Learning Affective Video Features for Facial Expression Recognition

SREEKANTH PAI
S7 CSB 30

GUIDE : SARITH DIVAKAR M
ASSISTANT PROFESSOR,CSE

CONTENTS

INTRODUCTION

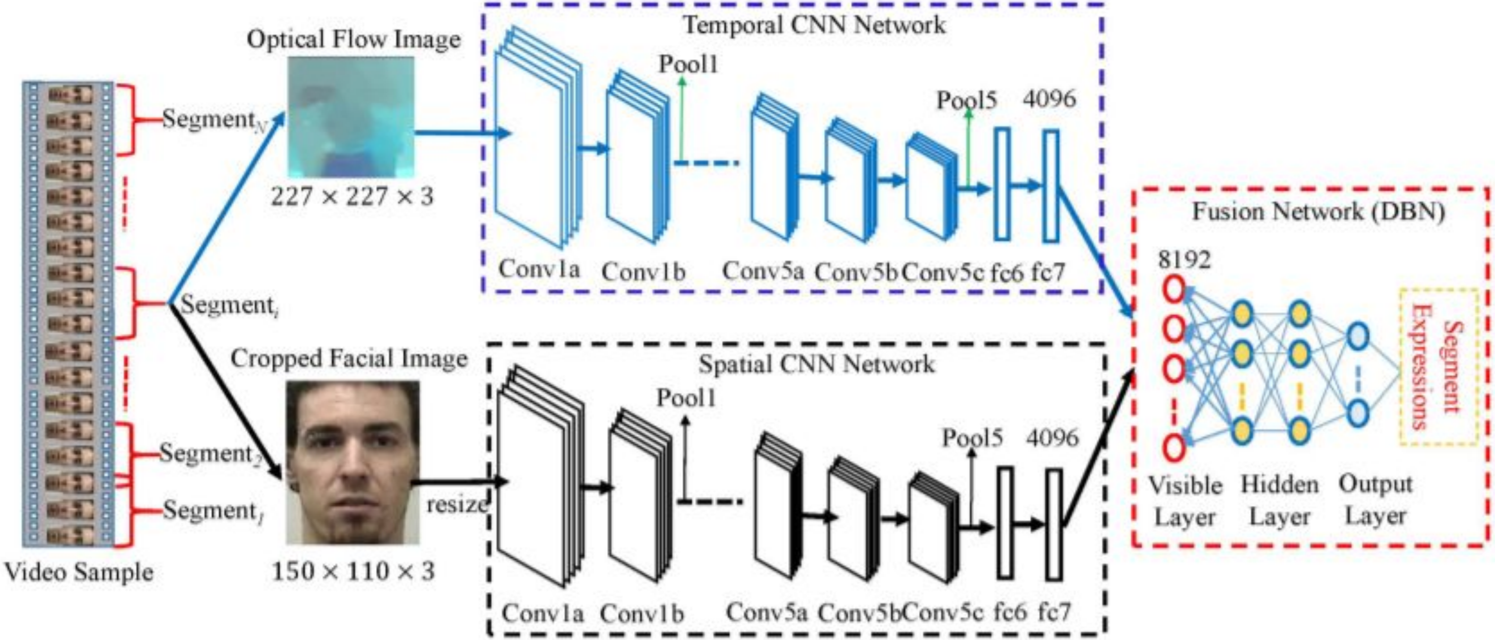
- ➔ Facial expression is one of the most natural nonverbal ways for expressing human emotions and intentions.
- ➔ FER has many potential applications such as human emotion perception, social robotics, human-computer interaction and healthcare.

RELATED WORKS

- ➔ ● HAND-DESIGNED FEATURE-BASED METHOD
- ➔ ● DEEP LEARNING-BASED METHOD

PROPOSED SYSTEM

- ➔ A hybrid deep learning model, comprising a spatial CNN network, a temporal CNN network and a deep fusion network built with a DBN model, to apply for FER in video sequences.
- ➔ To deeply fuse the spatial CNN features and temporal CNN features, we employ a deep DBN model as a deep fusion network to learn a joint discriminative spatio-temporal segment-level feature representation for FER.


