Deep Neural Networks are Really Undefeatable for Human Conflicting and Non-conflicting Event Detection

Devang S Pandya  
Information & Communication Technology  
Adani Institute of Infrastructure Engineering  
Ahmedabad, India  
Devang.Pandya@aiim.ac.in

Bhoomika Rathod  
Information Technology  
Silver Oak college of Engineering & Technology  
Ahmedabad, India  
brathod446@gmail.com

Abstract— Human activity analysis has been really vital nowadays for the notification of any threats to the system as well as human. This paper aims to compare conventional AI approaches with recent deep learning approach for the conflicting and Non-conflicting human activity detection. In this paper, we propose the system for the detection of Conflicting human activities namely; Kicking, Punching and Pushing. These systems are developed with AI conventional approaches and Deep Neural Networks. Classifying various activities only into two subgroups of activities based on their intrinsic aggression is really a challenging task. The proposed system effectively uses the well-known features namely contour and optical flow. System is tested and validated with the well-known dataset using traditional AI approaches specifically Hidden Markov Model and Support vector machine. We obtain very good accuracy with SVM with polynomial kernel. However, Convolutional Neural network is found outperforming compared to traditional AI approaches and achieves the peak accuracy.

Keywords— Hidden Markov model, SVM, Convolutional Neural network, Optical flow, Contour

I. INTRODUCTION

For Public Security, intelligent video surveillance is being vital and need of the time to effectively monitor the various activity to prevent unlawful activity. Understanding captured human activity through mounted cameras requires accurate and strong features. Conflicting event detection is becoming more important for the purpose of notification to safeguard the human and also other resources/property. Intelligent visual surveillance has dragged the attention due to increased global security concerns & an ever increasing need for effective monitoring at places such as airports, railway stations, shopping malls, and military areas [1]. Event detection involves the analysis and the recognition of motion patterns to produce a high-level description of actions and interactions among objects (Humans). Motion capture systems are costly and only effective in limited areas [2]. Blob feature has been utilized by the Yang and Cheng to recognize human activity [3] Trajectory based human action recognition has been proposed in [4].

In [4] authors have extracted trajectories by tracking the detected spatio-temporal points. Kolekar and Dash successfully employed Hidden Markov model to recognize human activity [5]. Authors have used shape and optical flow features for the identification of human activity. An optical flow based approach for recognizing human actions and human-human interactions in video sequences has been addressed by Kumar and John[6]. In [6] Authors proposed a local descriptor built by optical flow vectors along the edges of the action performer. In [7] Ma and Liu successfully employed spatio-temporal motion - based model method which denotes the entire body using the distance between contour points and the centroid, however in [7] when several people enter the background at the same time, classification of motion should be a problem. Babiker and Khalifa combine the features like area, height, weight & centroid of bounding box of blob to classify human activity [8]. Optical flow has been utilized for single human activity recognition using SVM by Danafar and Niloofar in [9].

The aim of this paper is to separate Conflicting activity which can cause the harm or injury. In our proposed approach, we aim to classify the aggressive activities like kicking, punching, pushing, as Conflicting and pointing, handshaking, hugging as Non-conflicting behaviors. Paper does not target the individual classification of activity rather it targets the above group of activities which is either Conflicting or Non-conflicting.

The organization of the paper is as follows: Section 2 describes proposed system overview using traditional AI approaches. Section 3 details information about standard dataset. In section 4, proposed approaches for Conflicting & Non-conflicting human behavior detection system are discussed. The experimental results are elaborated in section 5. Later Convolution neural architecture is elaborated with its results in section 6 and paper is concluded in section 7.

II. SYSTEM OVERVIEW

We propose system which successfully detects Conflicting events viz. kicking, punching and pushing. Paper is technically split into two parts. First part (Section IV and V) explains three approaches specifically using contour, optical flow and then combining both these features. Hidden Markov model and Support vector machines both have been separately employed to measure the performance of the system. Second part (Section VI) describes the CNN architecture and its results for the conflicting and Non-conflicting event detection system.

Fig. 1 shows the Flowchart of overall proposed system. The proposed system extracts optical-flow and contours from each frame of the input video. To detect foreground, previous frame is compared with next frame and take absolute difference between gray frames respectively. Later, contour and optical flow are extracted from the frames. In the later stage, we classify the events using both the classifier SVM and
HMM individually and their results are compared. We also combine both the features and performed classification to analyze their combined effect.

