Combining Sequential Geometry and Texture Features for Distinguishing Genuine and Deceptive Emotions

Liandong Li[1], Tadas Baltrusaitis[2], Bo Sun[1], Louis-Philippe Morency[2]
[1] College of Information Science & Technology, Beijing Normal University
[2] Language Technologies Institute, Carnegie Mellon University
bnulee@hotmail.com, tb346@cl.cam.ac.uk, tosunbo@bnu.edu.cn, morency@cs.cmu.edu

Abstract

In this paper, we explore a new type of automatic emotion recognition task - distinguishing genuine and deceptive emotions from video clips. For this task, it is not enough only using static images clipped from the video data, as there's only subtle differences between two types of emotions, which makes it even harder for automatic analysis. To utilize the temporal information, we introduce temporal attention gated model for this emotion recognition task. Compared to texture features which describe the whole face area, the facial landmark sequences may also indicate the temporal changes of the face, thus we utilize them by encoding feature sequence unsupervisedly.

1. Introduction

Recently, automatic emotion recognition has been a popular research area in computer vision community. Previous research mostly focuses on classifying six basic emotion categories [7], which is not a hard task for human beings when dealing with acted data. Meanwhile, being able to recognize deceit and the authenticity of emotional displays is notoriously difficult for human observers because of the subtlety or short duration of discriminative facial responses [20]. The applications are however numerous, from determining deceiving behavior in police investigations, to monitoring the mental status of patients.

Traditional tasks for emotion recognition benefits from the labeled image data of six emotion categories. Even for video data, combing the features of each frame using simple sequential model or aggregation method can achieve good results. Thus, how to extract or learn better representations for image emotion data became very important. While for our new task, the amount of data we have is not enough to learn better features. On the contrary, the durations of moving of facial muscles or action unit (AU)[8] could be important.

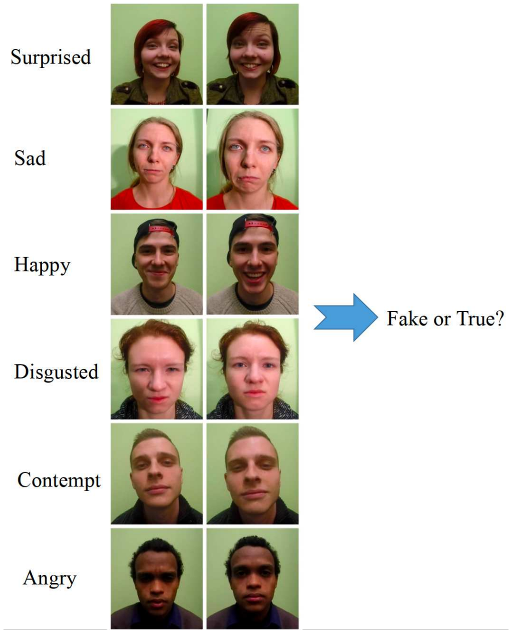

Figure 1. Examples of SASE-FE dataset. The left column is true or genuine expressions ,while the right column is fake or deceptive ones. We can see that it's difficult to recognize the differences for human, especially through only one frame.

To utilize the temporal information, we introduce temporal attention gated model (TAGM) [21] for this emotion recognition task. The TAGM model infer the temporal attention score of each time step, and update its hidden states based on the score. In our case, it would infer the important frames from the input video data, and form an representation of the video. Furthermore, we explore the the Long Short Term Memory networks (LSTM) autoencoder [19], which could learn to represent the video data unsupervisedly. Rather than only using texture features which describe the whole face area, we believe that the facial landmark sequences could reflect the temporal changes more directly. Thus, we combine the two models with two features as our final method. The experiment is conducted on a newly published dataset, the SASE-FE database [20]. See Fig.1 for examples of video frames.

