LBVCNN: Local Binary Volume Convolutional Neural Network for Facial Expression Recognition from Image Sequences

Sudhakar Kumawat[1] Manisha Verma[2] Shanmuganathan Raman[1]
[1]Indian Institute of Technology Gandhinagar, India, [2]Osaka University, Japan
[1]{sudhakar.kumawat, shanmuga}@iitgn.ac.in, [2]manisha.verma89@gmail.com

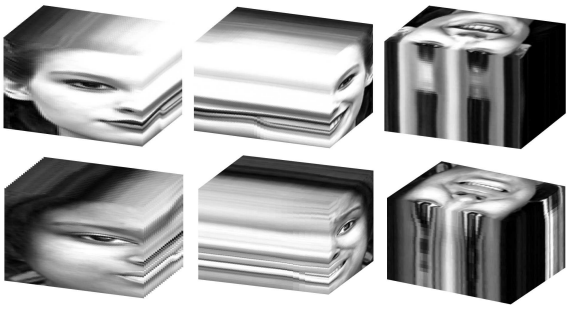
Abstract

Recognizing facial expressions is one of the central problems in computer vision. Temporal image sequences have useful spatio-temporal features for recognizing expressions. In this paper, we propose a new 3D Convolution Neural Network (CNN) that can be trained end-to-end for facial expression recognition on temporal image sequences without using facial landmarks. More specifically, a novel 3D convolutional layer that we call Local Binary Volume (LBV) layer is proposed. The LBV layer, when used with our newly proposed LBVCNN network, achieve comparable results compared to state-of-the-art landmark-based or without landmark-based models on image sequences from CK+, Oulu-CASIA, and UNBC McMaster shoulder pain datasets. Furthermore, our LBV layer reduces the number of trainable parameters by a significant amount when compared to a conventional 3D convolutional layer. As a matter of fact, when compared to a $3 \times 3 \times 3$ *conventional 3D convolutional layer, the LBV layer uses 27 times less trainable parameters.*

1. Introduction

Facial expressions are subtle signals of a larger communication process. They express one's feelings in the form of facial muscle displacements. A simple smile can indicate our liking, while a frown might show our displeasure. Thus, understanding facial expressions is an important part of our communication. In computer vision, facial expression recognition deals with the problem of recognizing basic human expressions from video or image data. The problem has many applications in the field of computer science, medicine, psychology, and other related areas.

Part of the research on this problem is focused on recognizing facial expressions from static images [24, 22, 28, 21, 15, 5, 3, 2, 41]. Although this approach is effective in extracting spatial information, it fails to capture morphological and contextual variations of the expression process. Recent methods aim to solve this problem by using temporal image sequences and utilize both spatial and temporal variations to give better recognition systems [9, 18, 32, 27, 35, 16]. Very recent methods use geometric features such as temporal variations in facial landmarks along with temporal image sequences to achieve state-of-the-art results [6, 37, 11, 38, 4]. Facial landmarks boost the accuracy of models by supplying discriminant information that steer the expression recognition process, especially with deep learning. However, detecting accurate facial landmarks is a problem by itself. Difficult visual conditions such as illumination, resolution, and alignment may further make facial landmarks detection difficult. Recently, Steger *et al.* studied the effects of trivial image distortions like rotation and Gaussian noise on the performance of facial landmarks detection algorithms [33]. The study, which is a first of its kind, showed that even state-of-the-art facial landmarks detection models like Uricar [34] and Kazemi [13] are vulnerable to image distortions. This emphasizes the need for a method that can be used in the domain of facial applications such as facial expression recognition and has an accurate performance at par with the state-of-the-art methods, while not using facial landmarks.



(a) (b) (c)

Figure 1. Space-time transitions in the third dimension for (a) XY, (b) YT, and (c) XT spaces.

In this paper, we propose a simple deep 3D Convolutional Neural Network (CNN) that can be trained end-to-

end on temporal image sequences without using any extra information such as facial landmarks. Our work is inspired from Volume Local Binary Patterns (VLBP) [40] and recently proposed Local Binary Convolutional Neural Network (LBCNN) [10]. VLBP takes a 3D neighborhood of each pixel of every frame of a video and generates the corresponding 3D LBP map. LBCNN replaces the conventional 2D convolutional layer of CNN by a Local Binary Convolutional (LBC) layer that exploits LBP concept in a CNN architecture. Normally, a video sequence is understood as a stack of XY planes along T axis, but it is easy to see that it can also be seen as a stack of XT planes along Y axis and YT planes along X axis. The XT and YT planes too have information about the space-time transitions as shown in Fig. 1. Our proposed network that we call Local Binary Volume Convolutional Neural Network (LBVCNN) captures these transitions by using three small networks LBVCNN-XY, LBVCNN-XT, and LBVCNN-YT. Each of these small networks consists of our newly proposed Local Binary Volume (LBV) layer which is a 3D variant of the Local Binary Convolution (LBC) layer of the LBCNN network. The three 3D convolutional neural networks LBVCNN-XY, LBVCNN-XT, and LBVCNN-YT are trained on the three orthogonal sides XY, XT, and YT respectively of a video cuboid. Finally, these fully trained networks are combined and then fine-tuned. The main motivation behind this idea is that the local texture information is significant in spatial structure (facial texture) as well as in minor spatio-temporal fluctuations (see Fig. 1).

The main contributions of this paper are summarized as follows.

