2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Video-based Emotion Recognition using Aggregated Features and Spatio-temporal Information

Jinchang Xu
Beijing University of Posts and Telecommunications
Beijing, China
Email: xjc1@bupt.edu.cn

Yuan Dong
Beijing University of Posts and Telecommunications
Beijing, China
Email: yuandong@bupt.edu.cn

Lilei Ma
Beijing University of Posts and Telecommunications
Beijing, China
Email: lileima@bupt.edu.cn

Hongliang Bai
Beijing Faceall Technology Co., Ltd
Beijing, China
Email: hongliang.bai@faceall.cn

***Abstract*—In this paper, we present a video-based emotion recognition system in the wild which consists of four pipeline modules: image-processing, deep feature extraction, feature aggregation and emotion classification. Our method focuses more on different feature descriptors. To obtain high-level features which are more discriminative in emotion recognition, we employ an aggregation of features extracted from different deep convolutional neural networks (CNNs). Furthermore, the long short-term memory network (LSTM) and 3D convolutional networks (C3D) are utilized to extract spatio-temporal features from videos in order to combine the spatial information and temporal information. Additionally, we evaluate our method on the 5th Emotion Recognition in the Wild Challenge in the category of video-based emotion recognition and the result shows our proposed system achieves better performance.**

I. Introduction

In the past several years, the automatic recognition of facial expressions has been an active research focus. Emotion recognition attracts more and more attention in computer vision due to its important role in many fields, such as human-computer interaction and psychological research [1], [2]. There are many studies based on emotion recognition technologies like detecting sadness in individuals [3], lie detection mechanisms [4], monitoring fatigue drive [5] and diagnosis of developmental disorders of children [6]. However, emotion recognition is challenging due to its difficulties in definition and classification of emotion expressions without contextual or psychological information. Thus, many researchers have done lots of work to recognize emotions in videos using computer vision technologies [7]–[11].

Traditional approaches for emotion recognition are based on hand-engineered features [12]. There is a rich literature on hand-crafted features extracted from images and videos for encoding facial emotions. For instance, Shan et al. [13] evaluate facial representation based on statistical local features, for person-independent facial expression recognition. Zhao et al. [14] identify facial expressions by use of local binary patterns on three orthogonal planes (LBP-TOP). The authors in [15] combine multiple visual descriptors with paralinguistic audio features for multimodal classification of video clips. In [16], Wu et al. adopt bag of features model to encode video clips.

Furthermore, with the rapid development of deep learning on image processing, a large number of techniques based on convolutional neural networks (CNNs) have been successfully applied to computer vision and become state-of-the-art for many vision tasks including image classification [17]–[19], image segmentation [20], object detection [21]–[23], and face recognition [24], [25]. Convolutional neural networks and deep features have also been used in emotion recognition [26]–[28]. The basic principle of deep learning is to learn hierarchical representations of input data such that the learned representations improve classification performance. Various pre-trained convolutional neural network models are utilized for extracting image features of the network's last few fully-connected layers, which performs well on transfer learning tasks [29]. However, such image-based deep features lack the time sequence information of videos. In [30], the authors propose a deep 3D convolutional network (3D ConvNet) to learn spatio-temporal features on videos datasets. Meanwhile, recurrent neural network architectures (RNNs) provide an attractive framework for propagating information over a sequence using a continuous-valued hidden layer representation [10]. And long short-term memory architectures (LSTM) have seen an explosion of recent interest as they yield high performance on a variety of sequence analysis tasks [31], [32]. As the task of emotion recognition from video is much more difficult than the general video recognition, the emotion recognition challenges such as the 2017 Emotion Recognition in the Wild Challenge (EmotiW2017) provide data for training and evaluating novel methods with the aim of providing a common benchmarking platform for researchers working on different aspects of affective computing [33].

In this work, we propose a video emotion recognition pipeline modules based on video modality only which consists