III. VIDEO DATASET

We conducted experiments on two datasets namely SDHA and our own developed (HAD-OD). These datasets are detailed in Table I. Dataset 1 consists of 60 videos and Dataset 2 consists of 10 videos.

**TABLE I. DATASET INFORMATION**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1-SDHA (HIGH LEVEL HUMAN INTERACTION)</td>
<td>60 videos, 4 seconds fps: 25 Resolution: 720*480 Static camera Color Videos</td>
</tr>
<tr>
<td>Dataset 2-HAD-OD (Human Activity Dataset-Own Developed)</td>
<td>10 videos fps: 25 Resolution: 720*576 Static camera Color Videos</td>
</tr>
</tbody>
</table>

IV. CONFLICTING AND NON-CONFLICTING EVENT DETECTION SYSTEMS

A. Scenario 1: Conflicting & Non conflicting human behavior detection using Optical-flow

Optical flow can be defined as observed motion of image objects in the consecutive frames induced by the camera or object movement. Differential methods for the estimation of optical flow are based on the assumption that intensity values of image objects do not change in the successive frame which can be described as follows

\[
f(x + v_x, y + v_y, t + 1) = f(x, y, t)
\]  

Where \( (v_x, v_y)^T(x, y, t) \) is called optical flow. There are well-known optical methods to compute optical flow like Lukas-kanade [9], Horn-Schunck [10] and Farenback [11]. Lukas and Kanade is differential optical method which assumes unknown optical flow is constant within some neighborhood window. This method is considered faster and found sensitive to the noise. Horn-Schunck optical flow produces dense optical flow and it is based on energy minimization. Optical flow method proposed by Farenback [11] approximates the neighborhood of frames by quadratic polynomials. Signal model represented in local coordinate system is described as follows

\[
f(x) \sim x^T A x + b^T x + c
\]

Where vector \( x \) contains variables \( x \) and \( y \), \( A \) is a symmetric matrix of unknowns which captures the information about the even part of the signal while \( b \) retains the odd part of the signal and \( c \) is an unknown scalar. We extract optical flow as a feature from every consecutive frame of the video. We compute the magnitude \( M_i \) of the optical flow for each frame as follows because optical flow magnitude is reported accurate and efficient in event detection [12]. Where \( f_i \) denotes the \( i \)th frame and \( v_x \) and \( v_y \) denote the optical flow at \((x, y)\) coordinates of the frame. Finally resultant sum of squared computes the magnitude.

\[
M_i = \sum_{(x,y) \in f_i} \sqrt{v_x^2(x,y)} + v_y^2(x,y)
\]

Since our classification problem targets the detection of Conflicting activity which involves aggressions and some rapid movements, it is natural to believe that magnitude of such activity exhibits this reflection. Fig. 2 depicts the behavior of the flow magnitude for the Non-conflicting and Conflicting activity. Fig. 2 left plot shows two peak due to the pointing activity happens in two phases. In the first phase person lifts up his hand, holds the hand pointing towards and finally moves the hand in actual position. In Conflicting event of kicking, we can observe continuous vibrations which causes fluctuations.
Fig. 2. Optical flow Magnitude of Non-conflicting and Conflicting activity

(a) Non-conflicting activity
(b) Conflicting activity

Fig. 3 represents the horizontal & vertical component of optical-flow for Conflicting/Non-conflicting human behavior describing handshaking and kicking events. Fig. 3 shows the clarity of dense optical flow obtained by Farenback algorithm.

(a) Horizontal component of Dataset 1
(b) Horizontal component of Dataset 2
(c) Vertical component of Dataset 1
(d) Vertical component of Dataset 2

Farenback algorithm requires the proper values of its parameters. We used scale of 0.5 to downsample the image by the factor of 2. We continue downsampling up to level 3 and use 3 number of search iterations per each pyramid level. Once the flow is computed, averaging is done over the window size of 15.

B. Scenario 2: Conflicting & Non conflicting human behavior detection using Contour

To determine the motion of a moving body, we use contours, which represent change in body’s boundary shape over time, and also represent the human in frames. Fundamentally, Contour is a curve joining all the continuous points, having same color or intensity. Because of issues like the changing light intensity, Noise in images, there is a problem in obtaining contour of foreground accurately. We obtained the multiple contours per one object (human). We consider each retrieved contour due to movement of actors and other objects in the frame. Next, we find the center points of all contours by calculating the moments of the obtained contours. After obtaining moments, we find the average of all zero order moments. The purpose is to find the distance between two humans as a feature of one frame using the center points. To carry out this task, we apply k-means clustering algorithm, which computes two fixed centroid points, the distance between these two centroid points is referred to as a distance between two humans in one frame. We calculate the cluster center as follows: Let \( X = \{x_1, x_2, x_3, \ldots, x_n\} \) be the set of data points and \( V = \{v_1, v_2, \ldots, v_c\} \) be the set of centers.

\[
VI_i = \left( \frac{1}{ci} \right) \sum_{j=1}^{ci} Xi
\]

where, ‘ci’ represents the number of data points in \( i^{th} \) cluster. Fig. 4 exhibits the above described steps. Fig. 4 subsequently shows background subtracted image, human representation using contour points and finally clusters.

