# Video-based emotion recognition using CNN-RNN and C3D hybrid networks

**Conference Paper** · November 2016

DOI: 10.1145/2993148.2997632



**4 authors:**

Yin Fan

**6** PUBLICATIONS **173** CITATIONS

SEE PROFILE

Xiangju Lu

**8** PUBLICATIONS **177** CITATIONS

SEE PROFILE

Dian Li

iqiyi.com

**1** PUBLICATION **156** CITATIONS

SEE PROFILE

Yuanliu Liu

iQiyi

**15** PUBLICATIONS **209** CITATIONS

SEE PROFILE

# Video-Based Emotion Recognition using CNN-RNN and C3D Hybrid Networks

Yin Fan, Xiangju Lu, Dian Li, Yuanliu Liu
iQIYI Co. Ltd,
Beijing, 10080, China
{fanyin, luxiangju, lidian, liuyuanliu}@qiyi.com

## ABSTRACT
In this paper, we present a video-based emotion recognition system submitted to the EmotiW 2016 Challenge. The core module of this system is a hybrid network that combines recurrent neural network (RNN) and 3D convolutional networks (C3D) in a late-fusion fashion. RNN and C3D encode appearance and motion information in different ways. Specifically, RNN takes appearance features extracted by convolutional neural network (CNN) over individual video frames as input and encodes motion later, while C3D models appearance and motion of video simultaneously. Combined with an audio module, our system achieved a recognition accuracy of 59.02% without using any additional emotion-labeled video clips in training set, compared to 53.8% of the winner of EmotiW 2015. Extensive experiments show that combining RNN and C3D together can improve video-based emotion recognition noticeably.

## Categories and Subject Descriptors
I.5.4 [**Pattern Recognition**]: Applications—*computer vision*, *signal processing*; I.4.m [**Image Processing and Computer Vision**]: Miscellaneous

## General Terms
Algorithms, Experimentation

## Keywords
Emotion Recognition; Long Short Term Memory network; 3D convolutional Network; Model Fusion

## 1. INTRODUCTION
With great progress in intelligent systems in recent years, emotion recognition remains the most important problem for human interaction. Emotion recognition is challenging due to the difficulties for definition and classification of emotion expressions for different people without contextual or psychological information. Nevertheless, many researchers have tried to identify emotions in videos based on computer vision technologies [3, 5, 8, 23].

The EmotiW challenge has been successfully held for four years since 2013 and has made great influences in the emotion recognition area. Previous winners usually focus on facial graph analysis [3, 23] or designing specific CNN-RNN networks [5]. Such classifier that takes video sequences as input is becoming more and more important for video event detection or prediction.

Traditional convolutional neural networks have a major limitation that they just handle spatial information. For example, in the EmotiW 2014 winner's work [4], all video frames are extracted from videos and regarded as the static images for further process. The aggregating image features of each video form a feature vector which neglects the important temporal video information.

Recently, deep 3-dimensional convolutional networks (C3D) made great progresses in dealing with various video analysis tasks. C3D can model appearance and motion information simultaneously and the C3D features with a linear classifier can achieve good performance on different video analysis benchmarks [6].

As for task of video-based emotion recognition, few work based on C3D structures is given in existing papers. We first take experiments to classify emotions based on C3D structures and we find that such C3D networks have a good performance in emotion recognition. The promising results inspire us to explore further experiments, including the experiments of hybridizing to LSTM models. As a result, such hybrid networks give the competitive results for emotion classification. The main contribution of the paper is the proposed hybrid CNN-RNN and C3D network[1].

A particular type of recurrent neural networks, the Long Short-Term Memory (LSTM) recurrent neural network is widely adopted [4, 5, 8]. LSTM has memory ability and suits for processing sequences with contexts well. An encoder LSTM can be used to map an input sequence into a fixed length vector representation. This vector representation is decoded using single decoder LSTM or multiple decoder LSTMs to handle different tasks such as emotion classification. Recently, LSTMs are also well developed and shown to be efficient to deal with various sequences versus sequences problems, such as audio analysis [26, 29], video captioning [24, 25], video action recognition [22], and so on.