- We propose a new network called Local Binary Volume Convolutional Neural Network (LBVCNN) that can be trained end-to-end on facial expression image sequences without using landmarks.
- Our network uses significantly fewer trainable parameters and has a lower computational cost when compared to the other conventional 3D CNN networks.
- We have validated the proposed method on CK+, Oulu-CASIA, and UNBC McMaster dataset.

The rest of the paper is organized as follows. Section 2 provides an overview of the relevant works. Section 3 discusses the architecture of our proposed network LBVCNN. Section 4 discusses the datasets used for the experiments along with the training and the implementation details. It also discusses comparison with the state-of-the-art approaches. Section 5 provides the conclusion and the future work that can be performed.

2. Related Work

Many existing techniques target facial expression recognition in images and video sequences [26]. Earlier works on facial expression recognition were concentrated on images [24, 22, 28, 21, 15, 5, 3, 2, 41]. However, they do not consider temporal variations. Facial expression process is a dynamic event which takes minute motion changes through time into account. Before the era of deep learning, handcrafted features were used to extract spatio-temporal information and to classify facial expressions. We give a brief overview of various methods that have achieved good performance on facial expression video sequences below.

Hand-Crafted Feature-Based Methods. For facial expression analysis in video sequences, many image-based features are extended in order to get temporal features along with spatial information such as LBP-TOP [40], 3D-HOG [14], and 3D-SIFT [29]. Jain *et al.* used conditional random fields and manually created shape-appearance features for temporal modeling of each facial shape [8]. Sanin *et al.* proposed spatio-temporal covariance descriptors using Riemannian locality preserving projection approach for action and gesture recognition [27]. Wang *et al.* proposed an Interval Temporal Bayesian Network (ITBN) for capturing complex spatio-temporal relations among facial muscles [35]. Liu *et al.* proposed an expressionlet-based spatio-temporal manifold method for dynamic expression recognition [18]. Ptucha *et al.* proposed a Manifold-based Sparse Representation (MSR) for expression recognition by mapping features in low dimensional manifolds using supervised locality preserving projections [24]. Recently, Sikka *et al.* proposed a Latent Ordinal Model (LOMo) for facial expression recognition in videos [31]. LOMo integrates features extracted from SIFT around the facial landmarks and LBP using a weakly supervised classifier to learn the expressions as hidden variables.

Deep Learning-Based Methods. Deep learning-based models have achieved state-of-the-art results in facial expression recognition. Liu *et al.* applied 3D CNN with deformable action part constraints (3D CNNDAP) to the problem of expression recognition [17]. Recent models use geometric features like facial landmarks to further boost the accuracy. Jung *et al.* proposed two separate networks called DTAN and DTGN and jointly fine-tuned the two networks to achieve state-of-the-art performance [11]. The DTAN network is a simple 3D convolutional network that captures spatio-temporal information from temporal image sequences. The DTGN network is a fully-connected network that captures temporal variations in facial landmarks. Guo *et al.* improved Jung *et al.*'s result and trained a spatial network (MSCNN) and a temporal network (PHRNN) separately and jointly fine-tuned them [38]. MSCNN is a simple convolutional network on peak expression images. PHRNN is a collection of subnets (recurrent neural networks) that

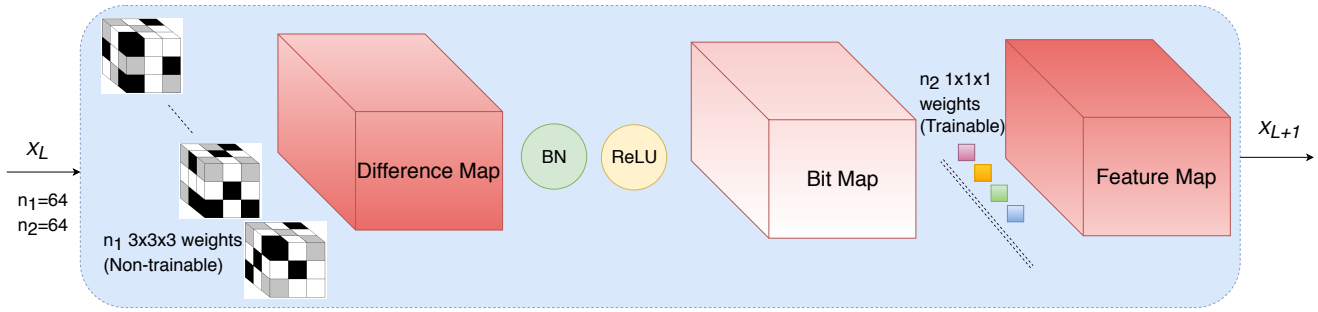


Figure 2. The proposed Local Binary Volume (LBV) block. BN- Batch Normalization. ReLU- Rectified Linear Unit

are connected in a binary tree-like structure. Facial landmarks are divided into four parts and passed at the bottom of this structure and the outputs of the subnets are concatenated at the next layer. The process is repeated for upper layers and the final layer is a softmax classification layer.

3. Our Approach

We propose a 3D convolutional neural network based architecture. Our idea is inspired from Volume Local Binary Pattern (VLBP) [40] and the recently proposed Local Binary Convolutional Neural Network (LBCNN) [10]. We give a brief description of the works that inspired our model below. A detailed description of our network is given in Section 3.1.