Video Sample
Segment$_N$
Segment$_i$
Segment$_2$
Segment$_1$
Optical Flow Image
227 × 227 × 3
Cropped Facial Image
150 × 110 × 3
resize
Temporal CNN Network
Pool1
Pool5
4096
Conv1a
Conv1b
Conv5a
Conv5b
Conv5c
fc6
fc7
Spatial CNN Network
Pool1
Pool5
4096
Conv1a
Conv1b
Conv5a
Conv5b
Conv5c
fc6
fc7
Fusion Network (DBN)
8192
Visible Layer
Hidden Layer
Output Layer
Segment Expressions

METHODOLOGY

A. GENERATION OF CNN INPUTS

- ➔ Video sample with different durations are divided into a certain number of fixed-length segments as inputs of CNNs
- ➔ The division augments the amount of training data to some extent.

METHODOLOGY(Contd)

INPUTS OF TEMPORAL CNNs

- Inputs for temporal CNNs are generated by extracting optical flow images between consecutive frames in a video sequence.
- The values of motion field dx , dy are transformed into the interval [0, 255] by

$$\tilde{d}_{x|y} = a d_{x|y} + b, \quad \text{where } a = 1, b = 128$$

- The transformed flow maps are conserved as an optical flow image containing three channels, which corresponds to motion $\tilde{d}_x$, $\tilde{d}_y$ and the optical flow magnitude
- It produces an optical flow image with size of 227 × 227 × 3.

METHODOLOGY(Contd)

INPUTS OF SPATIAL CNNs

- ➔ A cropped facial image of 150 × 110 × 3 for each frame in a video segment, as in [23].
- ➔ Robust real-time face detector [38] is firstly leveraged to perform face detection to crop a facial image from each frame in a video segment.
- ➔ A cropped image of 150 × 110 × 3 containing facial key parts, such as head, nose, mouth, etc., is obtained from a facial image
- ➔ Cropped facial image is resized into 227 × 227 × 3 as inputs of spatial CNNs

B. SPATIO-TEMPORAL FEATURE LEARNING WITH CNNs

- ➔ The spatial and temporal CNNs have the same structure as the original VGG16 [16]
- ➔ To realize the task of spatio-temporal feature learning with CNNs,the pre-trained VGG16 is fine tuned on target video-based facial expression data
- ➔ Existing VGG16 parameters are copied to pre-train on a large scale ImageNet data to initialize the temporal CNN network and the spatial CNN network
- ➔ Then, we replace the fc8 layer in VGG16 with a new class label vector corresponding to six facial expression categories used in our experiments.

B. SPATIO-TEMPORAL FEATURE LEARNING WITH CNNs(Contd.)

- ➔ The two CNNs are retrained indiviually using standard back propagation strategy.
- ➔ Back propagation technique to solve the following minimizing problem so as to update the CNN network parameters:

$$\min_{W,\theta} \sum_{i=1}^{N} H(\text{softmax}(W \cdot \Upsilon(a_i; \vartheta)), y_i),$$

C. SPATIO-TEMPORAL FUSION WITH DBNs

- ➔ 4096-D outputs(Temporal and Spatial) of their fc7 layers are directly concatenated into a total 8192-D vector as inputs of the fusion network built with a deep DBN mode.
- ➔ This deep DBN model is used to capture highly non-linear relationships across spatial and temporal modalities, and produce a joint discriminative feature representation for FER.

C. SPATIO-TEMPORAL FUSION WITH DBNs(Contd.)

Two-step strategy to train the DBN fusion network:

- ➔ An unsupervised pre-training is conducted in the bottom-up way by means of a greedy layer-wise training algorithm.

$$1w = \varepsilon(< vihj >data - < vihj >model)$$

- ➔ A supervised fine-tuning is performed to update the network parameters with back propagation.

$$L(x, x') = \|x - x'\|_2^2 ,$$

D. VIDEO-BASED EXPRESSION CLASSIFICATION

- ➔ The output of DBMs last hidden layer represents the jointly learned discriminative spatio-temporal feature representations in video segments.
- ➔ Average-pooling approach is applied on all divided segments in a video sample to produce a fixed-length global video feature representation for FER
- ➔ Linear SVM classifier is adopted to perform the final FER tasks in video sequences.

EXPERIMENTATION

FER experiments are performed on three public video-based facial expression datasets, i.e., the BAUM-1s database , the RML database and the MMI database .