Unlike C3D networks, a few works are given for video-based emotion recognition using CNN or RNN structures in recent papers [4, 5, 19]. Such deep networks reach top competitive


*ICMI '16,* November 12-16, 2016, Tokyo, Japan
 ISBN 978-1-4503-4556-9/16/11…$15.00
DOI: http://dx.doi.org/10.1145/2993148.2997632

[1]The source code is available at: https://github.com/lidian007/EmotiW2016

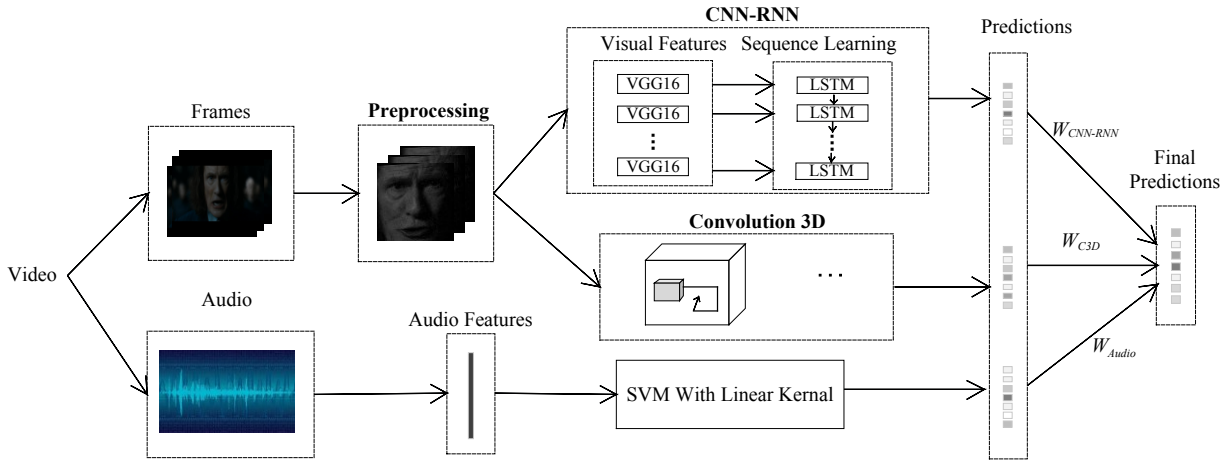



**Figure 1. The overview of the system**

results in the history of EmotiW challenges [5, 27]. Therefore we take similar LSTM models in our work while applying C3D structures at the same time. Without using any additional videos with emotion labels in our training set, the hybrid approach outperforms the state-of-the-art results by a significant margin. We achieved a recognition accuracy of 59.02%, compared to an accuracy of 53.8% as the winner of EmotiW 2015.

## 2. THE PROPOSED METHOD

The overview of the system is shown in Figure 1. The hybrid network has two core parts: CNN-RNN and C3D. We also add an audio classifier into the system. In this section we first introduce LSTM, which is a CNN-features-based spatio-temporal RNN model. Then we describe the properties of C3D structure, which can be directly used as a spatio-temporal model. At last, we give the details of building a hybrid classifier on LSTMs and C3Ds.

### 2.1 LSTM – A Special RNN

A general Recurrent Neural Network (RNN) can model temporal information by transforming a sequence of inputs to a sequence of outputs. Given an input sequence $(x_1, x_2 ..., x_n)$, a RNN computes the output sequence $(z_1, z_2 ..., z_n)$ via the following equations [5, 8]:

$$h_t = g(W_{xh} x_t + W_{hh} h_{t-1} + b_h)$$

$$z_t = g(W_{hz} h_t + b_z),$$

where $g$ is the hidden activation function, such as a sigmoid or hyperbolic tangent, and $h_t$ is the hidden state.

Although RNNs have been widely used in many tasks such as handwriting recognition [17] or speech recognition [18], they have difficulties in learning long-term dependencies due to the vanishing and exploding gradient problem. A Long Short Term Memory (LSTM) network is a special kind of RNN which is capable of addressing this long-term dependencies.

The core of LSTM networks is that it can remember a value for an arbitrary length of time. A LSTM unit is equipped with gates that determine when the input is significant enough to remember, when the unit should continue to remember or forget the value, and when the unit should output the value. We use LSTM unit described in [8] in our method (Figure 2), which iterates between the following operations:

$$i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + b_f)$$

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o)$$

$$g_t = \phi(W_{xc} x_t + W_{hc} h_{t-1} + b_c)$$

$$c_t = f_t \odot c_{t-1} + i \odot g_t$$

$$h_t = o_t \odot \phi(c_t),$$

where $\sigma$ is the sigmoid function, and $\phi$ is the hyperbolic tangent function.

