International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

2D CNN and Gated Recurrent Network for Dynamic Hand Gesture Recognition with A Fusion of RGB-D and Optical Flow Data

Sunil A. Patel, Ramji M. Makwana

Abstract*: The dynamic hand gesture is an essential and important research topic in human-computer interaction. Recently, Deep convolutional neural network gives excellent performance in this area and gets promising results. But the Researcher had focused less attention on the feature extraction process, unification of frame, various fusion scheme and sequence-to-sequence prediction of a frame. Therefore, in this paper, we have presented an effective 2D CNN architecture with three stream networks and advances weighted feature fusion scheme with the gated recurrent network for dynamic hand gesture recognition. To obtain enough and useful information we have converted each RGB-D video to 30-frame and 45-frame for input. We have calculated an optical flow for frame-to-frame by given RGB video and extract dense motion features. After finding proper motion path, we have assigned more weight to optical flow features and fuse this information to the next stage and gets a comparable result. We have also added a newest Gated recurrent network for temporal recognition of frame and minimize training time with improved accuracy. Our proposed architecture gives 85% accuracy on the standard VIVA dataset.

***Keywords*: 2D Convolutional neural network, Gesture recognition, Optical flow, RGB-D data, Gated recurrent unit, weighted fusion**

I. INTRODUCTION

Gesture recognition is an essential research topic in computer vision with number of applications in sign language recognition, virtual reality, human-computer interaction and so on. Now, we have applied hand gesture recognition in the car to provide a touchless interface. The recent survey conducted by TNS India Pvt. Ltd for the SaveLIFE foundation in India has noticed near 20% people have died in 2017 due to the use of the mobile phone during driving [1]. Our aim is to control all the secondary device using hand gesture and keep eye on the road during steering. The research paper proposed in [2] [3] tries to improve result using a different CNN based approach for hand gesture recognition on skeleton dataset. Our focus is on rising recognition accuracy of each class. In the earlier hand recognition task, features were extracted using the hand-crafted technique for improving accuracy. But the challenge remains the same because of the diversity and flexibility of gesture, the speed of action is changing, recognition time and gesture similarity.

Recently, Deep convolutional neural network-based technique can be used for identification of gesture because of strong implicit feature extraction. To overcome the challenges in gesture recognition, to find proper motion path in a video is first. RGB and Depth information can give the only appearance and strong edge response in a video. We have calculated optical from RGB video and get the appropriate motion of hand movement in the video. This addition of optical flow boosts the performance of our model [4]. As per the architecture represented in Fig.1. Firstly, we have converted all the video to a fixed number of 30-frames or 45-frames to acquire the proper knowledge of information. Next, to that, the optical flow information is calculated from RGB video to get proper motion path present in the video [5]. Then these frames are passing one-by-one to three different streams of 2D CNN model for learning and extraction of a deep feature. These deep features are merged using various fusing scheme to enhance the performance of our model. These merged features are passing to the gated recurrent network for integrating past and present information and temporal recognition. The last stage is SoftMax classifier which generates a probability score for each class. The most important four contributions of our paper are summarized below:

- The standard way for assign input from the video is 16 frames. We have tested our model with 30-frames and 45-frames. Therefore, more meaningful information with robust features we extracted. We elect 30-frames and 45-frames as a standard and convert all video to this same number of frames.
- We have used RGB video and Depth video as our input in the proposed model. We added third input optical flow for finding proper motion information in the video. We have calculated optical flow from RGB video with a consecutive fixed number of sequential frames. Optical flow gives dense motion and eliminates irrelevant information present in the video.
- Our Proposed model become a three-stream network with RGB, Depth and Optical flow as an input. The next step is extracting features from each frame using 2D CNN approach and blended these features together with several times of convolution and pooling. We have fused these features with several existing fusion schemes and proposed weighted scheme and compare their result.
- Our fused features with joint information

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Sunil A. Patel, Gujarat Technological University, Ahmedabad, Gujarat, India, Computer Engineering Department.
Dr. Ramji M. Makwana, Managing Director, AIIVINE PXL Pvt. Ltd, Rajkot, Gujarat, India.

Retrieval Number: J91850881019/19©BEIESP
DOI: 10.35940/ijitee.J9185.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