V. EXPERIMENT RESULTS

The Hidden Markov model (HMM) is one such statistical model, which interprets the (non-observable) process by analyzing the pattern of a sequence of observed symbols. There are three basic problems viz. Evaluation, Decoding and Learning can be solved using HMM [13], [14].

In proposed system, we use Gaussian HMM, which is Hidden Markov model with Gaussian emissions, because our observation sequences are continuous themselves (Gaussian distributed). In Training Phase, we feed training images to the Gaussian HMM and parameters are estimated. During Testing Phase, such label is assigned to the observation sequence (\( O \)) which largely matches with the given trained model (\( \lambda_i \)) of each class as described mathematically in equation 5.
event\_label = \arg \max_{j=1,k} P(O|\lambda_j) \quad (5)

Number of states considered in HMM also largely affect the accuracy of the system. Accuracy is calculated using evaluation parameters Precision and Recall. We have also utilized SVM for the classification of Conflicting and Non-conflicting event. Accuracy analysis is depicted in Table II and III respectively. We have tested on both the datasets. Dataset 2 is recorded video through the camera stationed in front of the activity area creating almost zero angle while dataset 1(SDHA) is through mounted camera at some height. We maintain training-testing ratio of 80%-20%. We performed training only on dataset 1, while testing is carried out on both the datasets. Table II reflects that using Hidden Markov model classification we achieve good precision of nearly 83% with optical flow and 55%, 65% with contour. We need to tune the number of states of HMM. With the combination of both the features, we achieve precision of 88% however on our dataset 2 it reduces to 62%. Thus, contour adversely impact the performance of the combined feature and accuracy declines because of its sensitivity to noise. SVM with polynomial kernel learning is found sound compared to HMM since in our experiments, number of states are created automatically from the observed features rather than manual. Thus, HMM accuracy is even found more fluctuating.

### Table II. Accuracy of classification using HMM

<table>
<thead>
<tr>
<th>Features</th>
<th>Dataset</th>
<th>Number of HMM component</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contour</td>
<td>SDHA</td>
<td>5</td>
<td>55%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>HAD-OD</td>
<td>5</td>
<td>65%</td>
<td>65%</td>
</tr>
<tr>
<td>Optical-flow</td>
<td>SDHA</td>
<td>3</td>
<td>83%</td>
<td>83%</td>
</tr>
<tr>
<td></td>
<td>HAD-OD</td>
<td>3</td>
<td>82%</td>
<td>72%</td>
</tr>
<tr>
<td>Optical-flow &amp; Contour</td>
<td>SDHA</td>
<td>5</td>
<td>88%</td>
<td>83%</td>
</tr>
<tr>
<td></td>
<td>HAD-OD</td>
<td>5</td>
<td>62%</td>
<td>61%</td>
</tr>
</tbody>
</table>

### Table III. Accuracy of classification using SVM with Polynomial kernel

<table>
<thead>
<tr>
<th>Features</th>
<th>Dataset</th>
<th>Polynomial Degree</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contour</td>
<td>SDHA</td>
<td>3</td>
<td>40%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>HAD-OD</td>
<td>3</td>
<td>57%</td>
<td>56%</td>
</tr>
<tr>
<td>Optical-flow</td>
<td>SDHA</td>
<td>3</td>
<td>83%</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>HAD-OD</td>
<td>3</td>
<td>85%</td>
<td>78%</td>
</tr>
<tr>
<td>Optical-flow &amp; Contour</td>
<td>SDHA</td>
<td>4</td>
<td>87%</td>
<td>83%</td>
</tr>
<tr>
<td></td>
<td>HAD-OD</td>
<td>4</td>
<td>78%</td>
<td>78%</td>
</tr>
</tbody>
</table>

Table III depicts the performance of SVM classifier with the individual feature as well as combination of both the features. We achieve good accuracy with the optical flow and combination of both the features with the regularization parameter C=1000. It is important to note that only SVM with polynomial kernel succeeds to achieve such good rate of accuracy which is quite comparable with HMM. With polynomial degree 4, and using both feature combined, precision of 87% and 78% is reported on SDHA and HAD-OD dataset as shown in Table III.