Local Binary Pattern. Local Binary Pattern (LBP) was proposed by Ojala *et al.* [23]. It computes a binary pattern using each pixel of an image. Every pixel of the image is treated as a center pixel and thresholded with neighborhood pixels. It assigns 0 or 1 to a neighborhood pixel if it is lesser or greater than the center pixel, respectively. Illumination invariance is an important property which makes LBP robust and it has been used in many computer vision problems for feature extraction. For a center pixel I_c and a neighboring pixel I_i ($i = 1, 2, .., p$), LBP can be formalized as follows.

$$\text{LBP}_{p,r} = \sum_{i=1}^{p} F(I_i - I_c) \times 2^{i-1} \tag{1}$$

$$F(I) = \begin{cases} 1, & I \geq 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here, p and r are the number of neighboring pixels and the radius, respectively. After construction of local binary pattern map, a histogram is created to form the feature descriptor which can be used for classification.

Volume Local Binary Pattern. In order to make LBP useful for dynamic video sequences, Zhao and Pietikainen proposed Volume LBP (VLBP) for dynamic texture recognition [40]. VLBP takes a 3D neighborhood of each pixel of every frame and generates the corresponding 3D LBP. To make it computationally simple, LBP is extracted from three orthogonal planes (XY, XT & YT) corresponding to a center pixel and called as LBP-TOP (Local Binary Pattern - Three Orthogonal Planes). Finally, all the three LBP histograms are concatenated in order to form a feature descriptor which can be fed into a classification algorithm. The feature descriptor combines motion features with spatial features and extracts significant information from the video sequences. Note that, although LBP-TOP is computationally cheap, it is not equivalent to VLBP [40]. This is because, it does not take into account all the pixels in the 3D neighborhood of a center pixel as done by VLBP. In LBP-TOP, only the co-occurrences of the local binary patterns on three orthogonal planes are taken into account [40].

Local Binary Convolutional Neural Network. Xu *et al.* [10] proposed Local Binary Convolutional Neural Network (LBCNN). In this network, the conventional convolutional layer of CNN is replaced by a Local Binary Convolutional (LBC) layer which is a generalized version of simple LBP. The LBC layer broadly consists of two sub-layers. The first layer involves convolving the input with fixed non-trainable filters of size 3×3 in order to get a difference map, followed by a ReLU activation to get an approximate local binary bit-map. The non-trainable filters contain values sampled from the set {-1,0,1} using Bernoulli distribution. The second layer is trainable and involves 1×1 convolutions on the output of the first layer in order to get feature maps. This architecture significantly reduces the number of trainable parameters as it involves training of only 1×1 filters.

3.1. Local Binary Volume Convolutional Neural Network

In order to make LBCNN useful for dynamic video sequences, a straight-forward way would be to apply it in an LBP-TOP fashion. This can be done by using three separate LBCNN networks for XY, XT, and YT planes of the video cuboid and taking the third dimension as channels, and finally combining them and fine-tuning the integrated network. However, we found experimentally that such an approach fails to fully capture the spatio-temporal variations along all the dimensions. Table 1 shows the results when

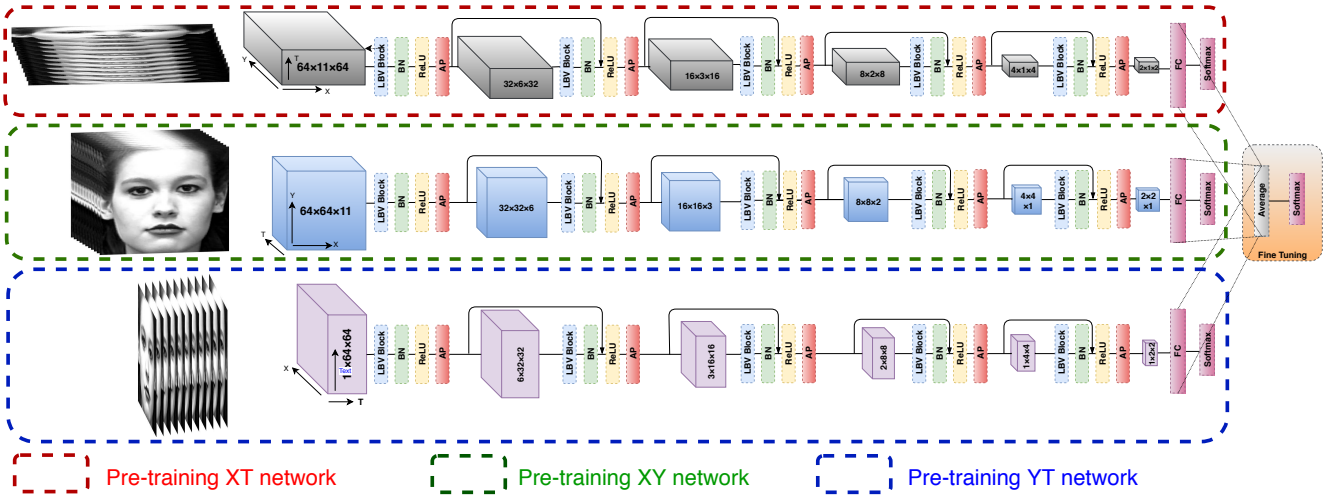


Figure 3. The proposed LBVCNN network architecture. BN- Batch Normalization, AP- Average Pooling, ReLU- Rectified Linear Unit. Red, green and blue dashed boxes represent individual LBVCNN-XT, LBVCNN-XY, and LBVCNN-YT networks.

LBCNN is applied on the CK+ and Oulu-CASIA datasets in LBP-TOP fashion. Here, the evaluation is performed using 10 fold cross validation. We think that the subtle structural difference (as discussed previously in [40]) between VLBP and LBP-TOP is responsible for such a phenomenon. In order to solve this problem, we propose a 3D variant of the LBC layer and integrate it with our newly proposed LBVCNN network as discussed below.