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

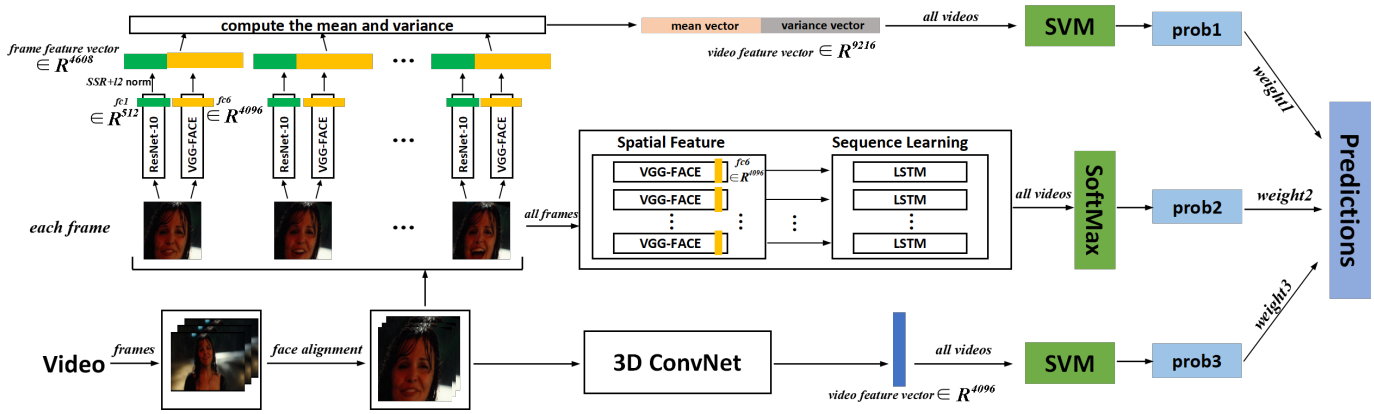

Fig. 1. The overview of our proposed pipeline modules. It consists of four parts: image-processing, feature extraction, feature aggregation, and emotion classification. We first extract all frames from each video. Then faces are detected and aligned. After that, we apply two different convolutional neural networks (CNNs) VGG-FACE and ResNet-10 to extract high-level emotional features with spatial information. Each of these features (4096-dimension fc6 of VGG16 and 512-dimension fc1 layer of ResNet10) is separately normalized using Signed Square Root (SSR) and L2 normalization. We compute the mean and variance of feature dimensions over all video frames and concatenate these features (9216-dimension) to represent a video sequence. Moreover, long short-term memory architectures (LSTM) based on VGG-FACE features and 3D convolutional network (C3D) are utilized to make full use of spatio-temporal features. Finally, we combine the predictions of all the classifiers with assigned weights to make the final predictions.

of four parts: image-processing, feature extraction, feature aggregation and emotion classification. This method focuses more on combining the features extracted from several different neural networks such as VGG and ResNet while making full use of spatial and temporal information from videos by employing CNN-LSTM and 3D ConvNet. Moreover, we present our solution to the EmotiW2017 audio-video emotion recognition challenge.

II. Proposed Method

Our emotion recognition system consists of an image-processing module, a feature extraction module, a feature aggregation module, and finally an emotion classification module. The deep features can describe semantic information of the input data by concatenating the different convolutional neural network features. Furthermore, in order to combine the spatial and temporal information, we apply the LSTM units and 3D convolutional networks. An overview of the pipeline is illustrated in Fig. 1.

A. Image Preprocessing

Firstly, we extract all the frames from each video. After that, multi-task CNN (MTCNN) [34] is used to detect face bounding boxes and landmarks. It returns a list of frames containing human faces and five landmark points including the position of eyes, nose, and mouth. Then the faces are aligned by affine transformation based on two-eye coordinates. Finally, the images are cropped and scaled to suit the input size of our pre-trained feature extraction networks.

B. Feature Extraction

As for the feature extraction module, we explore multiple networks in order to obtain high-level emotional features with spatial and temporal information. The details are listed as follows.

*1) **CNN Architectures**:* As CNNs can be trained to extract high-level facial texture features which are more discriminative in emotion recognition than other hand-crafted features, we use the pre-trained model mainly for feature extraction to classify static images containing emotions. With deep structures learning representations that better generalizing to other datasets, we choose the VGG model (VGG-FACE) [35] which is a 16-layer CNN architecture trained on celebrities face recognition tasks as one of our feature extraction network. Except for the VGG-like architectures, we also employ the residual networks. We observe that deeper architectures like ResNet-50 and ResNet-101 [36] get over-fitted much more easily for emotion recognition task. This might be due to the small amount of labeled data available for the emotion tasks. For this reason, we just use the ResNet-10 network [37] which only consists of 4 residual blocks. We add another fully connected layer 1 (fc1) between the global pool layer and the last fully connected layer and fine-tune all the layers on our face recognition data. The outputs of the fully connected layer 6 (fc6) from the VGG16 network and the fc1 from the ResNet10 network are selected as the deep features for every frame. These features extracted by CNNs are frame-level and represent spatial information.