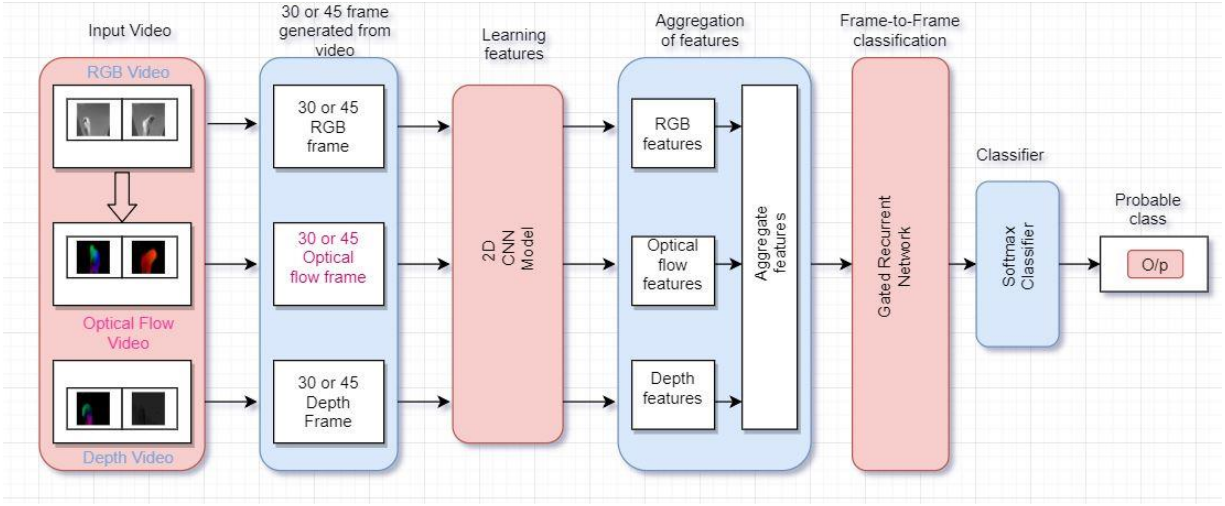

Fig. 1. This is a pipeline of architecture. In the first part input video are converted into 30-frame or 45-frame once. Then optical flow is calculated from RGB frame to find proper motion path of gesture. Then this converted frame is passing one-by-one to 2D CNN for feature extraction and learning. The processed features are merged and give input to gated recurrent network for many-to-one classification. Finally, the probability value is sent to input for SoftMax classifier for prediction of class.

of each frame are passing through a Gated recurrent network for sequence-to-sequence prediction and temporal recognition of the gesture. This is a many-to-one recurrent cellular network to merge past and present information in temporal sequence for each frame and predict probability score of each class on the last cell. This paper is formed as follows. In the section of II, we represent the related work of hand gesture recognition algorithm. In the section of III, we present our proposed work based on convolution and gated recurrent unit network. In the IVth section, we compute the result and compare with existing algorithm. In the Vth section, we can conclude with our main contribution of work and gives a way of future work.

II.RELATED WORK

In this section, we will try to represent current brief existent work on the classification of hand gesture recognition. All the conventional method typically extracts features explicitly with hand-crafted feature extraction. The early method typically uses the silhouette of an object [6], the contour of an object [7] with hand-crafted feature extraction and classified using a Hidden Markov Model. In [8] [9] often extract explicit features like a histogram of oriented gradients (HOG) and a histogram of optical flow (HOF) as a movement descriptor and support vector machine as a classifier. In the meantime, the hidden Markov model, support vector machine, K-nearest Neighbour [10] also applied for modeling a hand gesture for human. In [11] the author proposed a bag-of-words as a feature detector and use SVM as a classifier for RGB-D as an input to achieve recognition. The method which describes the support vector machine as a classifier and extracts features explicitly using a histogram of gradients (HOG) for RGB and depth data [12].

Recently, A Deep learning is used for gesture and activity recognition by using automatic implicitly learning complex features. [13] [14] [15] and [16] evaluates the performance of hand gesture using a three-dimensional convolutional neural network in video sequence along with space and time using support vector machine. CNNLSTM network is used to visualize the features using a deconvolutional neural network of the original input image [17]. The Two Stream network model is developed using 3D CNN by creating new convolutional fusion layer for video activity detection [18]. [19] and [20] propose a C3D model for extraction of space-temporal features simultaneously. In [21] have developed a deep learning based HGR system with a new strategy, by using a two-stage CNN architecture and proposed a hand segmentation architecture in the first stage of the network and a two-stream CNN for the second stage of the network which extracts useful features from raw and segmented images to obtain high classification accuracy even though they did not use depth information by taking advantage of an efficient data augmentation technique in two forms of online and offline. In [22] [23] first, use CNN for feature extraction and then use LSTM for prediction of the label.

Among discussed above method, the technique of [24] is similar to ours. But our proposed effort has four extensions. Very first is 30-frame and 45-frame unification of frames for finding proper motion fitting path. We also proved that a little bit greater number of frames compared to average frames can boost the performance. The second is find motion using optical flow to remove gesture irrelevant factor from the background. From this perception, the effort of [9] is similar to our proposed work. But they used a histogram of optical flow (HOF) to get motion information. In its place, we put more effort into finding the dense optical flow to acquire proper motion path information in RGB-D frame. For simultaneously

given RGB-D and optical flow data as an input in the network, we have made an alteration of our proposed model to three stream networks and merge the features of RGB-D and Optical flow data by using various fusion method. In the fusion point of view, our work is similar to [18] but we have added new weighted fusion-based scheme on statistical analysis and empirical analysis and assign more weight to optical flow.

Retrieval Number: J91850881019/19©BEIESP
DOI: 10.35940/ijitee.J9185.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The fourth one is we are using newest gated recurrent unit model for temporal recognition of frames.

III.PROPOSED METHOD

Compared with all the preceding works, our gesture recognition work, pointing to solve video-based dynamic recognition of hand gesture, faces many complexities in the extraction of important features. If inputs are video instead of images, then this task required more efforts because it's needed a temporal feature for learning. Appreciation due to the rapid growth of deep learning, it can automatically learn features from the spatial domain and temporal domain at the same time. In this context, we have implemented a method based on 2D CNN and GRU model can extract spaciotemporal features. Very first, we need to convert all the RGB video input data into 30-frame and 45-frame after examining the distribution of all the video to extract improved features. Then getting proper motion direction and removing undesirable impact on the background RGB video, we have calculated optical flow to focus only on the motion. Subsequently, 2D Convolutional Neural Network model can be used to obtain features individually in this video. After that, the feature of RGB-D videos and optical flow video, are fused for further refinement. Then, the fused features are passing one-by-one to Gated recurrent network for video level recognition. Resulting we acquire the last classification results by a SoftMax classifier. The details of our 2DCNN model implementation, calculation of optical flow, the 30-frame and 45-frame unification strategy, the fusion schemes, and the Gated recurrent unit will be presented in later subsections.