A. DATASETS

1) BAUM-1s

- ➔ The BAUM-1 database contains not only the six basic facial expressions (joy, anger, sadness, disgust, fear, surprise) and four mental states(unsure, thinking, concentrating, bothered)
- ➔ It comprises of 1222 video samples collected from 31 Turkish persons. Each video frame is 720x576x3

2) RML

- ➔ This database has the six basic facial expressions (angry, disgust, fear, joy, sadness and surprise)

- ➔ The RML database [21] consists of 720 video samples collected from 8 persons. Each video frame is 720×480×3

3) MMI

- ➔ The MMI database consists of 2894 video samples.
- ➔ 213 sequences have been labeled with six basic expressions from 30 subjects aging from 19 to 62

- ➔ RML database with less than 10 subjects, Leave-One-SubjectOut (LOSO) is used for experiments.All experiments adopted by subject-independent cross-validation strategy.
- ➔ BAUM-1s and MMI database with more than 10 subjects, Leave-One-Subject-Group-Out (LOSGO) with five subject groups is employed.
- ➔ RML database with less than 10 subjects, Leave-One-SubjectOut (LOSO) is used for experiments.

- Deep models are trained on the divided video segments so that the number of training data can be augmented.
- BAUM-1s database produce about 7000 segments from 521 video samples.
- RML database produce about 12, 000 segments are from 720 video samples.
- MMI database produce 4000 segments are from 213 video samples.

RESULTS AND ANALYSIS

- ➔ Performance of three different DBNs, including DBN-1 , DBN-2 and DBN-3 are verified.

DBN structure	BAUM-1s	RML	MMI
DBN-1	48.15	68.86	66.82
DBN-2	52.73	71.52	69.88
DBN-3	55.85	73.73	71.43

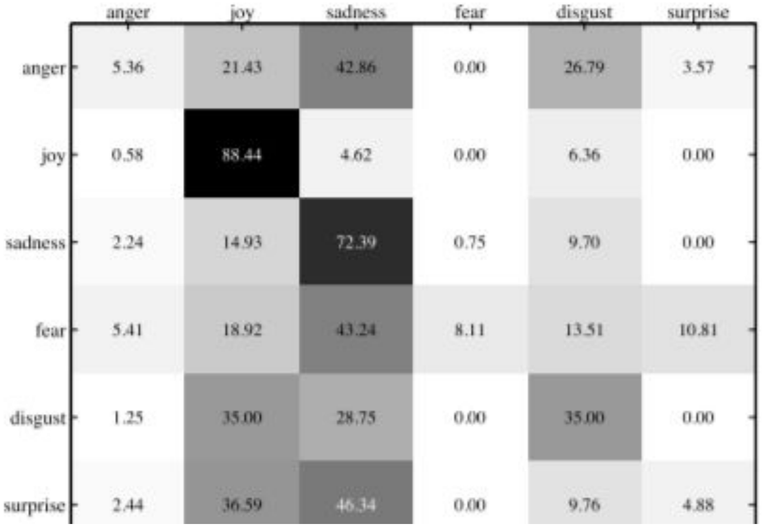
- ➔ In fusion network DBN-3 is adopted as the default structure of DBN for its best performance.

RESULTS AND ANALYSIS(Contd.)

- The spatiotemporal CNN+DBN features, which fuse spatio-temporal CNN features with DBNs, outperform the other two features.
- This indicates the effectiveness of fusing spatio-temporal features by using a deep DBN.
- DBNs are able to effectively discover the distribution properties of input spatio-temporal data, and learn the hierarchical feature representations of input spatio-temporal data.

