporal gait representation, called Gait Energy Image (GEI), to characterize human walking properties for individual recognition bygait. To address the problem of the lack of training templates, we generate a series of new GEI templates by analyzing the human silhouette distortion under various conditions like Principal component analysis followed by Multiple Discriminant Analysis for template generation.

Given a size-normalized and horizontal-aligned human walking binary silhouette sequence B(x,y,t) the grey level sequence G(x,y) is

$$G(x,y)=\frac{1}{N}\sum_{t=1}^{N} B(x,y,t)$$

Where N is the number of frames in complete cycles of the sequence, t is the frame number of the sequence, x and yare values in the 2D image coordinate.

We also propose a statistical approach to learn and recognize individual gait properties from the limited training GEI templates

GEI has several advantages over the representation of binary silhouette sequence. As an average template, GEI is not sensitive to incidental silhouette errors in individual frames. The robustness could be further improved if we discard those pixels with the energy values lower than a threshold. Moreover, with such a 2D template, we do not need to divide the silhouette sequence into cycles and perform time normalization with respect to the cycle length. Therefore, the errors occuring in these procedures can be therefore avoid.

1) Relationship of GEI With MEI and MHI: Both MEI and MHI are vectorimages,where the vector value at each pixel is a function of themotion properties at this location in an image sequence.MEI is a binary image which represents where motionhas occurred in an image sequence.

$E_\tau(x,y,t) = U_{i=0}^{\tau-1} D(x,y,t-i)$,

Where D(x,y,t) is a binary sequence indicating regions of motion, τ is the duration of time, t is the moment of time,x and y are values of 2D image coordinate. To represent regular human walking sequence, if D(x,y,t) is normalized and aligned as B(x,y,t), MEI $E_N(x,y,N)$ is the binary version of GEI(x,y).MHI is a grey level image that shows how the motion in the image is moving as below:

$$H_\tau(x,y,t) = \begin{cases} \tau & ; \text{ if } D(x,y,t) = 1; \\ \max(0, H_\tau(x,y,t-1) - \delta) & ; \text{ otherwise} \end{cases}$$

After the GEI templates are generated it is then fed into CNN for activity recognition.

E. View Based Algorithm with BMI and CNN

View based recognitions use visual templates for recognition and do not extract complex features from the image. Instead they retain the entire raw image as a single feature in high dimensional space. These templates are learnt under different poses and illumination conditions for recognition. With this in mind we build an idea of 2-D representation of action sequence by combining the image sequences into a single image called Binary Motion Image (BMI) to perform activity recognition. We test our method on Weizmann dataset focusing on actions that look similar like run,walk, skipetc. We also extend our method to 3-D depth maps using MSR Action 3D Dataset by extracting the BMI projections namely front, side and top views.

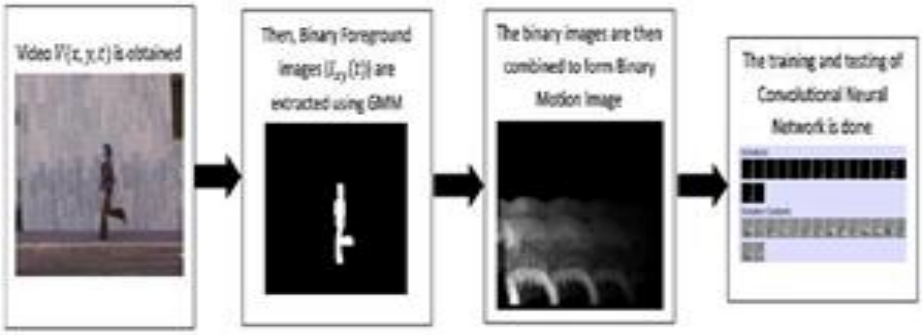

Fig.6 Overview of algorithm

This BMI representation has a number of characteristics like being local, the features have robustness to viewpoint changes and occlusions; being relatively sparse, they can be stored and manipulated efficiently. Further, by including both dynamic and static components (e.g., optical flow and gradient histograms), they can capture not only what kind of motion occurs, but also what kind of context and actors are present, without requiring reliable tracks on a particular subject. Various developments building on this general framework have yielded impressive results for realistic activities in Hollywood movies or YouTube videos.