A. Unification of frame for input in the system

The prerequisite of 2D CNN model, input to the system is the same width and same height of each frame. we first examine the VIVA dataset, then go to experiment on, and then convert all the RGB and Depth Video to the same number of frames to each video. This VIVA dataset has 1460 total video with 19 different class [12]. We organize three set for training, testing and validating data as illustrated in Table 1.

Table 1. Details Of VIVA Dataset For Three Different Sets: Training, Testing And Validating

Sets	Category of class	Gestures Video
Training	19	1168
Testing	19	146
Validating	19	146

The most significant part of this dataset is that some gestures are very similar in appearance and timing length of the video is not the same. Therefore, it is hard to differentiate some gesture with a less number of frames, resulting not finding proper motion. The most important thing is to find a balanced number of frames with preserving proper motion information. To examine all the 1460 video, we find that around 600 videos has frames less than 35 and other video has a length of more than 40. We examine that an average number of frames is near 37, but we select 45 and 30 as a higher level and lower level number of frames. We also compare the result with these 45 and 30 frames and gives judgment that 45-frames gives better result due to the proper motion path.

B. Motion Detection using Optical flow from RGB video

As stated in an earlier section, proper motion path supports higher recognition accuracy and remove gesture irrelevant information from the background the optical flow notion is used.

A pixel I (x, y, t) in a frame, where t is time. It moves by distance (dx, dy) in the exact next frame after dt time. So, this pixel has the same intensity in the next frame is represented as below

$$I(x, y, t) = I(x+dx, y+dy, t+dt) \quad (1)$$

By taking Taylor series approximation on right side, after removing common term and divide by dt final equation is:

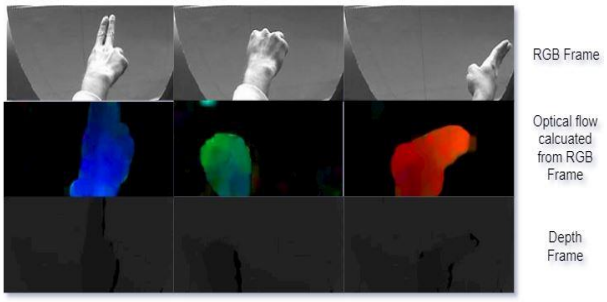

Fig. 2. First raw are RGB Frames. Second rows are calculated optical from RGB frames. Third row represent Depth information. Optical flow support exaction of motion path in video.

$$f_x u + f_y v + f_t = 0 \quad (2)$$

$$\text{where, } f_x = \frac{df}{dx}, f_y = \frac{df}{dy} \quad \text{and} \quad u = \frac{dx}{dt}, v = \frac{dy}{dt} \quad (3)$$

A represented in above for calculation of (u, v) is an optical flow equation. A simple solution after using least square fit method for 9 points and two unknown variables after solving the equation is determined as below:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i f_{x_i}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i f_{y_i}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix} \quad (4)$$

where, the value of (u, v) is an optical flow for consequent two frames. We have calculated a dense optical flow for all the frame in our existent RGB dataset and given input to our proposed model. Optical flow data emphasis only on the movement and strongly represent motion as per fig. 2.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

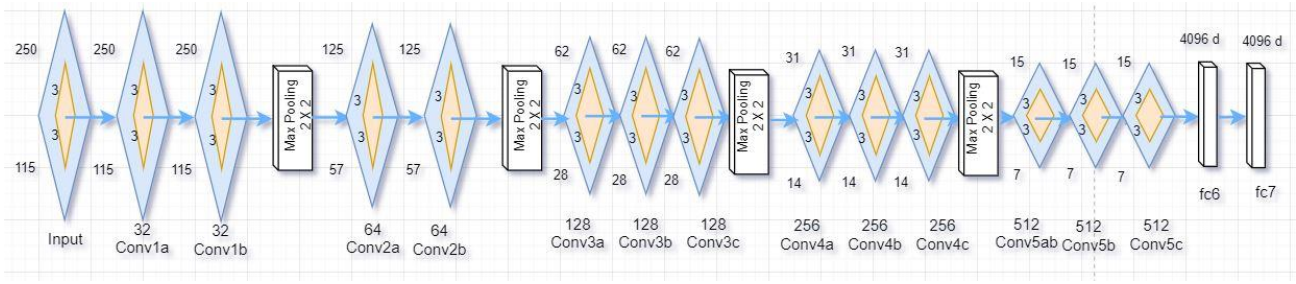

Fig. 3. The architecture of 2D CNN model. It involves 13 convolution layer, 4 Maxpooling layer and 2 fully connected layer, final size of feature vector is 4096-dim.