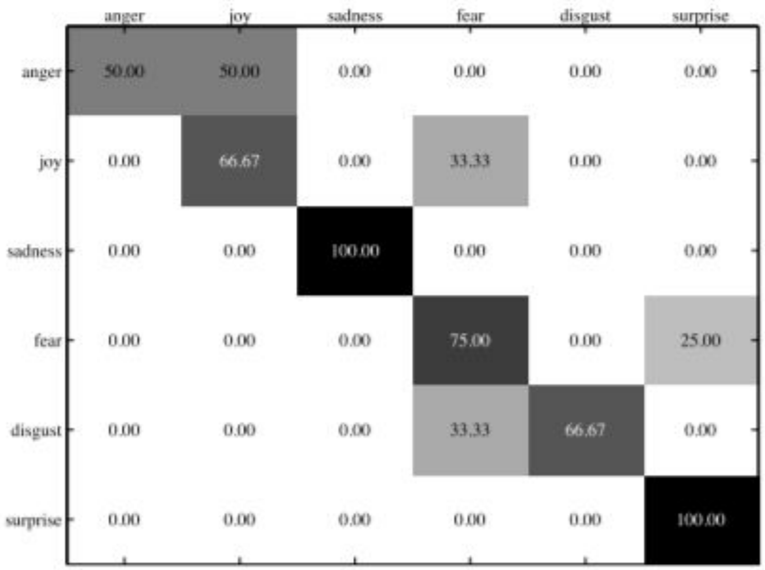
Features	BAUM-1s	RML	MMI
Spatial CNN	50.96	64.58	60.45
Temporal CNN	49.14	50.31	48.66
Score-level fusion	53.04	71.94	68.35
DBN fusion	55.85	73.73	71.43

➔ BAUM-1s dataset classifies "joy" and "sadness" with an accuracy of 88.44% and 72.39%, respectively, whereas other four facial expressions are identified badly with an accuracy of less than 35%.



Confusion matrix of recognition results with DBNs on the BAUM-1s dataset.

➔ MMI dataset classifies “sadness” and “surprise” with an accuracy of 100%, whereas the other expressions are identified with an accuracy of less than 75%.



Confusion matrix of recognition results with DBNs on the MMI dataset.

➔ RML dataset recognizes “disgust”, “sadness” and “surprise” well with an accuracy of more than 84%, whereas the remaining three facial expressions are distinguished with an accuracy of less than 80%.

	anger	disgust	fear	joy	sadness	surprise
anger	76.67	1.67	0.83	1.67	7.50	11.67
disgust	6.67	84.17	4.17	1.67	3.33	0.00
fear	1.67	2.50	72.50	16.67	6.67	0.00
joy	6.67	7.50	13.33	52.50	12.50	7.50
sadness	0.00	4.17	3.33	7.50	84.17	0.83
surprise	4.17	0.83	0.00	4.17	0.00	90.83

Confusion matrix of recognition results with DBNs on the RML dataset.

- ➔ Comparison with previous works on these three datasets. It is noted that these comparing works also employs subject-independent test-runs.
- ➔ Proposed method significantly outperforms the state-of-the-arts on these three datasets.

Datasets	Refs.	Features	Accuracy
BAUM-1s	S Zhalehpour[20]	LPQ	45.04
	Shiqing Zhang[15]	3D-CNN	50.11
	Ours	Spatio-temporal CNN+DBN	**55.85**
RML	NED Elmadany [42]	Gabor	64.58
	Shiqing Zhang[15]	3D-CNN	68.09
	Ours	Spatio-temporal CNN+DBN	**73.73**
MMI	M. Liu [14]	3DCNN-DAP	63.40
	B. Hasani [41]	Inception-ResNet	68.51
	Ours	Spatio-temporal CNN+DBN	**71.43**

Performance (%) comparisons of the-state-of-the-arts on the used three datasets.

CONCLUSION

- ➔ FER aims to analyze and understand human facial behavior, has become an increasingly active research topic in the domains of computer vision, artificial intelligence, pattern recognition, etc.
- ➔ FER has many potential applications such as human emotion perception, social robotics, humancomputer interaction and healthcare
- ➔ This paper proposes a hybrid deep learning model, which consists of the spatial CNN network, the temporal CNN network, and the DBN fusion network, to apply for FER in video sequences.
- ➔ Experiment results on three public video-based facial expression datasets, i.e., BAUM-1s RML, and MMI, demonstrate the advantages of our proposed method

REFERENCES

- K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ in Proc. ICLR, San Diego, CA, USA, 2015, pp. 1–14
- G. Gkioxari and J. Malik, ‘‘Finding action tubes,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 759–768.
- G. E. Hinton, ‘‘Training products of experts by minimizing contrastive divergence,’’ Neural Comput., vol. 14, no. 8, pp. 1771–1800, 2002.

THANK YOU