We use a similar framework with the one described in [8], which combines LSTMs with deep convolutional networks to train a model spatially and temporally for video sequences.

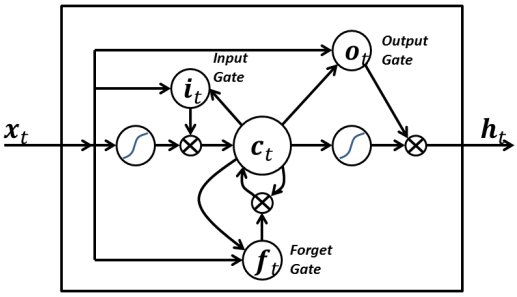



**Figure 2. A simple LSTM block with only input, output, and forget gates**

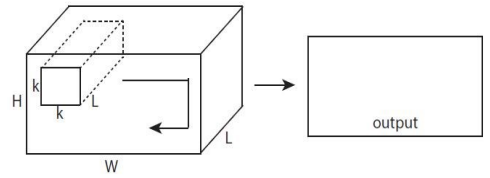
### 2.2 C3D – A Direct Spatio-Temporal Model

C3D can be understood as a 3D convolution on three channels. In such way, the results of C3D transform can be used as features for many tasks.

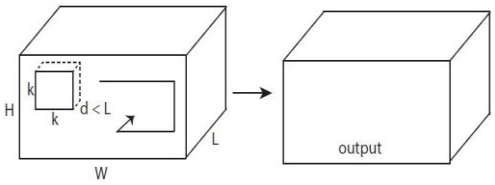
In our work, unlike CNN-RNN framework, C3D networks have a useful attribute that models both appearance and motion simultaneously.

In 2D ConvNets, convolution and pooling operations are only

spatially applied to 2D static images. While in our 3D ConvNets, the operations are performed spatio-temporally by adding an additional time dimension [6]. Hence such C3D networks preserve the temporal information of the input signals, resulting in a more distinctive result. The 2D and 3D convolutional frameworks are shown in Figure 3.




(a) 2D convolution on multiple frames




(b) 3D convolution on multiple frames

**Figure 3. 2D and 3D convolutions.** The input size of the frames is {H*W*L}. For the convolution layers, the kernel size is {K*K}, and the kernel temporal depth is d.

## 2.3 Hybrid CNN-RNN and C3D Networks

Previous work shows that either CNN-RNN or C3D model alone can achieve good performance in action recognition [5, 6, 8]. And we found that the CNN-RNN and C3D hybrid network can further improve the performance.

Considering that audio modality can bring ~3% gain in recognition accuracy [3], we also trained a SVM with the linear kernel using audio features extracted with the OpenSmile toolkit [7].

The CNN-RNN, C3D and audio SVM model were trained separately and their predication scores were combined into the final score. For each category, the prediction score obtained from different models are fused by a weighted summation. We select the parameters $W_{CNN-RNN}$, $W_{C3D}$ and $W_{Audio}$ according to the performance on the validation set. The category with the highest score is taken to be the final recognition result.

# 3. EXPERIMENTS

We participate in the EmotiW 2016 challenge and mainly evaluate the proposed method on the AFEW 6.0 database [1, 27].

In this section, the EmotiW 2016 challenge is briefly introduced and the implementation details of our method are given. Then the performance of our method is compared with the state-of-the-art. And we also test the sub-modules of our system, including different CNN features as inputs of LSTMs, and different parameters of LSTMs.

## 3.1 The EmotiW 2016 Challenge

The fourth Emotion Recognition in the Wild (EmotiW) 2016 challenge [27] has two sub-challenges: (1) video based emotion recognition on the AFEW 6.0 database [1]; (2) group level emotion recognition on the HAPPEI database [2, 27].

We only participate in the video-based emotion recognition sub-challenge. For this dataset, each video is labeled with one of seven emotions, that is, anger, disgust, fear, happiness, sad, surprise, and neutral. The task of the challenge is to assign an emotional label to each video in the test set. The challenge succeeds EmotiW 2013-15, and the major change in this year is the newly added reality TV videos in the test set.

In total, the AFEW 6.0 database contains 1750 video clips which are divided into three parts: 774 for training, 383 for validation, and 593 for test.