C. 2D CNN approach for feature extraction

Gestures are normally presented in the video, solving gesture recognition depend only on feature extraction. In the conventional CNN model, it concerns only spatial information present in a video which is not suitable. Hence, we add flow information as an extra feature to support the robustness of the model. We now introduce a network operation in detail. Each video clip as a volume V with size l x w x c x s. where l X w is image size, c number of channel and s >= 1 is a sequential number of frames. We have used three-stream network with RGB, Depth and optical flow as input, we convert each RGB, Depth and optical flow video to N number of frames where N=30 or 45. The first stream is used to find appearance information, the second stream is used depth information and third is used for exact motion information. The overall proposed CNN architecture has 5 convolution layers, 4 pooling layer, 2 fully connected layers, in each stream is illustrated in Fig. 3. Each convolutional layer has a different number of kernels. The first convolutional layer has 32 kernels with the size of 3 X 3-pixel and 2-pixel stride. The second convolutional layer contains 64 kernels with the same number of pixel and stride. The third, fourth and fifth layer has a total number of kernels is 128,256,512 with the size of 3 x 3. The final output of fully connected layer fc6 and fc7 is 4096-dim. The general size of each image in the video is 115 X 250. We did not apply to the crop of the frame because of suitable frame size. The size of Max-pooling is 2 X 2, and it is used to reduce the size of a factor of 2. We have also used Rectifier Linear Unit (ReLu) and batch normalization for faster training. The final output of 4096-dim vector for different three streams is clubbed and passed to the gated recurrent network for frame-to-frame generation of probability score.

D. Aggregation of features

In this section, we consider a different method for fusing features from Three-Stream network as shown in Table 2. Our intention here is to fuse these three networks such that pixel response at the same spatial location is fused. From these three streams, we get appearance information of hand from RGB, sharp edge feature from the depth and find motion from left-to-right, up-down from optical flow. [18] use different fusion scheme and aggregate the obtained feature to improve the performance, which creates an inspiration for us Table 2. Different fusion scheme for feature integration using MAX, SUM, AVG, CONCATE and WEIGHTED. that utilizing a fusion scheme can help for boosting the recognition accuracy. We have added a new weighted average scheme for fusion of feature because it gives more chance on those streams of feature which are appropriately related to the classification of gesture. We assign different weight for each stream and empirically check the result of different gesture. And finally conclude that due to more weight on optical flow increase the performance. Some gestures are strictly recognized due to the support of optical flow. Here X^a, X^b, X^c is RGB, Depth and OF feature vector with size 4096-dim. According to final response displayed in Fig. 4. It shows that we get a strong response after fusion with the maximal weight of optical flow stream. The recognition of gesture there is a motion of hand or finger is more precious. Some gestures are strictly recognized due to the support of optical flow because of similarity in the gesture.

A fusion function f: X_t^a, X_t^b, $X_t^c \rightarrow Y_t$, fuses three feature maps $X_t^a \in R^{H1 \times W1}$, $X_t^b \in R^{H2 \times W2}$ and $X_t^c \in R^{H3 \times W3}$ at time t for each frame and produce output $Y_t \in R^{H \times W}$ where W, H are width, height for respective feature map. We consider late fusion in our architecture and for simplicity assume that H=H1=H2=H3, W=W1=W2=W3.

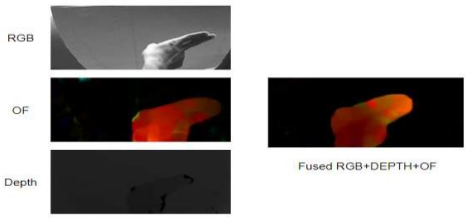

Fig. 4. Overall fusion after RGB, Depth and Optical flow, Strong response after assign more weight to optical flow.

E. Temporal recognition using gated recurrent unit

The input of this layer is fusion features with size 4096-dim after the convolution. GRU is an improved version of RNN and LSTM and trains faster due to a fewer number of gates [25]. As per the nature of the CNN, it can extract meaningful information only from a single frame. The video is a collection of frames therefor to blend all the past information and present information.

The output of each stream is 4096-dim vector. Final output after aggregation is 4096-dim Vector.

NO	Fusion	Mathematical Representation
1	Max fusion	$Y^{max} = f^{max}\{X^a, X^b, X^c\}$ means $Y_{i,j,d}{}^{max} = \max\{x_{i,j,d}{}^a, x_{i,j,d}{}^b, x_{i,j,d}{}^c\}$
2	Sum fusion	$Y^{sum} = f^{sum}\{X^a, X^b, X^c\}$ means $Y_{i,j,d}{}^{sum} = x_{i,j,d}{}^a + x_{i,j,d}{}^b + x_{i,j,d}{}^c$
3	Avg fusion	$Y^{avg} = f^{avg}\{X^a, X^b, X^c\}$ means $Y_{i,j,d}{}^{avg} = \frac{X^a_{i,j,d} + X^b_{i,j,d} + X^c_{i,j,d}}{3}$
4	Concate fusion	$Y^{concat} = f^{concat}\{X^a, X^b, X^c\}$ means $Y_{i,j,2d}{}^{concat} = x_{i,j,d}{}^a$, $Y_{i,j,2d-1}{}^{concat} = x_{i,j,d}{}^b$, $Y_{i,j,2d-2}{}^{concat} = x_{i,j,d}{}^c$
5	Weighted fusion	$Y^{weighted} = f^{weighted}\{X^a, X^b, X^c\}$ means $Y_{i,j,d}{}^{weighted} = \frac{X^a_{i,j,d} * w1 + X^b_{i,j,d} * w2 + + X^c_{i,j,d} * w3}{w1+w2+w3}$ $\sum_{i=1}^{3} w_i = 1$

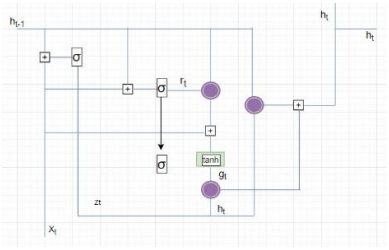

Fig. 5. Architecture Of Single Cell In GRU With Update And Reset Gate.

