# Dynamic Facial Expression Recognition based on Two-Stream-CNN with LBP-TOP

**DuoFeng[1], Fuji Ren[2]**

[1] Institute of Technology and Science, Tokushima University, Tokushima, Japan
duofeng_jp@yahoo.co.jp
[2] Faculty of Engineering, Tokushima University, Tokushima, Japan
ren@is.tokushima-u.ac.jp

**Abstract:** In this paper, we propose a dynamic facial expression recognition (FER) method based on two-stream architecture with both Spatial and Temporal Convolutional Neural Network (CNN) with LBP-TOP feature. The proposed system focuses on spatial feature with obvious expression frame and also focuses on the temporal information in all expressions sequence changed from non-expression frame. This two-stream architecture is validated in the field of action recognition in video by track the optical flow information on temporal part. In this paper, we utilized LBP-TOP feature which able to extract the spatial-temporal feature on facial expression change process, and is validated the effectiveness on this field. Our proposed method is evaluated using CK+. And the results are comparable to the state-of-the-art methods to prove the effectiveness of proposed architecture.

**Keywords:** LBP-TOP; Facial Expression Recognition; Two-Stream-CNN; Emotional Calculation.

## 1 Introduction

Create Human Machine Interaction (HMI) systems that able to reach the full emotional and social capabilities for rich and robust interaction with human beings will be a long and arduous but important task. Facial expression recognition is a vital research field for reach this goal, since the facial expression is one of the most important nonverbal channels for present emotions and intentions, which proved by psychological research. And psychologist identified the six basic facial expressions (viz. anger, disgust, fear, happiness, sadness, and surprise), which considered to be universal emotional expressions among human beings [1].

Based on the importance of facial expression in designing HMI and Human Robot Interaction (HRI) systems [2], Automated Facial Expression Recognition (FER) is still a challenging issue of study for decades. Numerous approaches have been proposed for automated FER, attempt to classify faces in a given single image or sequence of images as one of the six basic expressions (seven expressions for some research). Early research mostly focused on expression analysis from static facial images, and traditional machine learning approaches such as SVM, have been successful when classifying posed facial expressions in a controlled environment [3]. But the FER problem becomes even more difficult when we tried to recognize the expressions on sequence level. The three phases of a dynamic pattern facial expression: onset, peak and offset, the entire event is so quick in most of the times to makes the process of FER become more challenging [4]. Recently, there exist many annotated face databases have consecutive frames of exhibiting the expression from onset to offset, with either human acting basic expressions (CK+ [5], MMI [6], DISFA [7]), or spontaneously captured faces in an uncontrolled environment (DISFA [7]). With approaches of dynamic representation by extract the local spatio-temporal features such as LBP-TOP [8], HOG 3D [9], which extracted in local area and then pooled over the whole sequence or some part of segments, to obtain a representation with certain length independent of time resolution. The facial expression change could been represent in temporal level. Furthermore, with the help of Deep Neural Network (DNN), expression change been able to encode by 3D Convolutional ((3D-CNN) or Recurrent Neural Network (RNN) more promising results are reported in FER field. [10, 11].

In this paper, we propose a dynamic facial expression recognition method based on two-stream architecture with both Spatial and Temporal Convolutional Neural Network (CNN). In spatial part, we utilized the expression frame of each facial expression sequence and trained an Inception layer based CNN. In temporal part, we utilized LBP-TOP instead of optical flow as the temporal feature, and combined frames' features as channels of sequence to train the CNN. And fused two streams of CNN to recognize the facial expression sequence. FER experiments were conducted with CK+ database.

## 2 Related work

Traditional algorithms for automatic FER consist of pretreatment, feature extraction, and classification as three main modules. Conventional algorithms computing the hand-crafted feature such as Local Binary Patterns (LBP) [12], Local Phase Quantization (LPQ) [13] and more directly computing the facial landmarks [14] from faces to recognize the static facial expression. Then in order to recognize the facial expression on sequence level, and sequence labeling task could get batter