To present the above points, we organize the paper as follows: In Section 2 we review some related research. Then we describe the temporal models including the temporal attention gated model and the LSTM autoencoder in Section 3. The experiment settings are presented in Section 4, including baseline visual features and comparisons with other models. The final results and conclusions are given in Section 5 and 6.

2. Related Work

There are many works focusing on emotion recognition. Our work is motivated by the recent dynamic emotion recognition works and sequence models. We review them below.

Dynamic Emotion Recognition: Existing research mostly focused on classifying the six basic emotion categories[7]. Kahou et al. [15] used convolutional neural network and deep belief network for dynamic emotion classification. Liu et al.[17] used Grassmannian Manifold to get facial expression features, then they combined Riemannian Manifold and deep convolutional neural network (CNN) in [18]. Yao et al.[26] combined the CNN model with facial action unit aware features. Fan et al. [10] combines CNN-Recurrent Neural Networks (RNN) model and 3D CNN model and got the state-of-the-art result for facial expression recognition in videos. Most of the works focused on exploring visual features for emotion category classification, and directly using RNN or temporal aggregation methods to represent the temporal information. On the contrary, we explored using temporal attention model for new type of task: automatic recognition of deceptive facial expressions. Our model could learn to focus on more important frames, and at the same time learn the sequential relations of each time step.

Temporal Modeling: Some works have investigated sequential or temporal models for facial expression and action unit research. Baltrusaitis et al. [3] developed continuous conditional random fields for facial action unit detection. Liu et al. [16] developed spatio-temporal manifold learning for dynamic facial expression recognition. Recently, with the popular the representation learning methods, lots of emotion recognition researches [6, 10, 24] utilize RNN and its variants. The RNN [22] learn to construct a representation for each time step based on a current observation and the representation of previous time step. To address the gradient vanishing problem of plain-RNN when dealing with long sequences, the Long Short Term Memory networks (LSTM)[13] and the Gated Recurrent Units (GRU)[5] were proposed. They are equipped with a gating mechanism to balance the information flow from the previous time step and current time step dynamically. LSTM and GRU model contain more parameters to learn than RNN ones. Thus, they require more training data, which is not always available in emotion recognition settings. Inspired by the attention model, we explore the TAGM [21] model for emotion recognition task, which employs a gate to filter out the noisy time steps and preserve the salient ones.

Also, we explore an LSTM autoencoder [19] which is an unsupervised model and can utilize unlabeled data. The texture feature may face the problem of cross dataset issues, as it is more sensitive to the facial images. As to facial landmarks which are highly abstracted feature, it may be useful to learn more general representation using extra data. In this way, the LSTM autoencoder could be helpful.

3. Temporal Modeling

Instead of using the usual recurrent neural network model for emotion recognition, we explore the use of two temporal models: temporal attention gated model (TAGM) and autoencoder LSTM (eLSTM). The TAGM is a combination of temporal attention model and gated recurrent neural network, and has shown good results in spoken digit recognition, text-based sentiment analysis and visual event recognition [21]. The eLSTM can encode different length sequences features into a fixed length vector without supervision. Thus the eLSTM is good at utilizing unlabeled or weak-labeled data [19].

3.1. Temporal Attention Gated Model

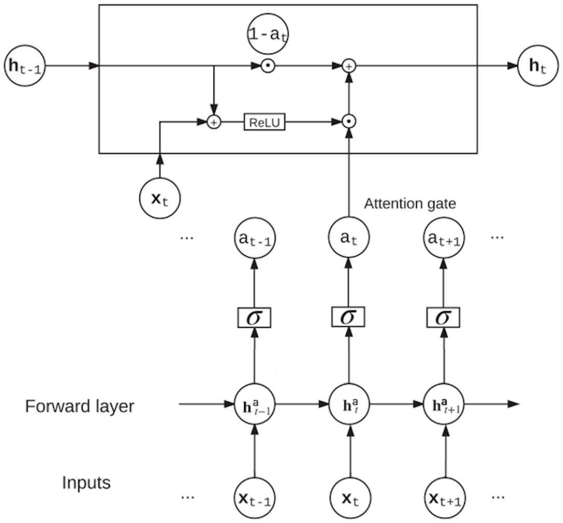

Figure 2. The components of TAGM model. The upper part is the recurrent attention-gated unit and the lower part is the temporal attention module.