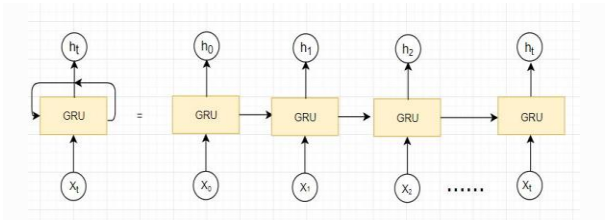

Fig. 6. Unrolling of Gated Recurrent Network. Each aggregated feature is passing one-by-one to this unit for temporal recognition for gather past and present information.

To join, all the feature Gated recurrent network is used for temporal extraction of features. The architecture of a single GRU block is characterized in Fig. 5. The memory block inside a hidden Gated Recurrent Unit (GRU) is controlled by a reset gate, memory cells, and update gate. The equation below describes the activation of a different part of the memory block and the different gate of the recurrent hidden GRU Layer. The detail working of the recurrent network is illustrated in Fig. 6. Where all the feature vector after CNN is passing one-by-one to each recurrent unit for temporal feature extraction and recognize proper motion path at different timestamp t. We have used 30-fame or 45-frame so, the total number of GRU block is 30 or 45. This is a many-to-one recurrent network; therefore, it produces final output on the last block.

F. Classification and training

The last part of our network is classification. Which convert video level representation of temporal frame to many-to-one network with the gated recurrent unit. The output of SoftMax classifier is the probability distribution of each class. The entire three stream 2D CNN and GRU network is trained using backpropagation through time (BPTT). The negative log-likelihood loss function is used to measure the difference between actual output and the predicted output. This loss function can be used for adjustment of weight for network.

$$L(p, y) = -\sum_{i=1}^{N} p_i \log y_i = -\sum_{i=1}^{N} p_i \log f(x_i, w) \qquad (5)$$

In the training phase, the idea is to minimize the cost function L reaches its minimum value with optimize weight value {w}. where N is total input video, x_i is input, $y_i = f(x_i, w)$ is a probability value of expected class. And p_i represents ground truth value. The network optimize weight according to stochastic gradient descent method. We extract the feature using the CNN model with varying number kernel and a varying number of strides. We do not fine-tune our proposed model and therefore, all network parameter is that are not pretrained with weights are initialized with random value drawn from a zero-mean and normal distribution with σ=0.01. we have used momentum 0.9 and weight decay 0.005 to train a network. We have set 0.0001 an initial learning rate and reduce it after 10000 iterations with the ratio of 0.8. After 70,000 training process stops automatically.

IV.EXPERIMENTAL RESULT

In this section, we validate the efficiency of our proposed method by a number of experimentations. In the first section 4.1, we concisely introduce our dataset used for the experiment. After that in section 4.2, an experimental environment for running model is explained, next to the evaluation for the recognition percentage is provided in Section 4.3 , then the conducted experiment that illustrate the impact of different feature extraction model, the fusion scheme, the effect of the unification of 30-frames and 45-frames strategy, the comparisons with other methods and our proposed method and confusion matrix are given in Section A-I.

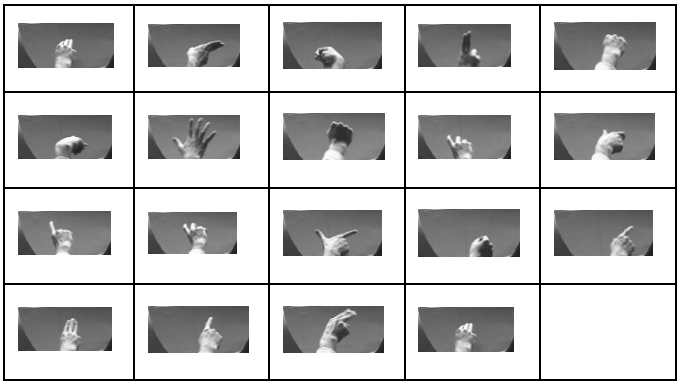

Fig. 7. VIVA dataset with snapshot of 19 different class.

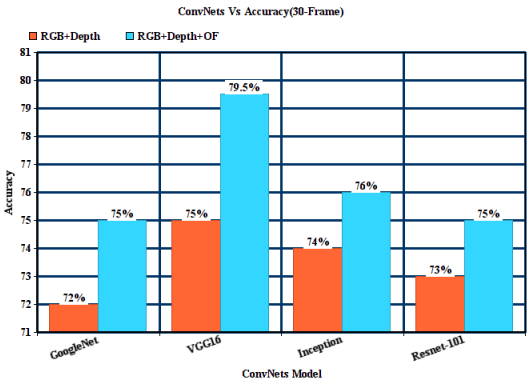

Fig. 8. Result comparison with existing feature extraction model with 30-frame as an input with optical flow and without optical flow.

