Video-Based Facial Expression Recognition using a Blend of 3D CNN and ConvLSTM

Sumeet Saurav
Academy of Scientific & Innovative Research
CSIR-Central Electronics Engineering Research Institute
Pilani, India
sumeet@ceeri.res.in

Tarun Kumar
Department of Computer Science
Birla-Institute of Technology and Science
Pilani, India
kumar1998.tarun@gmail.com

Ravi Saini
Integrated System Group
CSIR-Central Electronics Engineering Research Institute
Pilani, India
ravi@ceeri.res.in

Sanjay Singh
Cognitive Computing Group
CSIR-Central Electronics Engineering Research Institute
Pilani, India
sanjay@ceeri.res.in

***Abstract*—The 3-Dimensional Convolutional Neural Network (3D CNN) and Long Short-Term Memory Network (LSTM) have consistently outperformed many approaches in video-based Facial Expression Recognition (VFER). The vanilla version of the fully-connected LSTM (FC-LSTM) unrolls the image to a one-dimensional vector, which results in the loss of vital spatial information. Convolutional LSTM (ConvLSTM) overcomes this limitation by performing LSTM operations in terms of convolutions without performing any unrolling, as in the case with FC-LSTM. Motivated by this, in this paper, we propose a neural network architecture that consists of a blend of 3D CNN and ConvLSTM. The proposed hybrid architecture captures spatial-temporal information to produce competitive accuracy on three publicly available FER databases, namely the CK+, SAVEE, and AFEW. The experimental results demonstrate excellent performance without using any external emotion data with an added advantage of having a simple model with a comparatively fewer number of parameters and model size. Our designed FER pipeline is a suitable candidate for automatic recognition of facial expressions in real-time on a resource-constrained embedded platform.**

***Index Terms*—Video Facial expression recognition (VFFE); 3D convolutional neural networks (3D CNN); long short-term memory (LSTM); convolutional LSTM (ConvLSTM).**

I. Introduction

Human-computer interaction (HCI) has seen a lot of development in recent years. Video-based facial expression recognition (VFER) remains the most central component of HCI with a variety of methods proposed to date. However, VFER remains a challenging problem due to several limitations associated with the physical factors such as age, gender, ethnic background, and lack of contextual information as-well-as variations due to changes in lighting conditions, poses, and occlusion. Recently, 3-dimensional convolutional neural networks (3DCNN) have shown great ability in FER tasks by modeling spatial and temporal information simultaneously. These 3DCNN architectures have often been combined in different ways with recurrent neural network (RNN) models to capture long term dependencies and feature enhancement, as demonstrated in [1], [2].

978-1-7281-6916-3/20/$31.00 ©2020 IEEE

Some of the above hybrid models of 3DCNN and RNN perform very well in the VFER task. Based on the performance of ConvLSTM in several computer vision applications, in this paper, we explore the combination of 3DCNN and Convolutional Long-Short Term Memory (ConvLSTM) for the classification of facial expressions in a video sequence. ConvLSTM as proposed in [3], shows its superiority over fully-connected LSTM (FC-LSTM) in number of spatial-temporal tasks like Moving MNIST Dataset, Radar Echo Dataset [4]. The main advantage of ConvLSTM is that the input-to-state and state-to-state transitions are convolutional, and thus, are inherently more suited to images, as there is no loss of spatial correlation as compared to the FC-LSTM, which operates on vectorized features. We evaluate our approach on three well-known FER databases, namely CK+ [5], SAVEE [6], and AFEW [7]. Without using any additional videos with emotion labels in our training set, our approach achieved competitive performance.

A typical FER algorithmic pipeline consists of three main stages, namely the face detection, feature extraction, and feature classification. The face detection stage often accompanied by the alignment is more or less similar in all the algorithms. The techniques mainly used are the Viola-Jones and more recent deep learning-based face detectors (OpenCV). Traditionally, the feature extraction stage is done using Local Binary Patterns (LBP) [8], Histogram of Oriented Gradients (HOG) [9], Local Phase Quantization (LPQ) [10]. Since these methods are hand-crafted for their specific application, they do not generalize well in different imaging conditions such as lighting, occlusion, subjects' ethnicity, etc. More recently, works based on CNNs have been proposed for facial expres-

sion recognition and have demonstrated to achieve superior performance.

The remainder of the paper is organized as follows: Section 2 provides an overview of the related works on video-based facial expression recognition. Section 3 explains our proposed hybrid of 3D CNN and ConvLSTM. Experimental results and analysis are presented in Section 4, followed by conclusive remarks in Section 5.