BMI combines the image sequence using the following equation:

$BMI(x,y) = \sum_{t=1}^{n} f(t)\, I_{xy}(t)$

Where BMI(x,y)is the BMI, $I_{xy}(t)$ is the binary image sequence containing the ROI and f(t) is the weight function which gives higher preference to more recent frames and n is the total number of frames. Here, the quadratic function t^2 is used as a weight function for best looking results. Lastly, a bounding box around the image is used to extract only the region of interest in the image and to discard the black background. The BMI is post processed by applying a normalization operation. The weight function provides a means of depicting the flow of motion in an action or its optic flow. In this way, both the spatial and temporal dimensions of the activity performed are modelled using BMI.

F. 2-D Weizmann Dataset

The Weizmann database is selected for testing purpose. It contains activities performed by 9 individuals from which we selected 5 actions namely Jump,Run, Side and Walk. These are selected so as to judge our method on similar looking actions. Forthis Database BMI is calculated as described aboveand this will serve as input to CNN classifier. Matlab is used for extracting BMI. The below figure showsthe side and skip action and corresponding BMI.

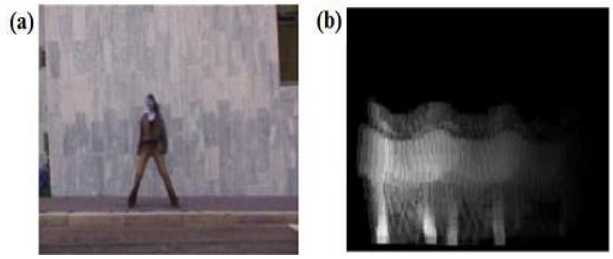

Fig.7. (a) Side action (b) Corresponding BMI

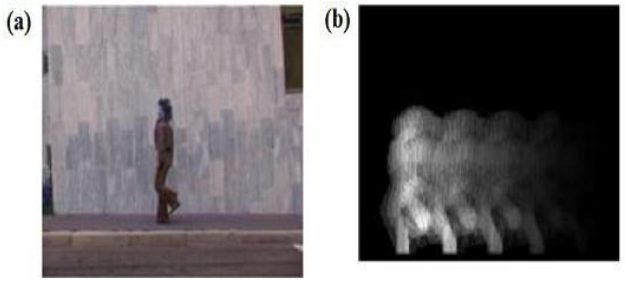

Fig.8(a) Skip Action (b) Corresponding BMI

G. MSR Action 3D Dataset

For Human Activity Recognition from 3-D data, this database is used. Consisting of 10 people performing 20 actions with each action performed by an individual, 3 BMI s ate obtained. The first image is the depth map of the forward kick and the second is from which three BMI s are calculated for front, side and Top view respectively.

Fig.9 Forward Kick Depth Map

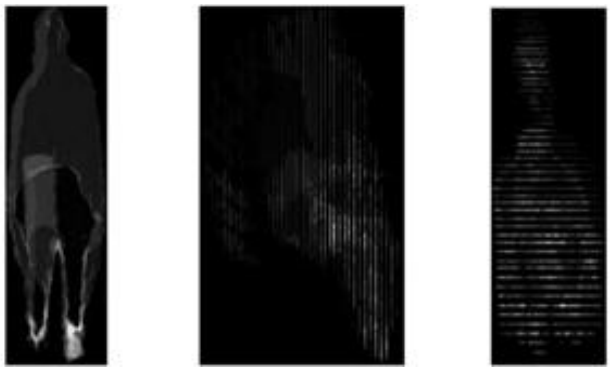

Fig.10 From left Front View BMI, Side View BMI, Top View BMI for fig.9

H. Observations From Above Methods:

Motion History Image (MHI) method is found to be robust in recognition and simple in computation. But, this method cannot solve the overwriting or motion self-occlusion problem. This also contains part of the moving persons that has background aberration due to motion and noise which lead to incorrect results, such as over-segment, motion ambiguity, and distortion. Moreover, motion at low speed cannot be easily detected. MHI cannot solve motion self-occlusion or overwrite problem, since previous motion is deleted or overwritten by the later motion information, if the motion has opposite directions in action. This direct GEI matching approach is sensitive to distortion in silhouettes generated from image sequences that are recorded under different conditions. However, with one GEI template per individual, learning cannot be performed. Even with several templates per individual, if they are from similar conditions, the learned features may be overfit to the training templates. Though the performance of the DMHI representations and the employed feature vectors for recognition are satisfactory, we find that this method cannot perform well or faces difficulties in some cases. When two or more than two persons are in-view, this method (along with other presented methods in this paper) cannot recognize properly, especially when all of them are moving in different directions. Also, like the MHI, it cannot recognize properly if the person is walking towards the camera's optical axis or if it moves something like diagonal directions.