We utilize the Temporal Attention-Gated Model

(TAGM) [21] which is able to capture the temporal information by using fewer parameters. The TAGM has a temporal attention module and a recurrent attention-gated unit. The temporal attention module is employed to measure the relevance of each time step of a sequence to the final decision. The goal of the recurrent attention-gated units is to learn a hidden sequence representation which integrates the temporal attention scores. See Fig.2 for the architecture of the TAGM model we use.

To model the influence of each time step, we infer the attention score a_t using an RNN:

$$a_t = \sigma(m \cdot h_t^a + b_a) \tag{1}$$

Here m is the weight vector of the RNN hidden states and b_a is the bias term. A sigmoid function is employed as the activation function σ at the top layer of the attention module in Equation 1 to constrain the attention weight to lie between [0, 1]. h_t^a is the hidden representations of the RNN model:

$$h_t^a = g(W_r x_t + U_r h_{t-1}^a + b_r) \tag{2}$$

The ReLU functions are used as the activation functions g. The inferred attention weights a_t serve as the attention gate for the following Recurrent Attention-Gated Units to control the involved information flow.

In order to integrate the attention scores in the recurrent network units, TAGM uses an attention gate to control how much information is incorporated from the input of the current time step based on the salience and relevance to the final task. Formally, given an input sequence $x_{1,...,T} = x_1, ..., x_T$ of length T in which $x_t \in \mathbb{R}^D$ denotes the observation at the t-th time step, the attention score at time step t is denoted as a_t , which is a scalar value that indicates the salience of current time step to the final decision. The recurring process where the hidden state h_t at time step t is modeled as a convex summation:

$$h_t = (1 - a_t) \cdot h_{t-1} + a_t \cdot h_t' \tag{3}$$

Here, h_{t-1} is the previous hidden state and h_t' is the candidate hidden state value which fully incorporates the input information x_t in the current time step:

$$h_t' = g(W \cdot h_{t-1} + U \cdot x_t + b) \tag{4}$$

Here W and U are respectively the linear transformation parameters for previous and current time steps while b is the bias term. The rectified linear unit (ReLU) is used as the activation function g. Equation 3 uses attention score a_t to control the tradeoff between incorporating and skipping new information. High attention value will push the model to focus more on the current hidden state h_t' and input feature x_t , while low attention value would make the model ignore the current input feature and inherit more information from previous time steps. The learned hidden representation R_t at the last time step h_T of the sequence is further fed into the final classifier, or used as the final feature of TAGM model.

3.2. LSTM AutoEncoder

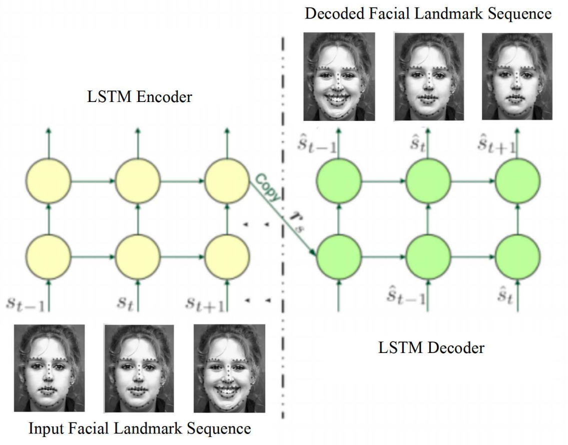

Figure 3. eLSTM model with facial landmarks. The model try to reconstruct the input landmark sequences in the reversed order.