**978-1-5386-6005-8/18/$31.00 ©2018 IEEE**

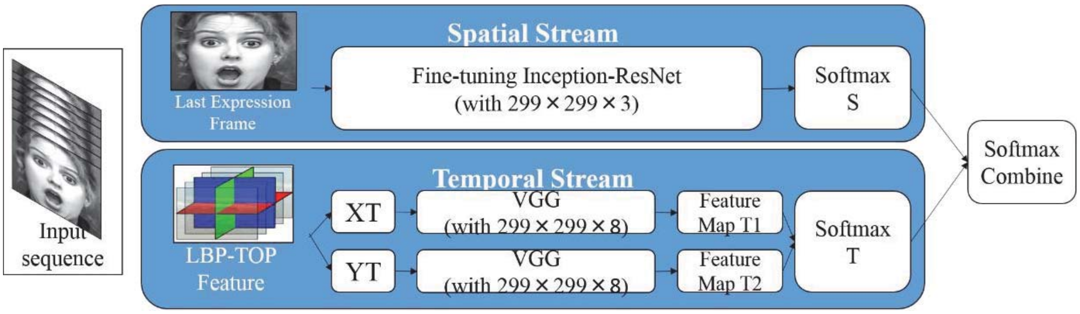

**Figure 1** Two-Stream CNN architecture

recognition result by extract the temporal relations of frames, these temporal relations has been studied using traditional methods such as Hidden Markov Models (HMM) [15], Conditional Random Fields (CRFs) [16, 17], or by another way to extract the spatio-temporal feature such as Optical Flow [18, 28] and LBP-TOP.

LBP-TOP (Local Binary Patterns from Three Orthogonal Planes) is a kind of spatio-temporal feature created by the LBP feature added the time information. Single image as only one orthogonal plane as XY, for image sequence they have space-time transition T to create the XT and YT planes. LBP-TOP kept this feature and proved to be very effective in expressing temporal feature [19].

In recent years, CNN have become the most popular approach among researchers in the field. Szegedy et al. [20] introduced GoogLeNet which is composed of multiple "Inception" layers. Mollahosseini et al. [11] has discussed about have used the Inception layer for the task for single frame FER and achieved state-of-the-art results.

Both action recognition and FER on sequence, there are mainly four DNN approaches: simple CNN, by treat the frames of sequence as the channels of input data; Recurrent Convolutional Network (LRCN) [21], by combining CNNs and RNNs which is both spatially and temporally deep; 3D-CNN [10, 22], by slide over the temporal dimension of the input data as well as the spatial dimension enabling the network to extract feature maps containing temporal information; Two-Stream-CNN [23] or Multi-Stream Deep Network [23], by fusing the multi-stream CNN each represent different spatial and temporal information.

At [22] Karen et al. proposed the two-stream architecture with spatial CNN of single frame and temporal CNN of optical flow, then discussed two fusion methods: averaging or training a multi-class linear SVM.

## 3 Proposed Method

By considering extraction of the temporal relations of facial expression sequence, we referenced action recognition field's research and proposed the Two-Stream architecture. Our proposed method, extracts both spatial and temporal features of the sequences in two streams with two end-to-end neural networks. Another component of our method is utilized LBP-TOP feature replaced optical flow as temporal feature. In the following, we explain each of the aforementioned units in detail. The Two-Stream-CNN architecture has shown in Figure 1.

### 3.1 Spatial Stream CNN

Since the AlexNet [25] got great success on the ILSVRC test, CNN also got great development. GoogLeNet [20] and ResNet [26] bring the CNN architecture going deeper. Inception-ResNet [27] combined those two architectures got better result. Paper [11] has shown the Inception architecture also have remarkable results in FER. The proposed spatial CNN utilized Inception-ResNet-v2 architecture from [27]. Inception-ResNet-v2 has totally 572 layer so we utilize fine-tuning method to avoid training a large number of parameters.

In our proposed architecture, the spatial part extract the feature on static frame of facial expression. Normally, the expression sequence in our database changed from non-expression frame to expression frame, and showing the changing process. From the sequence of facial expression, we select the final frame as the expression frame for our spatial stream CNN to extract.

Also we register facial frames in our work. In many FER research, facial landmarks were extracted to locate the facial components (such as eyebrows, lip corners, eyes, etc.) for next process. In our proposed system, we extract 68 facial landmark point, by Active Shape Model (ASM) [34]. We use these points to register facial frames to average face frame with affine transformation. After register process, the frames were resized to 299×299 pixels for next analysis. We keep the size of frame with a high resolution instead lower resolution for utilize the fine-tuning method to our network.

### 3.2 Temporal Stream CNN

The temporal part aimed to extract the change information from the expression sequence. Instead of using RNN or 3D-CNN architecture network, we extracted the spatio-temporal feature from sequence and utilized VGG liked architecture to extract the temporal information. Optical flow were extracted as spatio-temporal feature in action recognition research. But in our method we selected LBP-TOP as the spatio-temporal feature.