### VI. DEEP NETWORK FOR EVENT DETECTION SYSTEM

Deep learning is being popular now a days in solving the computer vision problem. As CNN has been proven very effective in many Classification problem, it is adopted in this work for robust behavior recognition. There are well Known CNN architecture exist in the literature [15]. It is basically a special class of multilayer perceptron. CNN architecture basically consists of Convolutional layer, Pooling layer and Fully connected layers. Inputs are fed to the input layer, where Convolutional process takes place and generally input size grows in the feature map dimension. Subsequently, the feature matrix is down-sampled, and the feature map is obtained by an excitation function. Fig. 8 depicts the structure of the CNN adopted which almost resembles VGG-16 wherein couple of Convolutional layers are followed by the Pooling layer. Frames of 60 activity video of standard SDHA dataset are supplied as inputs to the CNN. During hidden layer and output layer, we subsequently use non-linear "relu" and "softmax" activation functions of CNN network for learning purpose. Table IV depicts the detailed Network configuration for proposed work. Since CNN requires large number of samples for it’s to be trained properly which is carried out by Data augmentation. Input image is of dimension 120x160x3.

### Table IV. Network Configuration

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Patch size / Stride / No.of Kernel</th>
<th>Output Shape</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>(120,160,3)</td>
<td>(120,160,3)</td>
<td>57,600</td>
</tr>
<tr>
<td>Conv - 1</td>
<td>3*3/1/32</td>
<td>(120,160,32)</td>
<td>998,400</td>
</tr>
<tr>
<td>Conv - 2</td>
<td>3*3/1/64</td>
<td>(120,160,64)</td>
<td>1,228,800</td>
</tr>
<tr>
<td>Max Pool - 1</td>
<td>3*3/1/64</td>
<td>(60,80,64)</td>
<td>307,200</td>
</tr>
<tr>
<td>Conv - 3</td>
<td>3*3/1/128</td>
<td>(60,80,128)</td>
<td>614,400</td>
</tr>
<tr>
<td>Conv - 4</td>
<td>3*3/1/128</td>
<td>(60,80,128)</td>
<td>614,400</td>
</tr>
<tr>
<td>Max Pool - 2</td>
<td>3*3/1/128</td>
<td>(30,40,128)</td>
<td>153,600</td>
</tr>
<tr>
<td>Fully Connected</td>
<td></td>
<td>(128,1)</td>
<td>128</td>
</tr>
</tbody>
</table>

Fig. 5. Confusion matrix - SDHA(Left) and HAD-OD(Right) datasets(5 epochs)

Fig. 6. Accuracy graph on SDHA and HAD-OD (5 epochs)

Fig. 7. Training & Validation loss for 5 epoch
Events are classified into either of two classes namely Conflicting and Non-conflicting. Confusion matrices and graphs showing in Fig. 5,6 reports the accuracy of 97%, 99% and 96%, 100% with 5 epochs for SDHA and HAD-OD. Fig. 7 reflects the significant reduction in training and validation loss. With 10 epochs, validation accuracy reaches at peak of 100% as shown in Fig. 9 and 10 while testing loss becomes negligible for both the datasets which is depicted in Fig 11. Although, the lighting conditions, and environment are different for datasets SDHA and HAD-OD, CNN succeeds to achieve peak accuracy. Both the datasets are divided into training-testing ratio of 80%-20%. It is clearly observed that CNN outperforms AI conventional approaches with the margin of more than 10%.

VII. CONCLUSION

In this paper, we proposed Conflicting event detection approach using optical flow and contour using AI traditional approaches specifically HMM and SVM and also compared the performance with recent CNN architecture. Conducted experiments clearly show that dense optical flow is found very stable and sound feature compared to contour since contour is more sensitive to noise and illumination variation. Experimental results validate that joining of both the features improve the overall accuracy on the SDHA dataset. It is essential to note that SVM with polynomial kernel succeeded to generalize with the new unseen dataset and less sensitive to the problem of overfitting compared to HMM. CNN is undoubtedly found quite superior compared to SVM and HMM and succeeds to achieve the peak accuracy. It is clearly evident that Convolutional Neural networks technique completely outperforms the AI traditional approaches for conflicting event detection. In future, we would like to investigate various deep learning architecture with varied datasets.

VIII. REFERENCES