II. Related Works

Techniques available in the literature for facial expression recognition (FER) could be broadly classified into two major categories viz image-based methods and video/sequence-based methods. Static image-based techniques for FER use an end-to-end trainable convolutional neural network (CNN) for the classification of facial expressions using peak expression images. For instance, the work presented in [11] used an ensemble of multiple deep networks for FER in a single static image. In [12], the authors proposed a technique for FER using a weighted fusion of features extracted from the gray-scale and local binary pattern (LBP) encoded facial images. Expression related features from gray-scale facial images were extracted using ImageNet pre-trained VGG16 CNN and, that from the LBP encoded facial images using a shallow 2D-CNN trained from scratch. The authors in [13] presented a multi-channel deep neural network that learns and fuses the spatial-temporal features for recognizing facial expressions in static images. Here optical flow image is extracted between the peak expression face image and neutral face image and used as temporal information. Moreover, the emotional face image alone is used as the spatial information.

The video/sequence-based methods for facial expression recognition, on the other hand, extract both spatial and temporal information from an expression sequence, either using a hybrid of CNN & LSTM or a 3DCNN. In [14], the authors presented an integrated deep learning framework for facial expression recognition. Two deep learning models, one extracting the temporal appearance features from gray-scale images, and the other extracting the temporal geometrical features using landmark-based geometrical features, are first trained separately. Once trained, the models are combined using a novel integration method to boost the recognition accuracy. In [15], the authors proposed a technique for video-based facial expression recognition using a combination of deep neural network (DNN) and conditional random field (CRF). Here DNN has been used to capture the spatial relationship among the expression images, whereas a linear change of CRF is used to capture the temporal relations. In another work discussed in [16], a new technique has been proposed for facial expression recognition in videos. The proposed technique captures spatiotemporal information using a combination of CNN and nested LSTM. The spatiotemporal convolutional features from the image sequences are extracted using 3D CNN and the dynamics of the facial expressions are captured by nested LSTM with two-layered architecture. The first layer LSTM called T-LSTM is used to model the temporal dynamics of the spatiotemporal features in each convolutional layer, and the second layer LSTM called C-LSTM is used to facilitate the integration of the outputs of all T-LSTMs together. This helps in encoding the multi-level features present in the intermediate layers of the network. In yet another work presented in [17], the authors have used a 3DCNN architecture that learns to extract the static information using RGB sequences and dynamic information from optical flow sequences. The performance of the presented framework was tested on three publicly available FER databases namely the CK+ database, AFEW database, and SAVEE database. Two types of techniques were used for the extraction of optical flow. The first technique used for optical flow computation is the regular optical flow and uses the Gunner Farneback's algorithm. The second type contains the accumulative motion information of facial muscle movement, so it is called accumulative optical flow. In [18], the authors have proposed several schemes for facial expression recognition in an emotion clip using a hybrid of 2D-CNN & LSTM and C3D & LSTM. Moreover, a fusion of these schemes has also been explored to enhance the performance utilizing the effectiveness of the different schemes. Moreover, a hybrid of recurrent neural network (RNN) and 3D convolutional networks (C3D) has been explored for VFER in [1]. A technique for VFER using 3D CNN incorporating deformable parts learning has been demonstrated in [19]. Zhao et al. [20], proposed a method for facial expression recognition in a video sequence using Local Binary Pattern (LBP) features extracted from three orthogonal planes (LBP-TOP). Their presented LBP-TOP operator extracts expression features from a sequence of frames in the form of histograms computed from three orthogonal planes. In [21], the authors presented a technique for VFER using a fuzzy ARTMAP neural network (FAMNN). Hyper-parameters of the FAMNN has been determined using the particle swarm optimization (PSO). In [22], the authors proposed a technique for VFER using a Cross-channel Convolutional Neural network (CCCNN). In [23], the authors proposed a technique for VFER using a part-based hierarchical bidirectional recurrent neural network (PHRNN). Moreover, a multi-signal convolutional neural network (MSCNN) has also been proposed to extract spatial features from still frames of the expression sequence. Finally, both PHRNN and MSCNN fused extracts the partial-whole, geometry-appearance, and dynamic-still information, effectively boosting the performance of facial expression recognition. A manifold learning-based framework has been explored for VFER by Liu et al. [24]. Scheme for VFER using transfer learning and spatial-temporal fusion has been demonstrated by Ouyang et al. [25] whereas a hybrid of 2D CNN and RNN has been utilized in [26]. The work presented in [27], have proposed a fusion of audio and visual expression related information for FER. The also analyzed the impact of noise on the final decision of the system.

III. Proposed Method

Figure 1 shows the block diagram representation of our proposed facial expression recognition framework. As could