I. Overcoming Disadvantages Using BMI

In above system the system uses Binary Motion Image (BMI) and Convolutional Neural Networks to perform human activity recognition. Image re sampling, noise reduction and color changes apply the pixel value noise. Gray image value is to change the original image and image filtering and improve the visualization and brightness of the particular place and also to improve the visual appearance and manual datasets. After this the median filtering to find the value average for the image to analysis.

V. COMPARITIVE STUDY OF ACTION RECOGNITION METHODS

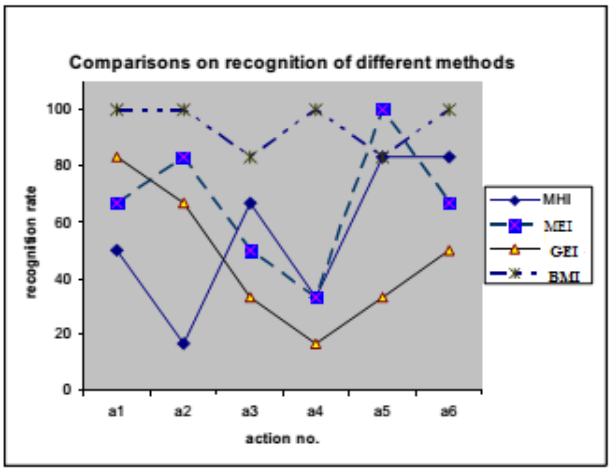

Fig.11 Comparison of methods

Figure above depicts these results in a graph that demonstrates that the BMI method visibly outshines other methods in activity recognition.

Table I Dataset Comparison

Dataset	Testing Condition	Rec. Rate
MSR ACTION 3D	30 fps	93.8
	15 fps	78.8
2-D WEIZMANN	30 fps	94.4
	15 fps	86.2
PA_86	30 fps	95.9
	Training 30 fps; Testing: 20% less data for a1~a4	83.4
PA_810 (mixed data)	Training 30 fps; Testing: 20% less data for p1~p4	82.5
	Training and Testing: 20% less data for p1~p4	82.5
MHI & DMHI	30 fps	55.6
	15 fps	38.9

In order to demonstrate the robustness of BMI against different speeds of actions, we considered various combinations of ten aerobics from eight subjects. Table I shows the recognition results for various combinations of datasets like MSR ACTION 3D and 2D WEIZMANN and PA. Here, 'P' denotes 'Person' and 'A' denotes 'Activity'; PA_810 means this dataset compriseswith10activities from 8 different individuals. Initially, we considered two different speeds for the video sequences 30 frames per second (fps) and 15 fps. So, by considering 15 fps, i.e., taking half of the frames compared to 30 fps for a period, we have lost some important correlations among the pixels. This is a reason to have a bit lower recognition.

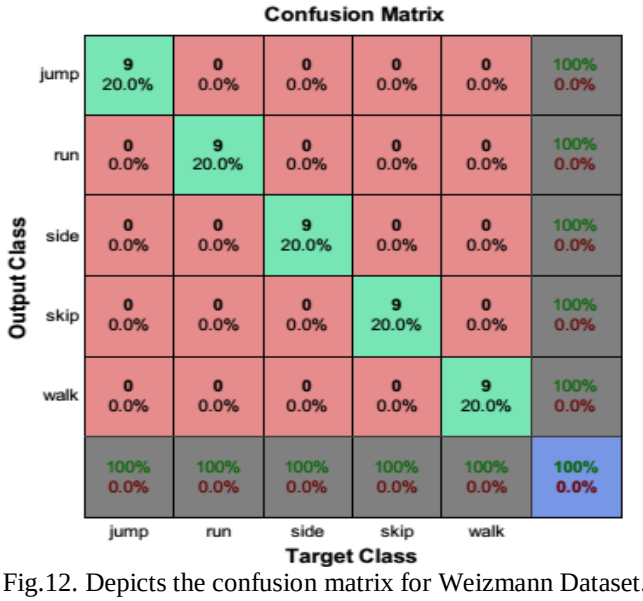

Fig.12. Depicts the confusion matrix for Weizmann Dataset.