The amount of video data is relatively limited compared to image data. Thus, we introduce LSTM autoencoder (eLSTM) [19] as an unsupervised way of learning the temporal representation. In Fig.3 we show the time-unfolded encoder and decoder parts of eLSTM.

The goal of eLSTM is to encode the input visual feature sequence, $s = s_t : t = 1, 2, ...,$ for example the locations of facial landmarks in our experiments. The sequences may have variable lengths in time. eLSTM observes the entire feature sequence s, and encodes it to a feature vector r_s . To learn the encoder, a decoder LSTM tries to reconstruct the normalized input visual features in the reverse order. The reconstruction error is then estimated in terms of the mean-squared error, and used to jointly learn both the encoder and decoder LSTMs. The reversed output reconstruction benefits from low range correlations which make the optimization problem easier. The input to the encoder at each time step t, is the output of the decoder at time step $t-1$, i.e. s_{t1}.

3.3. Combined model

In our experiments, we find that TAGM model and eLSTM model show different discriminative ability on different emotion categories. To utilize the diversity, we combine the two models together, as shown in Fig.4. The hidden states of TAGM and eLSTM are concatenated as a final representation of the image sequence. On top of the joint representation, we use a sigmoid classifier

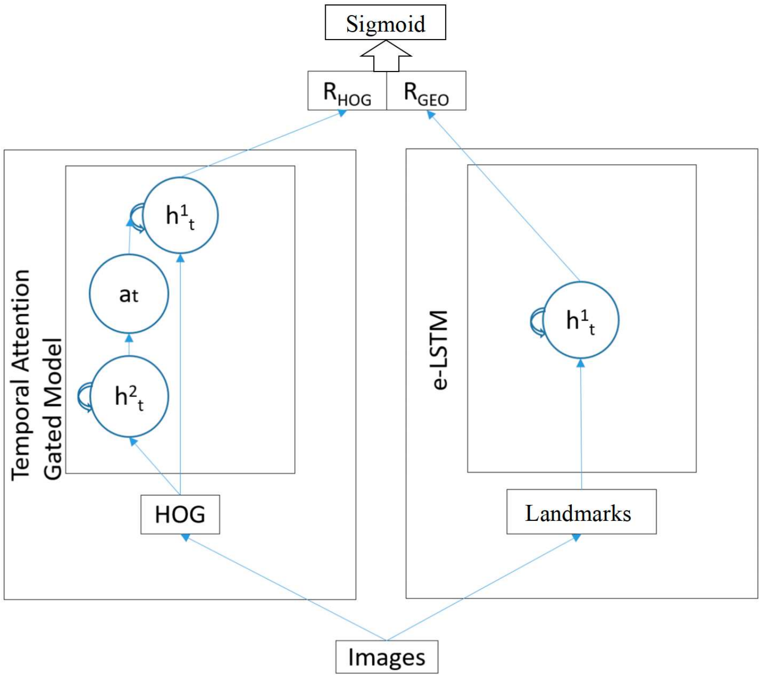

Figure 4. The architecture of final model, which combines the TAGM and eLSTM models. t is the current timestep. R_{HOG} is the hidden state of HOG + TAGM model while R_{GEO} is the hidden state of Facial landmarks + eLSTM model. The joint representation $R_{HOG} + R_{GEO}$ is then used for the final prediction.

4. Experiments

4.1. Settings

Datasets: We conduct the experiments on the ChaLearn Fake-vs-True emotion challenge [14] dataset SASE-FE[20]. The emotion challenge dataset contains video set of 50 subjects. For each subject, there are 12 videos representing 6 basic emotions (angry, happy, sad, disgust, contempt, surprise) for genuine and deceptive expressions. Each video was recorded with a high-resolution camera with 100 frames per second and is about 3-4 seconds. In order for the subjects to express these emotions, they were shown videos which are meant to induce these emotions and were acted accordingly. In each video, subjects started from a neutral emotion and the length of this neutral emotion is not predefined.