By consider the space-time transition T with LBP, LBP-TOP could extract features on three orthogonal planes as XY, XT, YT. Set *t* as the length of expression sequence, LBP-TOP get the XT and YT feature with length of *t-2*. For each frame in sequence LBP-TOP needs one frame before and one frame after this frame, totally three frames to extract the temporal information. We set *t*=10 for each sequence by down sampling or up sampling from original dataset sequence. And extracted 8 frames XT and YT LBP-TOP. The XY orthogonal plane feature extracted by LBP-TOP (also the normal LBP feature) shown the static feature of each frame. Since we extract the static feature by spatial stream part from original expression frame, only XT, YT orthogonal planes LBP-TOP feature needed and as input of our temporal stream.

Each orthogonal plane was trained by a VGG architecture CNN. Input of each VGG network with all frames XT or YT orthogonal plane feature, as 8 channels, to create feature maps. As the Figure 1 shown, we trained the end-to-end net by combining two feature maps before doing softmax. Table I shows the architecture of temporal stream CNN. In Table I, treat *t* as length of LBP-TOP feature per sequence, *Classes* as number of facial expression classes, Conv3 as a convolution layer with (3×3) kernel size. At the end of two orthogonal plane net, we concatenated the output, and then calculated the *loss.*

**Table I** The architecture of temporal stream CNN.

| LBP-XT | LBP-YT | Output Shape |
|---|---|---|
| Input | Input | (299, 299, *t*) |
| Conv3×2+Maxpool | Conv3×2+Maxpool | (144, 144, 64) |
| Conv3×2+Maxpool | Conv3×2+Maxpool | (72, 72, 128) |
| Conv3×3+Maxpool | Conv3×3+Maxpool | (36, 36, 256) |
| Conv3×3+Maxpool | Conv3×3+Maxpool | (18, 18, 512) |
| Conv3×3+Maxpool | Conv3×3+Maxpool | (9, 9, 512) |
| GlobalAvgPool | GlobalAvgPool | (512) |
| Concatenate | | (1024) |
| FC+Dropout | | (512) |
| Softmax | | (*Classes*) |

## 3.3 LBP-TOP vs. Optical Flow

In our proposed method, we utilized LBP-TOP feature to replaced optical flow, which utilized in action recognition field normally.

Optical flow is the pattern of apparent motion of objects between two consecutive frames caused by the movement of object, and showing the movement of points between frames by 2D vector field. Optical flow works on several assumptions: the pixel intensities of an object do not change between consecutive frames; and neighboring pixels have similar motion. Different with action recognition the body part or objects motion work like rigid motion in consecutive frames, facial part (eyes, mouth) changing between consecutive frames with tiny, non-rigid and complex. Also the illumination changes is a problem of FER in real environment. The illumination changes have a great influence on the optical flow.

However, same with LBP, one of the most important property of the LBP-TOP is its robustness to monotonic gray-scale changes caused. LBP-TOP compares pixels with neighboring pixels, and this relation was hard been influenced by illumination variations. Another important property is its computational simplicity, which makes it possible to analyze images in challenging real-time settings. LBP-TOP has been proved to be very effective on FER [8].

Figure 2 shows the LBP-TOP feature image. With a frame set (*t-1, t, t+1*), LBP-TOP output three feature images. The XT and YT showing the X-direction and Y-direction variations with time information.

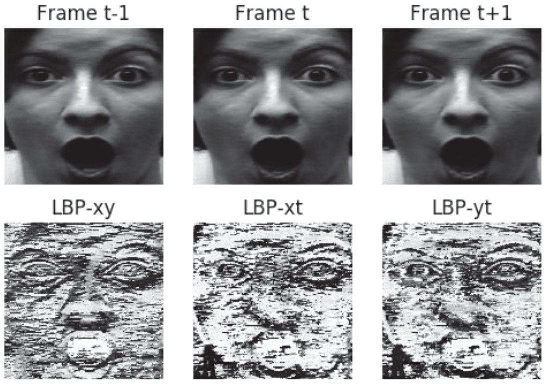

**Figure 2** An example of LBP-TOP feature.

## 4 Experiments and Results

Our method mainly aimed with FER on sequence level, databases of independent, irrelevant still images containing only facial expressions cannot be examined by our method. By considering the openness and the comparability, we evaluate our proposed method on CK+ [5] database.