Method	Accuracy (%)	
	CK+	Oulu-CASIA
LBCNN-XY	91.2	72.89
LBCNN-XT	85.65	69.26
LBCNN-YT	86.1	69.24
LBCNN(joint)	92.52	74

Table 1. Results on the CK+ and Oulu-CASIA dataset when the LBCNN network is applied in a straightforward way in the LBP-TOP fashion on the sides of video cuboid.

Local Binary Volume Layer. We propose a 3D variant of the LBC layer of LBCNN network that we call Local Binary Volume (LBV) layer (Fig. 2). The LBV layer is simple and very powerful in capturing subtle spatio-temporal variations in temporal image sequences. The LBV layer consists of two sub-layers. The first layer involves convolving the input with fixed non-trainable 3D filters of size $3 \times 3 \times 3$ in order to get a 3D difference map, followed by a ReLU activation to get an approximate 3D local binary bit-map. The non-trainable 3D filters contain values sampled from the set $\{$-1,0,1$\}$ using Bernoulli distribution. The number of elements from the set $\{$-1,1$\}$ determine the sparsity of the 3D filter. The second layer is trainable and involves $1 \times 1 \times 1$ convolutions on the output of the first layer in order to get the 3D feature maps. As proved in LBCNN [10], we show experimentally that our LBV layer approximates 3D convolutional layer of the conventional 3D-CNN.

Before discussing the complete structure of our proposed network, we discuss the usefulness of the ensemble of networks in deep learning and its relevance to the problem of recognizing facial expressions from videos.

Usefulness of ensemble of networks. Training and fine-tuning CNNs is difficult as it requires experimenting with many hyperparameters, and data splits and is highly subject to overfitting. An ensemble of independently trained networks can improve the predictions by reducing the overfit and can avoid the possible poor test result of a single network [7]. However, in a data fusion ensemble model, multiple networks are necessary to analyze the heterogeneous input data [1]. In other words, independent networks learn different data modalities to make a collective classification decision. In general, spatial information of video for each frame is captured by XY plane, whereas the temporal variations can be observed using YT and XT planes. Fig. 1 shows the variations observed along all the three directional planes. Approximately only half of video cubes are shown in order to illustrate the variations clearly along all the three planes. In Fig. 1, space-time visual motion impression of rows and columns can be observed using only XT and YT planes especially around eyes and lips. By combining the information from all these three planes using an ensemble of convolutional neural networks, we can extract appearance and motion information separately.

Local Binary Volume Convolutional Neural Network. Fig. 3 shows the general architecture of our proposed network LBVCNN. It consists of three small 3D CNNs that we call LBVCNN-XY, LBVCNN-XT, and LBVCNN-YT which are shown as dashed lines in Fig. 3. The networks LBVCNN-XY, LBVCNN-XT, and LBVCNN-YT capture spatio-temporal information from the three orthogonal sides

of a video cuboid XY, XT, and YT respectively. We use ResNet like structure for all the networks. More details on the input-sizes, parameters, hyperparameters, and the structures of all the networks are given in Section 4.

Fusion fine-tuning. For fine-tuning, we drop the final softmax layer from each of the three fully trained networks LBVCNN-XY, LBVCNN-XT, and LBVCNN-YT. Further, we combine the three networks by an element-wise average of the output of the fully connected layers which is then connected to a final softmax layer for classification. The fine-tuning network is shown in Fig. 3. Note that during the fine-tuning process, the entire LBVCNN network (Fig. 3) is fine-tuned at a very low learning rate.

Space-time complexity analysis of the LBV layer. On comparing our proposed LBV layer with the convolutional layer (of size $3 \times 3 \times 3$) of the traditional 3D CNN network, we can see that the LBV layer has 27 times less trainable filters. This is due to the fact that only the second layer of LBV (with $1 \times 1 \times 1$ size filters) is trainable while the first layer has fixed non-trainable filters of size $3 \times 3 \times 3$.

Furthermore, the 3D convolution operation in LBV (first layer) contains just addition and subtraction operations due to the presence of -1, 0, and 1. This is in contrast to multiplicative floating point operations in a traditional 3D convolution layer.

4. Experiments

To evaluate our model, we conducted extensive experiments on the three popular facial expression recognition datasets - CK+, Oulu-CASIA, and UNBC McMaster shoulder pain. We start by discussing data pre-processing and augmentation.

Data preprocessing. In order to process the data through the proposed network, we perform a few preprocessing steps. Note that each video consists of varying number of frames. Therefore, in order to account for varying temporal lengths, we used the video normalization method from [42]. The method converts the video sequences of arbitrary lengths into a fixed length sequence (11 in our case). The normalized fixed length temporal patterns preserve the characteristics of the original video well [42]. Thus, it will not affect the performance of the model. Note that another recent FER work, DTAGN [11] uses the same method for video normalization with a fixed length of 11 sequences. These 11 frames represent a neutral to peak expression. Face is extracted from each frame, cropped, and resized to $64\times$ 64 size. Therefore, each temporal image sequence is of shape $64 \times 64 \times 11$ (XYT). Sample frames of a happy and surprise expressions of a subject are shown in Fig. 4.