A. Dataset

The VIVA dataset comprises data about usual human action to provide a touchless interface in cars as given in Fig.7. Wholly gestures composed under varying illumination and low resolution. The dataset was designed with Microsoft Kinect device with the resolution of 115 X 250 pixels in length. The standard VIVA dataset has 1460 RGB and Depth video with 19 different class and anchored by 8 different anchors in the cars.

B. Experimental Environment

We have done an experiment on a PC with Intel Core i7-8086K processor and 6 cores, 16 GB RAM and Nvidia GeForce 940MX GPU. The tests of the 2DCNN-GRU model for the train a network and extraction of features are executed under Keras with OpenCV 4.0 python 3.7.2, and TensorFlow backend with cuDNN 8.0 framework on windows 10, 64-bit operating system.

C. Evaluation Criteria

The exact calculation of the recognition rate for a reasonable comparison of the result with this dataset is defined under:

$$\text{Accuracy} = \frac{\text{True Positive}}{\text{Total No of Video}} \qquad (6)$$

Where True Positive mean ground truth class of gesture is the same as the predicted gesture.

D. Comparison with different Feature Extraction ConvNets Model

We have checked the influence of optical flow to existent CNN model for feature extraction. The standard CNN Model are Google Net [26], VGG-16 [27], Inception [28] and ResNet-101 [29]. We have empirically evaluated the performance of different ConvNets with 30-frames and 45-frames, in both the cases with and without the use of optical flow. In the graph reported in Fig. 8 and Fig. 9, it shows that performance is boosting due to the support of optical flow. And finally, we conclude that 45-frames gives better performance than 30-frames with enough number of frames.

E. Comparison with different Fusion scheme

We start our effort by assessing and evaluating the consequence with our novel recommended method 2DCNN and GRU network with distinct modality covered in the earlier section. We independently train a network for each modality by similar architecture with late fusion. Fig. 10 encompasses the precise classification accuracy of numerous combinations of modalities. We examined that fusing a distinct pair of modalities increases their classification result. The greatest gesture recognition accuracy (85%) can be achieved by weighted based fusion. We also check and find that optical flow is most prominent for hand gesture recognition. We empirically evaluate that some gesture which has minor variation those are easily differentiated by using optical flow only. The various fusion for 30-frames and 45-frames is shown in Fig. 9. In all the cases, the best accuracy obtained by, in both framing sequence with weighted based fusion and optical flow.

F. Unified 30-Frame and 45-Frame comparison

In this subdivision, we validate the efficiency of our 30-frame and 45-frame approach in detail. The classification accuracy with 30-frame and 45-frame inputs with different fusion scheme are averaged and compared as described in section 4.5. In Fig. 11. Comparison of the result between the 30-frame and 45-frame are displayed. As per observation, it can see that recognition performance is increased with the involvement of optical flow data. And finally, we can say that our strategy for 45-frame is practical and accomplishes a fantastic advancement in RGB-D and optical flow data. It means that the motion path with proper information aids differentiating different hand gesture and increasing the accuracy in huge amount.

Retrieval Number: J91850881019/19©BEIESP
DOI: 10.35940/ijitee.J9185.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

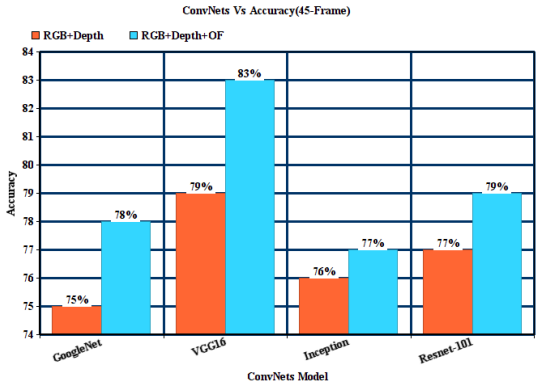

Fig. 9. Result comparison with existing feature extraction model with 45-frame as an input with optical flow and without optical flow. Performance improved with the support of optical flow in any feature extraction model.

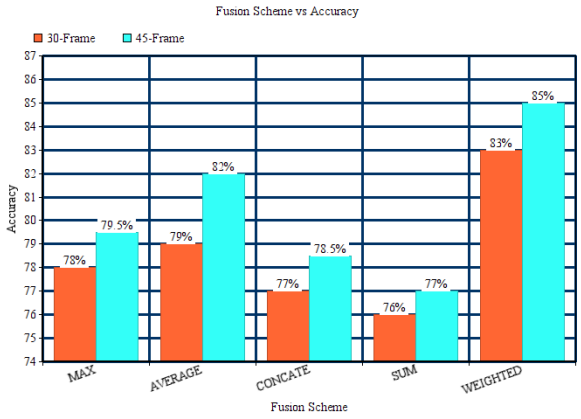

Fig.10. Fair comparison of different fusion scheme with 30-frame and 45-frame. It specifies that 45-frame with the support of optical flow and weighted scheme yield boost the performance.