### 4.1 Introduction of Facial Database

CK+: The extended Cohn-Kanade database (CK+) [5] contains 593 videos from 123 subjects. However, only 327 sequences from 118 subjects contain facial expression labels with seven expressions (angry, contempt, disgust, fear, happy, sad, and surprise). Sequences in CK+ start from the non-expression frame and end at one expression of the seven expressions with a variable sequence length from 6 to 71. In order to make the database compatible with our network, we considered down sampling or up sampling the sequence length into a ten frames sequence, totally 327 sequence on seven expression classes in our experiment.

### 4.2 Results

The input of each data were pretreatment into two part. In spatial part we resized the face frames to 299×299 pixels. The temporal part first down sampling or up sampling into ten frames per sequence. And then extract LBP-TOP feature into eight frames on XT and YT orthogonal planes, which also resized into 299×299 pixels. Each temporal CNN treat eight frames LBP-TOP feature as eight channels, and the input of whole CNN will be (299×299×3, 299×299×8, 299×299×8) as the (static frame input, XT input, YT input). The temporal stream CNN only gave one softmax output.

The two stream CNN outputs were averaged as Two-Stream-CNN output. Table II gives the accuracy when classifying into seven expressions. Table III gives the comparison the state of the art on CK+ database. Figure 3 show the resulting confusion matrices of our Two-Stream CNN .on CK+ database.

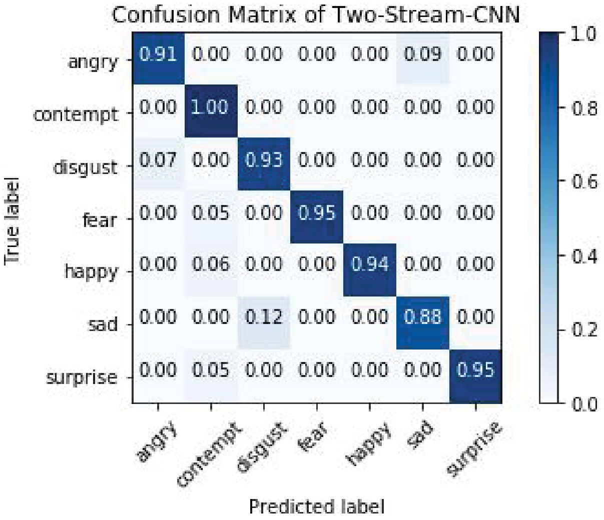

**Figure 3** The confusion matrices of Two-Stream CNN on CK+ database.

The spatial stream result got the accuracy of 92.3% on test data by recognize the static facial expression frame with fine-tuning Inception-ResNet. The temporal stream result got 78.8% on test data by recognize the LBP-TOP feature on XT and YT orthogonal planes with VGG liked networks. And Two-Stream-CNN combined both spatial and temporal networks and calculate the average outputs, got the accuracy of 93.7%.

**Table II** The accuracy (%) of individual and two-stream CNN

| CK+ Database | Spatial CNN | Temporal CNN | Two-Stream CNN |
|---|---|---|---|
| **Training data** | 99.6 | 99.4 | |
| **Testing data** | 92.3 | 78.8 | 93.7 |

**Table III** The comparison of the state of the art.

| Methods | Accuracy (%) on CK+ |
|---|---|
| Sparse representation [29] | 84.4 |
| SPSD [30] | 88.5 |
| AUDN [31] | 92.1 |
| Inception [11] | 93.2 |
| 3D CNN [32] | 92.4 |
| $l_p$ norm MKL multiclass-SVM [33] | 93.6 |
| Inception-ResNet+CRF [17] | 93.0 |
| 3DInception-ResNet+landmarks [10] | 93.2 |
| LBP-TOP+VLBP [8] | 95.2 |
| **Our Two-Stream-CNN** | 93.7 |

## 5 Conclusions

In this paper, we presented a Two-Stream-CNN for the task of facial expression recognition in sequence. Proposed Two-Stream architecture which extends from well-known Two-Stream-CNN for action recognition field. By utilized Inception-ResNet module to extract static expression frame as spatial stream part, and extract LBP-TOP feature with VGG liked network as temporal stream part.

Compare with the spatial stream CNN got 92.3% accuracy, the temporal stream CNN only got 78.8%. Because of CK+ only has 327 labeled sequences, and we can't utilize fine-tuning method with we treat frames' feature as channels of input data. We tried regularization method but still got overfitting easily. However, even if the accuracy of temporal stream network is not high, the two stream architecture shows that temporal network could still provide information between frames, and improve the overall accuracy, got 93.7% accuracy rate for Two-Stream-CNN.