Data augmentation. Data classification using a deep network requires a large amount of data to train the network in order to prevent overfitting. However, the datasets which have been used in this experiment contain only hundreds of videos. Hence, in order to increase the data, we perform data augmentation similar to [11]. Cropped and resized facial frames are rotated to 5°, 10°, 15°, -5°, -10°, and -15° angles. Frames are flipped and again rotated with the above six angles. Hence, a total of 14 times of original (1 original + 6 angles of original + 1 flipped + 6 angles of flipped) dataset has been created through data augmentation.

Construction of the LBV layer. As discussed in section 3.1, the LBV layer consists of two convolutional layers. The first layer is a 3-D convolution layer with 64 fixed non-trainable $3 \times 3 \times 3$ filters. This is followed by a convolution layer containing 64 trainable $1 \times 1 \times 1$ filters with a ReLU activation function in between. For the first layer, we construct a filter bank of 64 - $3 \times 3 \times 3$ filters. Each of the 64 3-D filters contains values from the set $\{$-1,0,1$\}$ sampled according to the Bernoulli distribution. Sparsity which is defined as the number of non-zero elements of each filter is kept as 0.9. Note that all the LBV layers in our experiments share the same 64 non-trainable 3D filters irrespective of the network (LBVCNN-XY, LBVCNN-XT, LBVCNN-YT, or joint network) that they are being used. Note that the 64 - $1 \times 1 \times 1$ trainable filters are not shared among the three networks and are learned independently.

Network architecture. Each of the small networks LBVCNN-XY, LBVCNN-XT, and LBVCNN-YT (see Fig. 3) take inputs of different sizes. Let XYT (in our case X=64, Y=64 and T=11) be the size of our video cuboid with XY being the spatial dimension and XT and YT being the temporal dimensions. The network LBVCNN-XY takes as input a volume cuboid of shape $64 \times 64 \times 11$ while the networks LBVCNN-XT and LBVCNN-YT take inputs of sizes $(64 \times 11 \times 64)$ and $(11 \times 64 \times 64)$ respectively. Rest of the network is same for all the three networks with the input layer followed by five consecutive LBV layers with a maxpooling layer after each LBV layer, except the last. The final LBV layer is followed by a fully connected layer of size 256 which is followed by a final softmax layer for classification. The architecture of the combined network for fine-tuning is discussed in Section 3.1 and shown in Fig. 3. Total number of parameters used in an individual network (XY, YT or XY) are .53 million where trainable parameters and non-trainable parameters are .08 million and .44 million respectively. Total number of parameters in the fusion network are 1.6 million where trainable and non-trainable parameters are .13 million and 1.47 million respectively.

Training. Our LBVCNN network architecture is shown in Fig. 3. At first, each of the subnetworks LBVCNN-XY, LBVCNN-XT, and LBVCNN-YT (see Fig. 3) were trained separately on XY-T, XT-Y, and YT-X cuboids respectively. All the subnetworks use adam optimizer with momentum 0.9, learning rate 1e-3, and are trained for 50 epochs. Finally, all the fully trained subnetworks are inte-

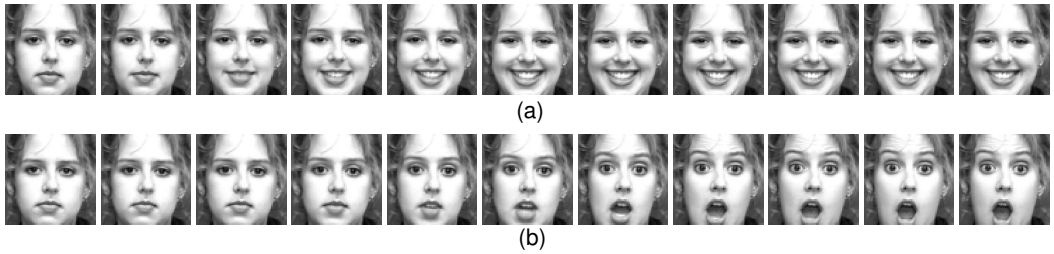


Figure 4. An example of the cropped and resized frames from CK+ dataset of (a) happy and (b) surprise emotions.

grated as shown in Fig. 3 for fine-tuning. The joint network is then fine-tuned for 100 epochs with SGD (Stochastic Gradient Descent) optimizer with momentum 0.9 and learning rate 1e-7. Throughout all our experiments, we maintain the batch-size of 16. The loss function used is categorical cross-entropy.

Testing. For testing, we adopt the k-fold cross-validation method. Details of the number of splits/folds created and the method used for their construction is provided in the description section of the datasets. Note that, while testing on a particular split/fold, we consider its unaugmented part only [11].

4.1. CK+ Dataset

Description of the dataset: Cohn-Kanade AU-Coded Expression dataset is a benchmark for facial expression recognition [12, 19]. This dataset is composed in a restricted environment where the subject is facing the camera with an empty background. Each video in the dataset starts with a neutral expression and ends with a peak expression. Each video is labeled as an expression of anger, contempt, disgust, fear, happiness, sadness, and surprise. The dataset contains a total of 327 videos collected from 118 subjects. Each video includes a varying number of frames. For the preparation of the dataset, the subjects are arranged by ID in ascending order. These subject IDs are then partitioned into 10 subsets by sampling in ID ascending order with a step size of 10 [18]. Nine subsets were used for training and the remaining one was used for validation [18]. This process is called as 10-fold cross-validation. The evaluation is performed in a subject independent way.