## 3.2 Implementation Details

As for data pre-processing, all faces of video frames are extracted and aligned using a similarity transform according to the facial key points. Additionally, for non-faces frames falsely detected by the face detector, a CNN-based face filter is applied for filtering out the mistakes.

For the CNN-RNN network used in the paper, the CNN features of the faces are taken from fc6 layer of VGG16-Face model fine-tuned with FER2013 face emotion database [13, 16]. For each iteration of RNN training, 16 face features are randomly selected as inputs. One layer LSTM network is applied with 128 embedding outputs.

For the C3D network, a series of a length of 16 sequent faces for each video clip is chosen as the inputs. The C3D net has 8 convolutions, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. Other parameters are similar to [6]. The specific C3D structure used in our implementation is shown in Figure 4.

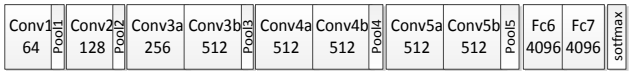



**Figure 4. C3D architecture.**

In the test stage, a series of 16 sequences is taken for each video by a stride of 8 frames. If a video clip doesn't have enough frames, we pad the last frame to get only one segment. The average score of all series of the video is the final reference of the prediction label.

## 3.3 Results of Submissions

The challenge allows 8 times submissions in total. For all the submissions, we use one audio model all the same as stated in section 2.3. And a different number of CNN-RNN models and C3D models are merged together. The results of our submissions are shown in Table 1.

**Table 1. Submissions on the validation and testing sets**

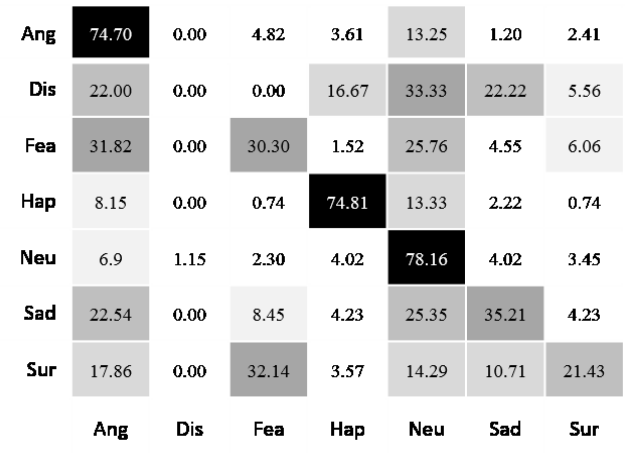
| Sub | Val | Test | Models 1Audio+ | TrainingData |
|---|---|---|---|---|
| 1 | 51.96 | 56.16 | 2 CNN-RNNs + 1 C3D | TrainingSet Only |
| 4 | - | 58.01 | 2 CNN-RNNs + 2 C3Ds | Training+Validation |
| 5 | - | 58.85 | 1 CNN-RNN + 2 C3Ds | Training+Validation |
| 6 | - | 59.02 | 1 CNN-RNN + 3 C3Ds | Training+Validation |

From Table 1, we can find that the more C3D models are taken to merge, a higher accuracy can be obtained. Due to the


(a) A Single VGG-LSTM model on validation set


(b) A Single C3D model on validation set


(c) Submission 1 on validation set




(d) Submission 6 on testing set

**Figure 5. Confusion matrices on AFEW6.0 dataset.**

computational complexities and the limited time for training, only a maximal number of three C3D models are tested. Interestingly, things are different for CNN-RNN network models. There is almost no difference in accuracy when taking one or more CNN-RNN models. The details of different CNN-RNN and C3D models are given in section 3.4 and 3.5.

Without audio information, a best single CNN-RNN model achieves an accuracy of 45.43%, while a single C3D can reach only 39.69%. An accuracy of 48.30% can be reached if the two models are fused. Also, we can find that the recognition accuracy is increased by adding more validation data in training phase. The final recognition accuracy is 59.02%, outperforming the baseline 40.47% and the 2015 winner team's final accuracy 53.80%. The confusion matrices of our submissions on the validation and testing sets are given in Figure 5.

## 3.4 CNN-RNN Classifier

### *3.4.1 Pre-processing of Video Frames*

If all the original video frames are taken as input directly, a poor distinctive ability is obtained for all kinds of emotions, with an average accuracy around 20%.