We evaluated our proposed method in well-known database CK+, and comparison the experiment result with many of the state-of-the-art methods, showing that our proposed method was effectiveness.

In future work, we will try other architecture of DNN (3DCNN, LRCN etc.) to improve temporal part. And we will do more test on other facial expression database, especially in the wild environment. We also considering utilize other stream feature to improve the result rate.

## Acknowledgements

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

## References

[1] Ekman P, Friesen W V. Constants across cultures in the face and emotion. Journal of personality and social psychology, 1971,17(2):124.

[2] Mollahosseini A, Graitzer G, Borts E, et al. Expressionbot: An emotive lifelike robotic face for face-to-face communication. Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on. IEEE, 2014:1098-1103.

[3] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. IEEE T PAMI,

2000,22(12):1424-1445.

[4] Tian Y, Kanade T, Cohn J F. Recognizing lower face action units for facial expression analysis. Automatic face and gesture recognition, 2000. proceedings. fourth ieee international conference on. IEEE, 2000:484-490.

[5] Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010:94-101.

[6] Pantic M, Valstar M, Rademaker R, et al. Web-based database for facial expression analysis. 2005 IEEE international conference on multimedia and Expo. IEEE, 2005:5.

[7] Mavadati S M, Mahoor M H, Bartlett K, et al. Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 2013,4(2):151-160..

[8] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence, 2007,29(6):915-928.

[9] Klaser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3d-gradients. BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008:275:1-10.

[10] Hasani B, Mahoor M H. Facial expression recognition using enhanced deep 3D convolutional neural networks. Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017:2278-2288.

[11] Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks. Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016:1-10.

[12] Shan C, Gong S, McOwan P W. Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing, 2009,27(6):803-816.

[13] Wang Z, Ying Z. Facial expression recognition based on local phase quantization and sparse representation. Natural Computation (ICNC), 2012 Eighth International Conference on. IEEE, 2012:222-225.

[14] Cootes T F, Edwards G J, Taylor C J. Active appearance models. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001(6):681-685.

[15] Yeasin M, Bullot B, Sharma R. Recognition of facial expressions and measurement of levels of interest from video. IEEE Transactions on Multimedia, 2006,8(3):500-508.

[16] Jain S, Hu C, Aggarwal J K. Facial expression recognition with temporal modeling of shapes. Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. IEEE, 2011:1642-1649.

[17] Hasani B, Mahoor M H. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. IEEE, 2017:790-795.

[18] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance. European conference on computer vision. Springer, Berlin, Heidelberg, 2006:428-441.

[19] Andrearczyk V, Whelan P F. Convolutional neural network on three orthogonal planes for dynamic texture classification. Pattern Recognition, 2018,76:36-49.

[20] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:1-9.

[21] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:2625-2634.

[22] Molchanov P, Yang X, Gupta S, et al. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4207-4215.

[23] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems. 2014:568-576.

[24] Wu Z, Jiang Y G, Wang X, et al. Fusing multi-stream deep networks for video classification. arXiv preprint arXiv:1509.06086,2015.

[25] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012:1097-1105.

[26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:770-778.

[27] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI. 2017,4:12.

[28] Allaert B, Bilasco I M, Djeraba C. Advanced local motion patterns for macro and micro facial expression recognition. arXiv preprint arXiv:1805.01951,2018.

[29] Lee S H, Plataniotis K N K, Ro Y M. Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. IEEE Transactions on Affective Computing, 2014,5(3):340-351.

[30] Taheri S, Qiu Q, Chellappa R. Structure-preserving sparse decomposition for facial expression analysis. IEEE Transactions on Image Processing, 2014, 23(8):3590-3603.

[31] Liu M, Li S, Shan S, et al. AU-aware Deep Networks for facial expression recognition. FG. 2013:1-6.

[32] Liu M, Li S, Shan S, et al. Deeply learning deformable facial action parts model for dynamic expression analysis. Asian conference on computer vision. Springer, Cham, 2014:143-157.

[33] Zhang X, Mahoor M H, Mavadati S M. Facial expression recognition using lp-norm MKL multiclass-SVM. Machine Vision and Applications, 2015,26(4):467-483.

[34] Cootes T F, Taylor C J, Cooper D H, et al. Active shape models-their training and application. Computer vision and image understanding, 1995,61(1):38-59.