Results: The total accuracy of 10-fold cross-validation of our model on the CK+ dataset is shown in Table 2. Note that in order to make the comparison fair, we do not consider image-based and 3D geometry based algorithms and models from the comparison tables. The top three models DTAGN [11], LOMo [31] and PHRNN-MSCNN [38] that have recently achieved state-of-the-art accuracy use facial landmarks. Our model achieves state-of-the-art accuracy when compared to the models like HOG 3D [14], Cov3D [27], and STM-ExpLet [18] that do not use facial landmarks. It achieves results comparable to landmark-based state-of-the-art models and better results when compared to the non landmark-based models. The confusion matrix of the combined network i.e., LBVCNN (joint) on CK+ dataset is reported in Table 3. Comparison of accuracy according to each emotion among four networks is shown in Fig. 6. The accuracy in the cases of angry, contempt, disgust, happiness, and surprise is good, but the performance for sadness and fear is relatively poor.

Method	Accuracy	Landmarks	Strategy
HOG 3D [14]	91.44	×	10 folds
TMS [8]	91.89	✓	4 folds
Cov3D [27]	92.30	×	5 folds
3DCNN-DAP [17]	92.40	✓	15 folds
STM-ExpLet [18]	94.19	×	10 folds
LOMo [31]	95.10	✓	10 folds
VLBP [40]	96.26	×	10 folds
DTAGN [11]	97.25	✓	10 folds
PHRNN-MSCNN[38]	98.50	✓	10 folds
LBVCNN-XY	95.31	×	10 folds
LBVCNN-XT	95.50	×	10 folds
LBVCNN-YT	95.19	×	10 folds
LBVCNN(joint)	97.38	×	10 folds

Table 2. Comparison of various methods on the CK+ dataset in terms of average recognition accuracy of seven expressions. Note that in order to make the comparison fair, we do not consider image-based and 3D geometry based algorithms and models. VLBP [40] results are for six expressions only.

	An	Co	Di	Fe	Ha	Sa	Su
An	97.63	0	2.37	0	0	0	0
Co	0	100	0	0	0	0	0
Di	0	0	100	0	0	0	0
Fe	0	0	0	88.05	7.97	3.98	0
Ha	0	0	0	0	100	0	0
Sa	2.62	2.62	0	2.62	0	92.14	0
Su	0	0	0	0	1.28	0	98.72

Table 3. Confusion matrix of LBVCNN (joint) on CK+ dataset.

4.2. Oulu-CASIA Dataset

Description of the dataset: The Oulu-CASIA dataset consists of six expressions (surprise, happiness, sadness, anger, fear, and disgust) from 80 subjects under visible light condition [39]. Subjects are between 23 to 58 years old and

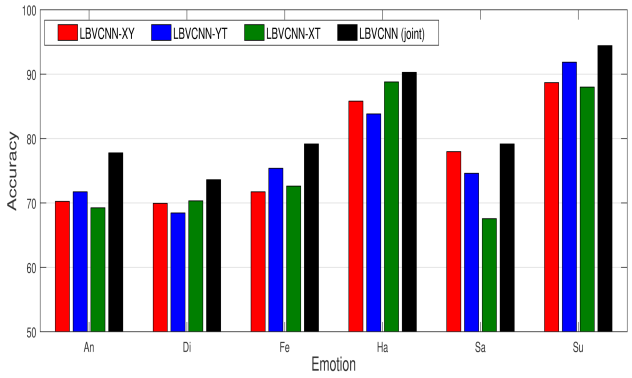


Figure 5. Comparison of accuracy according to each emotion among four networks on CK+ dataset.

73.8% of the subjects are males. It has a total of 480 video sequences, 6 each for 80 subjects. The dataset provides cropped version (only face) of the original frames. We have performed similar preprocessing as in CK+ dataset on cropped frames. Only 11 frames per video are considered and resized to 64 × 64. Each video sequence starts with a neutral expression and ends with a peak expression. Preparation of this dataset is done similarly to that of CK+ dataset.

Results: The total accuracy of 10-fold cross-validation of our model on the Oulu-CASIA dataset is shown in Table 4. Note that the models DTAGN [11], LOMo [31], and PHRNN-MSCNN [38] that have recently achieved state-of-the-art accuracy use facial landmarks. Thus, geometric features certainly boost the performance of expression recognition models. Our model achieves state-of-the-art accuracy when compared to the models like HOG 3D [14], AdaLBP [39], and STM-ExpLet [18] that do not use facial landmarks. It achieves results comparable to the landmark-based state-of-the-art models except the PHRNN-MSCNN[38] and better results when compared to all the non landmark-based models. The confusion matrix of the combined network i.e. LBVCNN (joint) is reported in Table 5. Comparison of accuracy according to each emotion among the four networks is shown in Fig. 5. The accuracy in the cases of fear, happiness, sadness, and surprise is good, but the performance for anger and disgust is relatively poor. In particular, there is a high degree of confusion among the expressions anger, disgust and sadness as they happen to look similar in particular facial region.

4.3. UNBC McMaster Shoulder Pain Dataset

Description of the dataset: Unlike CK+ and Oulu-CASIA datasets which are in controlled setting, UNBC McMaster dataset is in spontaneous setting [20]. This makes the task of facial expression recognition even more challenging. The dataset consists of real world videos of subjects with pain while performing guided movements of their affected and unaffected arms in a clinical interview. The videos are rated for pain intensity (0 to 5) by trained experts. Following [31], we labeled videos as "pain" for intensity above 3 and "no pain" for intensity 0, and discarded the rest. This resulted in 149 videos from 25 subjects with 57 positive and 92 negative samples. Following [25], a temporal window of 0.5 seconds is taken. The process of data pre-processiong and augmentation is same as that of CK+ and the Oulu-CASIA datasets. Unlike the case of CK+ and Oulu-CASIA datasets, the validation protocol used is "leave one subject out" which is same as the works mentioned in Table 6.