This depicts the outstanding performance of BMI.

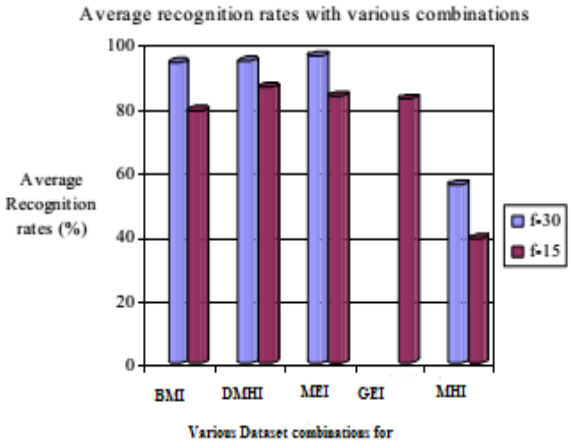

Fig.13 Graphical representation explicitly demonstrates that therecognition rates are agreeable even for 15 fps (F-15) and 30 fps (f-30) for the BMI. However, the recognition rates for the MHI method are lower.

VI. CONCLUSION

This paper presents the comparison of performances when DMHI, MEI, GEI and BMI are fed into CNN. This has also shown the robustness of DMHI representation. Though the performance of the DMHI representations and the employed feature vectors for recognition are satisfactory, we find that this method cannot perform well or faces difficulties in some cases. When two or more than two persons are in-view, this method cannot recognize properly, especially when all of them are moving in different directions. Also, like the MHI, it cannot recognize properly if the person is walking towards the camera's optical axis or if it moves something like diagonal directions. With one GEI template per individual, learning cannot be performed. Even with several templates per individual, if they are from similar conditions, the learned features may be overfit to the training templates.

In this paper, we have chosen a new view-based algorithm for recognizing human activities. Our method stacks all video frames into a single image to form the BMI which demonstrates the flow of motion of the action and is invariant to holes, shadows and partial occlusions. This method is then extended for activity detections using 3-D depth maps. The performance shown by our algorithm on both 2-D and 3-D datasets support our hypothesis. Our method includes a slight level of invariance to translation, rotation and scale changes mainly due to sub-sampling layer in CNN. Due to the use of binary foreground masks, the method is independent of dress style worn by individuals. Also the method is invariant to speed of the action performed. SO we concludethat BMIisvery efficient when compared to GEI, DMHI and MEI.

REFERENCES

[1] Analysis of Motion Self-Occlusion Problem Due to Motion Overwriting for Human Activity Recognition Md. Atiqur Rahman Ahad, JooKooi Tan, HyoungSeop Kim and Seiji Ishikawa Faculty of Engineering, Kyushu Institute of Technology, Fukuoka, Japan, JOURNAL OF MULTIMEDIA, VOL. 5, NO. 1, FEBRUARY 2010

[2] Application of SAD Algorithm in Image Processing for Motion Detection and SIMULINK Blocksets for Object Tracking, Menakshi Bhat1, Pragati Kapoor2, B.L.Raina3 1Assistant Professor, School of Electronics & Communication Engg.

[3] Individual Recognition Using Gait Energy Image , Ju Han and BirBhanu Center for Research in Intelligent Systems University of California, Riverside, California 92521, USA

[4] Human Activity Recognition using Binary Motion Image and Deep Learning, Tushar Dobhal, Vivswan Shitole, Gabriel Thomas, Girisha Navada.

[5] Arandjelovic R. and Zisserman A. (2012), "Three things everyoneshouldknow to improve object retrieval", in IEEE Conference, pp. 2911–2918.

[6] Brendel W. and Todorovic S. (2011), "Learning spatiotemporal graphs ofhuman activities," in IEEE International. Conference, pp. 778–785.