*2) **3D ConvNet**:* The operations in 3D ConvNet (C3D) are 3D convolution and pooling which are performed spatio-temporally by adding an additional time dimension compared to 2D ConvNet [27]. Applying 3D convolution on a video clip, we can preserve the time series information of the input frames. Thus, it has an advantage on spatial and temporal feature learning. We adopt the C3D network proposed in [30]. The 3D ConvNet has 8 convolutional layers, 5 pooling layers, two fully connected layers and a SoftMax output layer. We modify the last fully connected layer to suit our emotion recognition task. The output of the last fully connected layer which is a 4096-dimension vector is considered as the deep

feature for per video.

*3) **LSTM**:* Though recurrent neural networks (RNNs) can connect spatial information with temporal information by transforming a sequence of inputs to a sequence of outputs, they have difficulties in learning long-term dependencies due to the vanishing and exploding gradient problem. However, the LSTM network is capable of dealing with these problems and can remember the value for an arbitrary length of time. It is more complex, but easier to train. Different from the traditional LSTM cell, we adopt the LSTM units with bidirectional and residual structures [1], [38]. In this architecture, we take the fc6 layer output of the fine-tuned VGG-Face network [35], which is a 4096-dimension vector, as the extracted frame feature. And the features of randomly selected frames are sequentially traversed in a bidirectional LSTM so as to capture the temporal information and dynamic changes of facial textures.

C. Feature Aggregation

There are two types of feature representations: frame-level and video-level. The deep features extracted from our trained CNN models are on behalf of frames, therefore we encode these features into a feature vector which represents the entire video sequence. For each video frame, we normalize each of these features (4096-dimension fc6 of VGG16 and 512-dimension fc1 layer of ResNet10) separately using Signed Square Root (SSR) and L2 normalization. After that, we compute the mean and variance of feature dimensions over all video frames and concatenate these features to represent a video sequence. Since both the features of LSTM and C3D are video-level representations, we just use their features for classification.

D. Emotion Classification

The concatenating features from VGG-FACE and ResNet-10 are used to train a one-vs-rest multi-class linear Support Vector Machine (SVM) to classify each video with one of the 7 emotion classes. At test time, we compute the encoded features in the same way and obtain the SVM class predictions. For the C3D network, we also extract the features and input them to a multi-class linear SVM for training models. As for the LSTM network, we use the last time-step output of SoftMax as the video prediction. Finally, we employ a weighted sum of the class probabilities estimated by these classifiers to boost the performance of video-based emotion recognition.

III. Datasets and Implementation Details

A. Datasets

1) FER2013: The FER2013 dataset [39] consists of 35889 images with seven basic expressions: angry, disgust, fear, happy, sad, surprise and neural. All the images are split into three parts: 25889 for training, 5000 for validation, and 5000 for the test.

2) AFEW 7.0: The Acted Facial Expressions in the Wild (AFEW) 7.0 Dataset [33] is created by a simple sentiment analysis of the closed captions in movies and TV series. Being collected from the trimmed video clips of movies and TV series, they are more realistic and have more challenging conditions compared to videos of facial actions deliberately produced and captured in lab conditions [26]. The dataset is divided into three data partitions with seven basic expressions: Train (773 samples), Val (383 samples) and Test (653 samples).

B. Implementation Details

For VGG-FACE network, we crop the aligned faces to 128×128 pixels and then resize them to 224×224 pixels to suit the input image size of our network. As for ResNet-10 network, the input image size is set to 150×150, which is much smaller. We flip and rotate the input images in order to make data augmentation to avoid network over-fitting. The initial learning rate is set to 0.0001 and mini-batch size is set to 32 while the Stochastic Gradient Descent (SGD) is selected as our optimizing method. After extensive experiments on FER2013 dataset, we obtain the best VGG-Face model which can achieve a relatively high accuracy of 71.01% and the best ResNet-10 model with the accuracy of 70.27% on the test set while the model is not over-fitting or collapsing. After that, we fine-tuned these two models on the AFEW 7.0 dataset with the same image preprocessing. All the cropped face images from the same video clip are regarded as samples whose labels are the same with the video. When the single model can achieve a better accuracy, we extract and aggregate the features from the fc6 layer of VGG-FACE network and the fc1 layer of ResNet-10 network. Finally, we train a linear SVM based on these concatenated features.

For the C3D network, the input shape is like N×T×C×H×W, where N is mini-batch size, T is the time series, and C, H, W is denoted as the channel, height, and width of the image separately. A series of the length of 128 sequent faces for each video clip is chosen as the inputs. And the image size is set to 128×171 pixels. We fine-tune the C3D pre-trained model which were trained on Sports-1M dataset [40] using the AFEW 7.0 dataset [33]. The initial learning rate is set to 0.0001 while the optimizing method is SGD. After that, we extract the features from the last fully connected layer for each video and train a linear SVM classifier.