Results: The total accuracy of "leave one subject out" cross-validation of our model on the UNBC McMaster dataset is shown in Table 6. Note that the models MS-MIL [30], MIL-HMM [36], RMC-MIL [25], and LOMo [31] use

Method	Accuracy	Landmarks	Strategy
HOG 3D [14]	70.63	×	10 folds
AdaLBP [39]	73.54	×	10 folds
STM-ExpLet [18]	74.59	×	10 folds
DTAGN [11]	81.46	✓	10 folds
LOMo [31]	82.10	✓	10 folds
PHRNN-MSCNN[38]	86.25	✓	10 folds
LBVCNN-XY	77.40	×	10 folds
LBVCNN-XT	77.59	×	10 folds
LBVCNN-YT	76.09	×	10 folds
LBVCNN(joint)	82.41	×	10 folds

Table 4. Comparison of various methods on the Oulu-CASIA dataset in terms of average recognition accuracy of six expressions. Note that in order to make the comparison fair, only video based methods are included.

	An	Di	Fe	Ha	Sa	Su
An	77.78	6.94	4.17	0	11.11	0
Di	13.89	73.61	1.39	2.78	8.33	0
Fe	0	5.56	79.17	2.78	5.56	6.94
Ha	1.39	1.39	5.56	90.28	1.39	0
Sa	12.5	5.56	2.78	0	79.17	0
Su	0	1.39	4.17	0	0	94.44

Table 5. Confusion matrix of LBVCNN (joint) on Oulu-CASIA dataset.

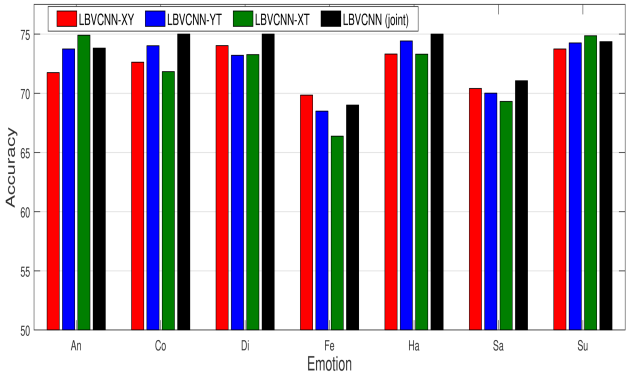


Figure 6. Comparison of accuracy according to each emotion among four networks on Oulu-CASIA dataset.

landmarks to achieve state-of-the-art results. Our method achieves better results than all, except the LOMo [31] where we achieve comparable results.

Method	Accuracy	Landmarks
MS-MIL [30]	83.7	✓
MIL-HMM[36]	85.2	✓
RMC-MIL[25]	85.7	✓
LOMo[31]	87.0	✓
LBVCNN-XY	84.76	×
LBVCNN-XT	83.20	×
LBVCNN-YT	83.48	×
LBVCNN(joint)	86.55	×

Table 6. Comparison of various methods on the UNBC McMaster shoulder pain dataset in terms of average recognition accuracy of pain and no pain expressions.

4.4. Feature Visualization

In this section, we visualize the learned feature maps of our LBVCNN model. Fig. 7 show the feature maps learned by our multi frame-based CNN in the first layer for expressions angry, happy, and surprise respectively on the CK+ database. For the sake of simplicity, only six out of sixty four filters feature maps are shown. Here, blue and red represent the high and the low response values. We observe that our model is able to capture the facial expression movements very effectively. Furthermore, the learned feature maps are consistent, for example the feature maps corresponding to the starting frame, which is a neutral frame for each emotion sequence, have approximately same visualization. Fig. 8 shows the failure cases from CK+ and UNBC McMaster shoulder pain datasets, respectively. We observe that, for UNBC McMaster dataset the videos with true label as "pain" and misclassified as "no pain" are high. The number of cases where videos with true label as "no pain" being misclassified as "pain" are very less.

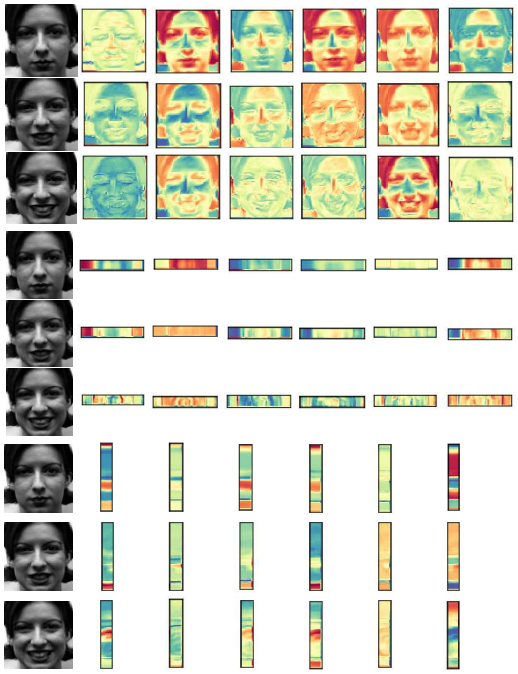


Figure 7. Feature maps learned by the LBVCNN-XY (left), LBVCNN-XT (middle), and LBVCNN (right) for the happy emotion. Blue and red represent the high and low response values.