[7] Brox T. and Malik J. (2011), "Large displacement optical flow: Descriptormatching in variational motion estimation," IEEE vol. 33, pp. 500–513.

[8] Efros A.A., Berg A.C, Mori G., and Malik J. (2003), "Recognizingactionata distance,"inIEEEconference vol. 2, pp. 726–733.

[9] Hu Y., Cao L., Yan S., Gong Y., and Huang T. S. (2009), "Action detection in complex scenes with spatial and temporal ambiguities," IEEE pp.128–135.

[10] Jhuang H., Gall J., Zuffi S., Schmid C., and M.J. Black (2013), "Towardsunderstanding action recognition," IEEE International Conference, pp. 3192

[11] A. Bobick and J. Davis, "The recognition of human movementusing temporal templates", IEEE Trans. On PAMI, vol.23, no.3, pp. 257-267, March 2001.

[12] Md. Atiqur Rahman Ahad, J.K. Tan, H.S. Kim, and S. Ishikawa,"Human activity recognition: various paradigms",International Conference onControl, Automation and Systems, pp. 1896-1901, Oct. 2008.

[13] D.Gavrilla, "The visual analysis of human movement:asurvey", Computer Vision and Image Understanding, vol.73, pp. 82-98, 1999.

[14] M. Pantic, A. Pentland, A. Nijholt, and T.S. Hunag, "Humancomputing and machine understanding of humanbehavior:asurvey", Int. Conf. on Multimodal Interfaces, pp. 239-248, 2006.

[15] R. Poppe, "Vision-based human motion analysis: an overview", Computer Vision and Image understanding, vol.108, no.1-2, pp. 4-18, Oct. 2007.

[16] Md. Atiqur Rahman Ahad, T. Ogata, J.K. Tan, H.S. Kim ,andS.Ishikawa, "Motion recognition approach tosolveoverwritingcomplex actions", 8th Int. Conference on Automatic Face and Gesture Recognition, Amsterdam, 6 pages, Sept. 2008.

[17] J. Liu and N. Zhang, "Gait history image: a novel temporaltemplate for gait recognition", IEEE Int. Conf .on Multimediaand Expo, pp. 663-666, 2007.

[18] H. Meng, N. Pears, and C. Bailey, "A Human ActionRecognition System for Embedded Computer VisionApplication", 3nd Workshop on Embedded ComputerVision(with CVPR), pp. 1-6, June 2007.

[19] R. Poppe, "Vision-based human motion analysis: an overview", Computer Vision and Image Understanding, vol. 108, no. 1-2, pp. 4-18, Oct. 2007.

[20] J. Liu and N. Zhang, "Gait history image: a novel temporal template for gait recognition", IEEE Int. Conf. on Multimedia and Expo, pp. 663-666, 2007.

[21] J. Han and B. Bhanu, "Individual recognition using gait energy image", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 2, pp. 316-322, 2006.

[22] T. Jan, "Neural network based threat assessment for automated visual surveillance", IEEE Int. Joint Conf. on Neural Networks, vol. 2, pp. 1309-1312, July 2004.

[23] K. Leman, G. Ankit, and T. Tan, "PDA-based human motion recognition system", Int. J. Software Engineering and Knowledge, vol. 2, issue 15, pp. 199-205, Apr. 2005.

[24] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Texture based description of movements for activity analysis" Third Int. Conf. on Computer Vision Theory and Applications (VISAPP 2008), Madeira, Portugal, vol. 1, pp. 206-213.

[25] H. Meng, N. Pears, and C. Bailey, "Motion Information Combination for Fast Human Action Recognition", 2nd International Conference on Computer Vision Theory and Applications, Spain, Mar. 2007.

[26] H. Meng, N. Pears, and C. Bailey, "Recognizing Human Actions Based on Motion Information and SVM", 2nd IEE International Conference on Intelligent Environments, pp. 239-245, 2006.

[27] Zhou, Q.; Aggarwal, J. K.; "Tracking and classifying moving jects from video", Proc of 2nd IEEE Intl Workshop on Performance Evaluation of Tracking and Surveillance (PETS'2001), Kauai, Hawaii, USA (December 2001).