Similar to C3D network, the length of video frames are fixed to 128 for our LSTM network and the features extracted from fine-tuned VGG-FACE model are concatenated as the inputs. We set the number of hidden layers to 28 in the neutral network. The training batches are set to 32 and initial learning rate is set to 0.001. For the optimization of LSTM network, we use Adaptive Moment Estimation(Adam) method.

IV. Experiments

A. Results on FER2013

In order to boost performance on EmotiW video-based emotion recognition task, we fine-tune our networks on FER2013 as our pre-trained models are mainly aimed at face recognition tasks which is emotion invariant. To select best models, several CNN architectures have been tested, including ResNet50 [19], ResNet-10 [37], VGG [41] and deep-face VGG model (VGG-FACE) [35]. We use training data and validation data for training, and evaluate the model performance with test data. From Table I, VGG-FACE model and ResNet-10 model gain high performance than the others. This results also show that the pre-trained models based on face recognition provide a good initialization than the ImageNet pre-trained models.

TABLE I
The Accuracy of fine-tuned models of different basis network on FER2013 dataset.

Network	Pre-trained Dataset	Test Accuracy
VGG16	ImageNet	67.86
VGG19	ImageNet	68.14
ResNet50	ImageNet	69.53
ResNet10	Our face recognition data	70.27
VGG-FACE	Celebrities	71.01

TABLE II
The Accuracy of different approaches on validation set accuracy.

Method	Validation Accuracy
Challenge baseline [33]	38.81
VGG-FACE	38.86
ResNet-10	37.14
Feature Aggregation(VGG-FACE + ResNet-10)	40.15
C3D	41.53
LSTM	39.84
Our final method	48.01

B. Results on AFEW 7.0

Table II presents the emotion recognition accuracy of the baseline approach provided by the EmotiW2017 and our proposed methods. The baseline method uses Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) which is a standard texture-based feature. They align the faces into a 128×128 grid and each frame is divided spatially into non-overlapping 4×4 blocks. Non-linear Chi-square kernel-based SVM is trained for emotion classification using the LBP-TOP concatenated features computed on the aligned frames from each block. The video-only baseline system achieves 38.81% and 41.07% classification accuracy for the Val and Test sets, respectively [33].

From the results of Table II, it can be seen that our single VGG-FACE model is able to work well compared with the sophisticated LBP-TOP method. Moreover, it can be observed that concatenating features from different networks give better performance than either of a single model. By aggregating the features of VGG-FACE and ResNet-10 networks, we can

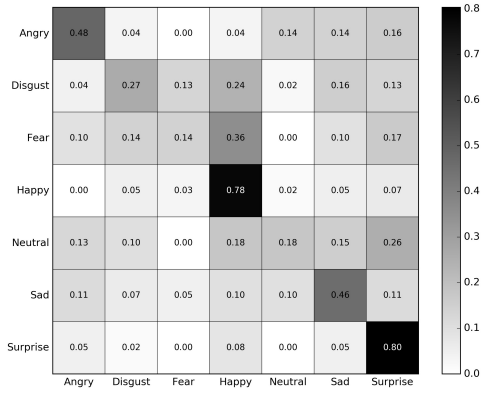

Fig. 2. The confusion matrix of results on the validation set.

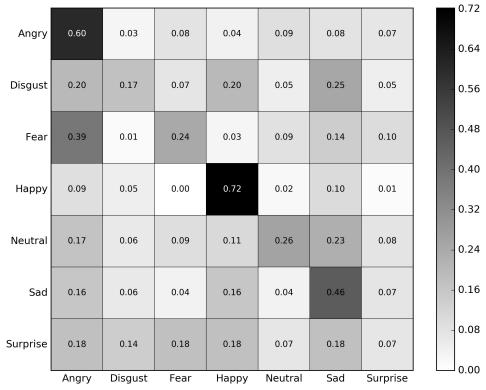

Fig. 3. The confusion matrix of results on the test set.

increase the accuracy to 40.15%. Furthermore, both the C3D network and LSTM network outperform better than the single CNN model, which proves that the spatio-temporal evolution of facial features is much more effective for emotion recognition. During the predicting phase, each model generates a score matrix which indicates the probability of every predicted sample belonging to the related emotions. According to the performance of each model, we assign a proper weight to each of them. The *weight1*, *weight2* and *weight3* is set to 0.5, 0.25 and 0.25 separately. With weighting all the scores, we can obtain an accuracy of 48.01%, which is obviously superior to the performance of each single model. Also, it surpasses baseline method by 9.2%.