Figure 8. Failure cases from the CK+ (left) and UNBC (right) datasets. P- Predicted, T- Target.

5. Conclusion

A novel 3D-CNN is proposed in order to recognize facial expressions from image sequences in an end-to-end fashion. The method can be performed directly on image sequences without any additional information such as facial landmarks. In particular, local binary volume layer (an efficient replacement of 3D-CNN layer) is proposed based on the concept of volume local binary pattern. LBV layer saves a significant number of trainable parameters when compared to conventional 3D-CNN layer. Our proposed network, LBVCNN, achieves comparable results on CK+, Oulu-CASIA and UNBC McMaster shoulder pain datasets. Most of the state-of-art methods use facial landmarks to extract geometric features. Since detecting landmarks is a difficult problem by itself and the problem becomes more complex with changes in illumination, resolution, and orientation, our work is of significant use as it does not use landmarks to drive the expression recognition process.

In future, we will seek utilization of local binary volume layer in other face video based computer vision problems such as face recognition and biometrics (e.g., age, ethnicity, gender recognition). In such problems, geometric features (e.g. facial landmarks) are used to boost the accuracy of the models. We shall explore other video based applications where additional features are required to boost the accuracy of the model.

Acknowledgments. Sudhakar Kumawat was supported by TCS Research Fellowship. Shanmuganathan Raman was supported by SERB Core Research Grant and Imprint 2 Grant.

References

[1] Zainal Ahmad and Jie Zhang. Combination of multiple neural networks using data fusion techniques for enhanced nonlinear process modelling. *Computers & Chemical Engineering*, 30(2):295–308, 2005.

[2] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James OReilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 302–309, 2018.

[3] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *ACM International Conference on Multimodal Interaction*, pages 423–426, 2015.

[4] Jiayu Dong, Huicheng Zheng, and Lina Lian. Dynamic facial expression recognition based on convolutional neural networks with dense connections. In *IEEE International Conference on Pattern Recognition*, pages 3433–3438, 2018.

[5] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *ACM International Conference on Multimodal Interaction*, pages 445–450, 2016.

[6] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734, 2013.

[7] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[8] Suyog Jain, Changbo Hu, and Jake K Aggarwal. Facial expression recognition with temporal modeling of shapes. In *IEEE International Conference on Computer Vision Workshops*, pages 1642–1649, 2011.

[9] László A Jeni, András Lőrincz, Zoltán Szabó, Jeffrey F Cohn, and Takeo Kanade. Spatio-temporal event classification using time-series kernel based structured sparsity. In *European Conference on Computer Vision*, pages 135–150, 2014.

[10] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local binary convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2017.

[11] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[12] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[13] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[14] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 2008.

[15] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2584–2593, 2017.

[16] Zi-Jun Li, Yu-Hung Liu, An-Sheng Liu, Yu-Huan Yang, Tso-Hsin Yeh, and Li-Chen Fu. Temporal-contrastive appearance network for facial expression recognition. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 2359–2364, 2018.

[17] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer vision*, pages 143–157, 2014.

[18] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.

[19] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 94–101, 2010.

[20] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *IEEE Face and Gesture*, 2011.

[21] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 558–565, 2017.

[22] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2016.

[23] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[24] Raymond Ptucha, Grigorios Tsagkatakis, and Andreas Savakis. Manifold based sparse representation for robust expression recognition without neutral subtraction. In *IEEE International Conference on Computer Vision Workshops*, pages 2136–2143, 2011.

[25] Adria Ruiz, Joost Van de Weijer, and Xavier Binefa. Regularized multi-concept mil for weakly-supervised facial behavior categorization. In *BMVC*, 2014.

[26] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[27] Andres Sanin, Conrad Sanderson, Mehrtash T Harandi, and Brian C Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *IEEE Workshop on Applications of Computer Vision*, pages 103–110, 2013.

[28] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, 26(4):1965–1978, 2017.

[29] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM International Conference on Multimedia*, pages 357–360, 2007.

[30] Karan Sikka, Abhinav Dhall, and Marian Stewart Bartlett. Classification and weakly supervised pain localization using multiple segment representation. *Image and vision computing*, 32(10), 2014.

[31] Karan Sikka, Gaurav Sharma, and Marian Bartlett. Lomo: Latent ordinal model for facial analysis in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5589, 2016.

[32] Karan Sikka, Tingfan Wu, Josh Susskind, and Marian Bartlett. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision*, pages 250–259, 2012.

[33] Andreas Steger and Radu Timofte. Failure detection for facial landmark detectors. In *Asian Conference on Computer Vision*, pages 361–376, 2016.

[34] Michal Uřičář, Vojtěch Franc, Diego Thomas, Akihiro Sugimoto, and Václav Hlaváč. Real-time multi-view facial landmark detector learned by the structured output svm. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, volume 2, pages 1–8, 2015.

[35] Ziheng Wang, Shangfei Wang, and Qiang Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429, 2013.

[36] Chongliang Wu, Shangfei Wang, and Qiang Ji. Multi-instance hidden markov model for facial expression recognition. In *IEEE Face and Gesture*, 2015.

[37] Aliaa AA Youssif and Wesam AA Asker. Automatic facial expression recognition system based on geometric and appearance features. *Computer and Information Science*, 4(2):115, 2011.

[38] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.

[39] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[40] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[41] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442, 2016.

[42] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Towards a practical lipreading system. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.