Optical flow is claimed to be a crucial component of video classification approaches. We additionally train our temporal models on optical flow images and perform late fusion akin to the two-stream hypothesis following [21, 22]. The performance of the system is around 25% for emotion classification, better than only using frames as inputs.

A face detector and face similarity transform is applied to pre-process video frames. Faces of video frames are detected by Viola-Jones cascades [28] and falsely detected non-face frames are filtered out by a face model trained based on CNN networks. Then face alignments are realized by a similarity transform based on facial point landmarks. The system takes aligned faces as inputs, give a competitive result. Also, optical flow of faces are tested following [21], but no improvements are obtained for emotion classification.

Note that, the visual frames that are inputs of the hybrid network refer to aligned faces that are already analyzed and extracted from original video frame sequences.

### *3.4.2 Pre-training model / CNN Feature Extractor*

As fine-tuning is widely adopted and proved to be very efficient to re-train a pre-training model using limited data for a special

task, we use the emotional data to fine-tune pre-training models for face CNN features. Here the Facial Expression Recognition database (FER2013) [16] was used for CNN pre-training.

The FER2013 database consists of 35889 images: 28709 for training, 3589 for public test, and 3589 for private test. We use training data and public test data for training, and evaluate the model performance with private test data.

To select a basis for the fine-tuned models, several mainstream CNN architectures trained with ImageNet are examined, including CaffeNet [10], GoogLeNet [11], VGG [13], and Residual Network [12], which perform well for general objects recognition. Deep-face VGG model (VGG-FACE) [14] is tested for comparisons with VGG model trained on ImageNet. The VGG-FACE model is trained on face images only, which is originally used for a face verification task. We also evaluate DCN network architecture described in [19] without any pre-training. The faces are resized to a fixed size of 128*171, which is the same as the size of the pre-trained ImageNet and VGG-FACE models. All the results of emotion fine-tuning performances from pre-training CNNs are shown in Table 2.

**Table 2. Testing Accuracies for All CNN architectures**

| | **Pre-Trained** | **Accuracy** |
|---|---|---|
| DCN | None | 64.29 |
| GoogLeNet | With ImageNet | 62.96 |
| CaffeNet | With ImageNet | 68.05 |
| VGG16 | With ImageNet | 68.24 |
| Residual Network | With ImageNet | 69.65 |
| **VGG16-FACE** | **With Faces** | **70.74** |

In Table 2, it's surprising to see that GoogLeNet works worse than all other network architectures. Probably the inception style network is not suitable for emotion recognition. Most of the models pre-trained using ImageNet database outperform the model without any pre-training due to the per-trained models providing a good initialization. Meanwhile, the model after pre-training using face database (VGG-FACE) achieves the best recognition accuracy.

### *3.4.3 LSTM Parameters*

In this subsection, CNN features of video faces based on the fine-tuned VGG16-FACE model are adopted. Such features are aggregated as LSTMs' inputs.

**Table 3. Validation Accuracies for Different CNN-RNNs**

| | | **Fc6** | **Fc7** |
|---|---|---|---|
| VGG-LSTM-OneLayer | **128** | **45.43** | 42.30 |
| | 256 | 44.39 | - |
| VGG-LSTM-Two-Layers | 256-256 | 41.26 | - |
| CaffeNet-LSTM-OneLayer | 128 | 40.99 | - |

We explore different LSTM parameter settings, including the different number of LSTM net's layers, different number of LSTMs' hidden units, and so on. The comparison on performance can be seen in Table 3.

From Table 3, we can find that with less number of embedded outputs, the LSTM net reaches a much better accuracy. That is, a simpler structure has a better classification ability. The reason may be that the training data is too limited.

## 3.5 C3D Network

Considering that in [3], the images with two face scales are used to further improve the performance, we adopt two kinds of face images, the face described in 3.4.1 (Face-ours) and the face provided in AFEW6.0 (Face-AFEW). We fine-tuned the pre-trained sport1m [20] C3D model with these two kinds of face images, respectively. The faces are resized to a fixed size of 128*171, which is the same as the size of the pre-trained sport1m C3D model.

Similar to LSTM classifier, the input of C3D was fixed to be a sequence of 16 frames. Table 4 shows the comparative results. We can find that the performances of C3D network are similar when different scales of faces are used for training, and the better accuracies can be reached when C3D and CNN-RNN models are merged. Moreover, when combining the models of two kinds of face images, the accuracy is further increased.