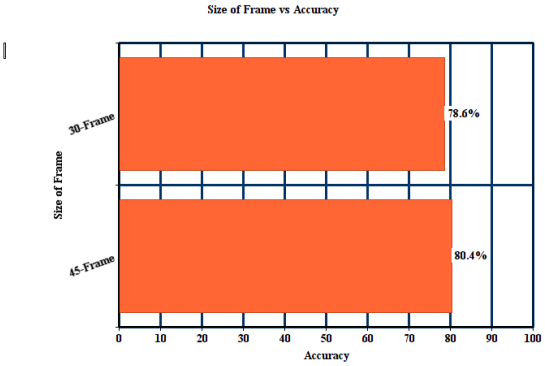

Fig. 11. Comparison of result with 2DCNN and GRU model with the input of 30-frame and 45-frame. It can specify that proper information of

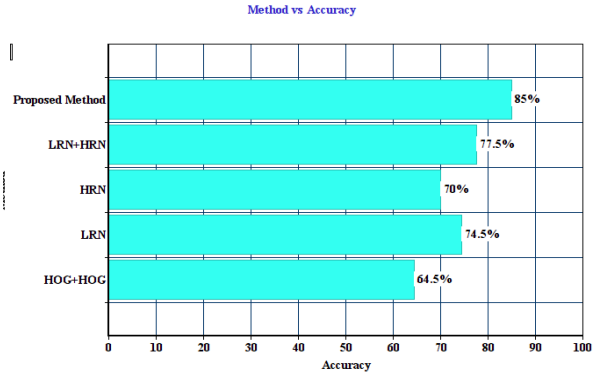

Fig. 12. The comparison results with proposed method and existing method

G. Unified 30-Frame and 45-Frame comparison

In this subdivision, we validate the efficiency of our 30-frame and 45-frame approach in detail. The classification accuracy with 30-frame and 45-frame inputs with different fusion scheme are averaged and compared as described in section 4.5. In Fig. 11. Comparison of the result between the 30-frame and 45-frame are displayed. As per observation, it can see that recognition performance is increased with the involvement of optical flow data. And finally, we can say that our strategy for 45-frame is practical and accomplishes a fantastic advancement in RGB-D and optical flow data. It means that the motion path with proper information aids differentiating different hand gesture and increasing the accuracy in huge amount.

H. Comparison with existing method

The precise classification rate of comparison with the various existing technique is specified in Fig. 12. We compare our proposed 2DCNN and GRU classifier with Ohn-Bar and Trivedi [12], HOG+HOG features and P. Molchanov [13], Use LRN, HRN and LRN+HRN convolutional neural network. Our method outperformed with classification accuracy with 85% respectively. The results indicate that our proposed architecture is outperforming than existing method given in the literature for dynamic hand gesture recognition in a car.

I. Confusion Matrix

we can evaluate the performance of all the gesture in a different class by using the confusion matrix. It is a ratio of correct prediction and total prediction in a whole gesture class. We can measure a number of misclassifications by using error-rate as shown in below.

$$\text{Classification-Accuracy}=\frac{Correct\ Prediction}{Total\ Prediction} * 100 \quad (7)$$

$$\text{Error-rate}= \left(1-\left(\frac{Correct\ Prediction}{Total\ Prediction}\right)\right) * 100 \quad (8)$$

Some gesture in a dataset is misclassified because of similarity in a gesture

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 3. The Average Confusion Matrix for Proposed 19 Gesture Classifier: R-right, L-left, U-up, D-down, CCW-Counter clockwise, CW-clockwise.

Class	1.Swipe R	2. Swipe L	3. Swipe D	4. Swipe U	5. Swipe V	6. Swipe X	7. Swipe +	8. Scroll R	9. Scroll L	10. Scroll D	11.Scroll U	12. Tap – 1	13. Tap - 3	14. Pinch	15. Expand	16.Rotate CCW	17. Rotate CW	18. Open	19. Close
1	85						2	4	2					1	2		4		
2		81	2		3		2		3				3	4	1	1			
3			80						1	4			3	2		2	3	2	3
4				87		2	2				2	3				2	2		
5		5			85	2	3					2	3						
6			2		4	79	6					3	2	2			2		
7	1	2				6	82									2	7		
8	7						2	90			1								
9		3							94		1				2				
10			3			2				91		4							
11						1					95	4							
12			3			1						86	6		3		1		
13	4		3	2	1	2						3	79	1	2	1	2		
14		2	4				2	2				2		84		2			2
15			2										5		82	4	6	1	
16			1	3			3			3			2	3	3	80	2		
17			1	4	3							2	2	2		5	79		2
18				2								1						89	8
19			2				2							3	1			5	87

As per the confusion matrix is shown in Table 3, it shows a major source of error occurred in different gesture class. Our gesture recognition algorithm is sometimes confused between Swipe X and Swipe +. The Rotate CW/CCW gesture accuracy also increased by 79% and 80% due to support of optical flow. The accuracy of 16 class is more than 80% and remaining has less than 80%. The overall accuracy of our proposed method is 85%.

V.CONCLUSION

We robustly recognize the gesturer by developing an effective method using Three stream 2DCNN and gated recurrent network. The input to the proposed method is unified with 30-frame or 45-frame for preserving proper motion information. To eliminate gesture irrelevant information the optical flow data is calculated from RGB data. The features are extracted using CNN from RGB-D and Optical flow and blended together using a weighted scheme to increase the performance. After all the fused features are passed to the gated recurrent unit for temporal recognition for a different timestamp. The output of last GRU block is connected to SoftMax classifier for prediction of final class label. The experimentation validates that the efficiency our proposed approach is excellent to some of the state-of-art techniques on VIVA dataset with 85% recognition accuracy. We have proved that 45-frame increase performance with preserving motion information. Future work will be extending the current three streams 2DCNN and GRU architecture to connectionist temporal classification (CTC) [30] loss model which train the whole network to each frame level classification. By using the CTC model its removed pre-segmentation and post-processing of video gesture data.