TABLE III
Accuracy of the each class on validation and test set.

Emotion	Validation Accuracy	Test Accuracy
Anger	48.21	60.20
Disgust	26.67	17.50
Fear	14.28	24.29
Happy	77.97	72.22
Neutral	17.95	25.91
Sadness	45.90	46.25
Surprise	80.32	7.14
Average	48.01	42.27

From the confusion matrix in Fig. 2 and Fig. 3, it can

be seen that our classifiers perform well on happy, angry and surprise while the performance on fear and disgust is poor mostly due to lacking training samples. Furthermore, the accuracy of surprise class on the test is much lower than on validation. One of the explanations is that the distribution of validation and test samples are different. In the validation phase, the number of samples for each class is nearly equal according to the provided data while the sample number of each class becomes extremely unbalanced in test phase according to the confusion matrix in Fig. 3. Meanwhile, our method achieves a better performance on the validation than the test as the results shown in Table III. The main reason should be that the random search may assign high weights to the over-fitted models in the fusion stage which would result in poor generalization performance. Except that, the accuracy of each class on test set is almost consistent with that on validation set.

V. Conclusion

In this paper, we present a spatio-temporal video emotion recognition pipeline to deal with the challenging emotion recognition in the wild problem. Our algorithm is only based on the video modality. We aggregate the deep features from different networks and achieves a higher validation accuracy than any of the single classifier. Moreover, we combine the spatial information and temporal information from video clips by considering their complementarity in recognizing emotions. By a fusion of our models, we achieve a better performance with respect to baseline method. In the future, work should focus more on exploring different modalities such as audios and activities which may have an important role in recognizing the emotion from videos.

VI. Acknowledgment

This work is supported by Chinese National Natural Science Foundation under Grants 61532018 and Beijing Faceall Technology Co.,Ltd.

References

[1] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016. 1, 3

[2] R. Srinivasan, J. D. Golomb, and A. M. Martinez, "A neural basis of facial action recognition in humans," *Journal of Neuroscience*, vol. 36, no. 16, pp. 4434–4442, 2016. 1

[3] M. S. Clark, K. R. Von Culin, E. Clark-Polner, and E. P. Lemay Jr, "Accuracy and projection in perceptions of partners recent emotional experiences: Both minds matter." *Emotion*, vol. 17, no. 2, p. 196, 2017. 1

[4] M. Iwasaki and Y. Noguchi, "Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements," *Scientific reports*, vol. 6, p. 22049, 2016. 1

[5] M. Ali, F. Al Machot, A. H. Mosa, and K. Kyamakya, "Cnn based subject-independent driver emotion recognition system involving physiological signals for adas," in *Advanced Microsystems for Automotive Applications 2016*. Springer, 2016, pp. 125–138. 1

[6] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3414–3421. 1

[7] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 451–458. 1

[8] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570. 1

[9] J. Wu, Z. Lin, and H. Zha, "Multi-view common space learning for emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 464–471. 1

[10] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 467–474. 1

[11] R. Khan and O. Sharif, "A literature review on emotion recognition using various methods," *Global Journal of Computer Science and Technology*, 2017. 1

[12] S. E. Kahou, P. Froumenty, and C. Pal, "Facial expression analysis based on high dimensional binary features," in *European Conference on Computer Vision*. Springer, 2014, pp. 135–147. 1

[13] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009. 1

[14] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007. 1

[15] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 517–524. 1

[16] J. Wu, Z. Lin, and H. Zha, "Multiple models fusion for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 475–481. [Online]. Available: http://doi.acm.org/10.1145/2818346.2830582 1

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 1

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 1, 4

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440. 1

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. 1

[22] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy *et al.*, "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2403–2412. 1

[23] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448. 1

[24] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996. 1

[25] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015. 1

[26] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 433–436. 1, 3

[27] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450. 1, 2

[28] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, 2017. 1

[29] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," 2014. 1

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," pp. 4489–4497, 2014. 1, 2

[31] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014. 1

[32] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. 1

[33] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *The ACM International Conference*, 2017, pp. 524–528. 1, 3, 4

[34] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. 2

[35] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6. 2, 3, 4

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. 2

[37] M. Simon, E. Rodner, and J. Denzler, "Imagenet pre-trained models with batch normalization," *arXiv preprint arXiv:1612.01452*, 2016. 2, 4

[38] G. Chevalier, "Lstms for human activity recognition," https://github.com/guillaume-chevalier/LSTM-Human-Activity-Recognition, 2016. 3

[39] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124. 3

[40] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732. 3

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 4