**Table 4. Accuracies on Validation for C3D Using Different Faces**

| | **Accuracy** |
|---|---|
| Face-AFEW(1 C3D) | 39.69 |
| Face-ours(1 C3D) | 38.97 |
| Face-ours(1 CNN-RNN + 1 C3D) | 46.49 |
| Face-AFEW(1 CNN-RNN + 1 C3D) | 42.82 |
| Face-ours + Face-AFEW(2 CNN-RNNs + 2 C3Ds) | 48.30 |

# 4. CONCLUSIONS

In this paper, we proposed a method for video-based emotion recognition in the wild. We used CNN-LSTM and C3D networks to simultaneously model video appearances and motions. Especially, we found that the combination of the two kinds of networks can give impressive results, which demonstrated the effectiveness of the method.

# 5. REFERENCES


[1] Dhall, A., Goecke, R., Lucey, S. and Gedeon, T. 2012. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia.

[2] Dhall, A., Goecke, R. and Gedeon, T. 2015. Automatic Group Happiness Intensity Analysis. *IEEE Transaction on Affective Computing*.

[3] Yao, A., Shao, J., Ma, N. and Chen,Y. 2015. Capturing AU-Aware Facial Features and Their Latent Relations for Emotion Recognition in the Wild. *ACM ICMI*.

[4] Liu, M., Wang, R., Li, S., Shan, S., Huang Z. and Chen, X.2014. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. *ACM ICMI*.

[5] Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal, C. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 467-474. ACM.

[6] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* .4489-4497. IEEE.

[7] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In*Proceedings of the 18th ACM international conference on Multimedia*. 1459-1462. ACM.

[8] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.2625-2634.

[9] Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç, Memisevic, R. and Mirza, M. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 543-550. ACM.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In ACM MM.

[11] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D. and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.1-9.

[12] He, K., Zhang, X., Ren, S. and Sun, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

[13] Simonyan, K., & Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[14] Parkhi, O. M., Vedaldi, A., & Zisserman, A. 2015. Deep face recognition. In *British Machine Vision Conference* (Vol. 1, No. 3, p. 6).

[15] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K. and Li, F.F., L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition. CVPR.* 248-255. IEEE.

[16] Carrier, P. L., Courville, A., Goodfellow, I. J., Mirza, M. and Bengio, Y. 2013 .*FER-2013 face database*. Technical report, 1365, Université de Montréal.

[17] Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H. and Schmidhuber, J. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, *31*(5), 855-868.

[18] Sak, H., Senior, A. W. and Beaufays, F. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*. 338-342.

[19] Kim, B. K., Dong, S. Y., Roh, J., Kim, G. and Lee, S. Y. 2016. Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach. In *Computer Vision and Pattern Recognition. CVPR.*

[20] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Li, F.F. 2014. Large-scale Video Classification with Convolutional Neural Networks.

[21] Ng, J., Hausknecht, M., Vijayanarasimhan S., Monga R., Vinyals O., Toderici G.2015. Beyond Short Snippets: Deep Networks for Video Classification. In *Computer Vision and Pattern Recognition. CVPR.* 4694-4702. IEEE.

[22] Sharma S., Kiros R., Salakhutdinov R.2016 Action Recognition using Visual Attention. Workshop track - ICLR.

[23] Kaya, H., Gürpinar, F., Afshar, S. and Salah, A. A. 2015. Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 459-466. ACM.

[24] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T.m. and Saenko, K. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*. 4534-4542.

[25] Pan, P., Xu, Z., Yang, Y., Wu, F. and Zhuang, Y. 2015. Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning. *arXiv preprint arXiv:1511.03476*.

[26] Graves, A., Mohamed, A. R. and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. 6645-6649. IEEE.

[27] Dhall, A., Goecke, R., Joshi, J., Hoey, J. and Gedeon, T. 2016. EmotiW 2016: Video and Group-level Emotion Recognition Challenges, ACM ICMI 2016.

[28] Jianguo L., Tao W., Yimin Z. 2011. ICCV: Face Detection using SURF Cascade. *In Computer Vision Workshops.*

[29] Fernández, S., Graves, A., Schmidhuber, J. 2007. An application of recurrent neural networks to discriminative keyword spotting. In *International Conference on Artificial Neural Networks*. 220-229. Springer Berlin Heidelberg