REFERENCES

1 Foundation, TNS India Pvt. Limited, "Distracted Driving in India-A Study On Mobile Phone Usage, Pattern & Behaviour," India, 2017.

2 Veni, Shalini AnantShanmugham, "Safe Driving using Vision-based Hand Gesture Recognition System in Non-uniform Illumination Conditions," Journal of Information and Communication Technology, vol. 12, no. 2, pp. 154-167, 2018.

3 Quentin De Smedt, H. W.-P., "SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset," Eurographics Workshop on 3D Object Retrieval, pp. 1-6, 2017.

4 Zhigang Tu, Wei Xie, Dejun Zhang,Ronald Poppe,Remco C. Veltkamp, Baoxin Li , Junsong Yuan, "A survey of variational and CNN-based optical flow techniques," Signal Processing: Image Communication, vol. 72, pp. 9-24, 2019.

5 Kanade, Bruce D. Lucas and Takeo , "An Iterative Image Registration Technique with an Application to Stereo Vision," In Proceedings of the DARPA Image Understanding Workshop, Washington, DC, USA, pp. 674-679, 1981.

6 M. Zobl, R. Nieschulz, M. Geiger, M. Lang, and G. Rigoll , "Gesture components for natural interaction with in-car devices," Gesture-Based Communication in Human-Computer Interaction, Springer, p. 448–459, 2004.

7 F. Parada-Loira, E. Gonzalez-Agulla, and J. Alba-Castro , "Hand gestures to control infotainment equipment in cars," IEEE Intelligent Vehicles Symposium, p. 1–6, 2014.

8 Jakub Konecny, Michal Hagara , "One-Shot-Learning Gesture Recognition using HOG-HOF," Journal of Machine Learning Research, vol. 15, pp. 2513-2532, 2014.

9 Gheissari, Somayeh DanafarNiloofar , "Action Recognition for Surveillance Applications Using Optic Flow and SVM," in Asian Conference on Computer Vision, Springer, Japan, 2007.

10 Fifin Ayu Mufarroha, Fitri Utaminingrum. Hand Gesture Recognition using Adaptive Network Based Fuzzy Inference System and K-Nearest Neighbor.in International Jornal of Technology, Vol. 3, pp.559-567, 2017.

11 H. Zhang, L.E. Parker , "Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos," IEEE Trans. Circuits Syst. Video Technol, vol. 26, p. 541–555, 2016.

12 Trivedi, E. Ohn-Bar and M, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," IEEE Trans. on Intelligent Transportation Systems, vol. 15, no. 6, pp. 1-10, 2014.

13 P. Molchanov, S. Gupta, K. Kim and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," IEEE Conference on Computer Vision and Pattern Recognition Workshops , pp. 1-7, 2015.

14 P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," IEEE Automatic Face and Gesture Recognition, pp. 1-8, 2015.

15 P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," IEEE Conference on Computer Vision and Pattern Recognition, p. 4207–4215, 2016.

16 T. Du, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri,, "Learning spatiotemporal features with 3d convolutional networks," IEEE International Conference on Computer Vision, pp. 1-9, 2015.

17 leni Tsironi, Pablo Barros, Cornelius Weber, Stefan Wermter, "An analysis of Convolutional Long Short-Term Memory Recurrent Neural Networks for gesture recognition," Neurocomputing, pp. 76-86, 2017.

18 C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," CVPR, p. 1933– 1941, 2016.

19 Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu, "3D Convolutional Neural Networks for Human Action Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 221-231, 2013.

20 D. Tran, L. Bourdev , R. Fergus , L. Torresani , M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in IEEE International Conference on Computer Vision, 2015.

21 Dadashzadeh, Amirhossein & Targhi, Alireza & Tahmasbi, Maryam, "HGR-Net: A Two-stage Convolutional Neural Network for Hand Gesture Segmentation and Recognition," CoRR, pp. 1-10, 2018.

22 A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014.

23 Jeff Donahue ,Lisa Anne Hendricks , Marcus Rohrbach, "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

24 Karen Simonyan, Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems, 2014.

25 Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," CoRR, pp. 1-9, 2014.

26 Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going deeper with convolutions," Computer Vision and Pattern Recognition , pp. 1-9, 2015.

27 Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Vision and Pattern Recognition, pp. 1-14, 2015.

28 C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna,, "Rethinking the Inception Architecture for Computer Vision," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826 , 2016.

29 K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

30] A. Graves, "Supervised sequence labeling, in Supervised Sequence Labelling with Recurrent Neural Networks," Springer, p. 5–13 , 2012.

AUTHORS PROFILE

Sunil A. Patel is a Research Scholar at the Gujarat Technological University, Ahmedabad. He received master's degree from S. P. University, Vallabh Vidyanagar in 2008. He is a Computer Vision researcher and his research interests includes visual representation learning, object recognition, action recognition, video analysis, and deep learning.

.

Dr. Ramji M. Makwana is a Managing Director of AIIVINE PXL Pvt. Ltd. He received Ph.D. degree from S. P. University, Vallabh Vidyanagar in 2011. He has authored several papers in major computer vision and multimedia conferences and journals. His research interests include Data mining, Soft computing and deep learning with applications on computer vision tasks, like object recognition, action recognition and Object tracking.