For this task, as the SASE-FE test set is not publicly available, we randomly use 80% of its train set for training, and use the left 20% for hold-out validation. Then, we test the model performances on the challenge validation set. The performances is judged by the average binary classification accuracies across all 6 emotion categories.

Implementation: We extracted all the face images using the OpenFace [4] library. We performed a similarity transform to align all images to a common reference frame using tracked facial landmarks, with a resolution of 112×112. For each video of image sequences, we select 10 frames close to the rear with an interval of 20 frames.

We use RNN, LSTM and GRU as baseline temporal models. For them and TAGM model, we set the hidden state dimension to 16 and use a 0.25 dropping rate drop-out method for the output representation. For eLSTM, we set the temporal and layerwise drop-out rate to 0.5. All sequential models were trained using the AdaDelta [27] gradient descent with the initial learning rate of 1.0. Early stopping regularization on the held-out train set was used to avoid over-fitting during the training. We follow the mini-batch training mode, batches of which are set to 16 samples. Separate models are trained for each emotion categories. All implementations were based on the tensorflow library [1].

We also use linear support vector machine (SVM) and multilayer perceptron (MLP) as baseline models. The visual features of each timestep are concatenated and feed to SVM or MLP. The MLP has two fully connected layer, with 1024 and 2 dimensions respectively.

4.2. Visual Representations

2D Facial Landmarks: The geometric feature is based on the theory of emotion action unit defined by Ekman [9]. With it, Valstar et al. [25] successfully detected AU detection and then achieved a better result than some complex descriptors. We use the 2D geometric feature from Open-

Face [4] library, which describe the geometric realtions of all the facial landmark points. The geometry feature of a video is by taking the mean feature values of every frame.

HOG: We use the histogram of oriented gradients descriptor (HOG) as the basic visual features. HOG feature is widely used in emotion and facial action unit as it is capable of describing the local object appearance and shape within an image. [2]

To detect oriented edges, HOG feature utilizes the image gradient. The facial image is divided into 11×11 cells, and for the pixels within each cell, a histogram of gradient directions is compiled according to the computed gradient G and θ. The descriptor is the concatenation of these histograms. For improved accuracy, the local histograms are contrast-normalized by calculating a measure of the intensity across a block (2×2 cells), and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination and shadowing. Following the setting of Baltrusaitis et al.[2], we also perform person-specific normalization for HOG feature by subtracting the mean HOG feature of the same person.

CNN: The convolutional networks (CNN) have been very popular among computer vision community. To explore the use of CNN in this task, we utilize two CNN models: the facial expression recognition model [23] trained from FER-2013 dataset [11], and the facial action unit (AU) detection model [12] trained from the BP4D dataset [28]. For CNN feature, we downsampled the images to 48×48.

The first convolutional layer of expression-CNN filters the 48×48 input patch with 64 kernels of size 5×5. The second convolutional layer takes as input the response-normalized and max-pooled output of the first convolutional layer and filters it with 64 kernels of size $3 \times 3 \times 64$. The third, fourth, and fifth convolutional layers are connected to one another without any intervening pooling or normalization layers. The third convolutional layer has 128 kernels of size $3\times3\times64$ connected to the (normalized, pooled) outputs of the second convolutional layer. The fourth and fifth convolutional layers both have 128 kernels of size $3 \times 3 \times 128$. The fully connected (FC) layers have 1024 neurons each. The rectified linear unit activations are applied to the output of every convolutional or fully connected layer.

The first convolutional layer of AU-CNN model has a kernel size of $5 \times 5 \times 64$, followed by a max-pooling layer of kernel size 3 3 with a stride of 2 in each dimension. The following convolutional layers have the kernel size of $5\times5\times64$ and $4\times4\times128$ respectively. Finally, the output of the third convolutional layer is fed into the last hidden layer of the network, a fully connected linear layer with 3072 neurons. Dropout technique is applied to this fully connected layer, with a dropout probability of 0.2. The output of this layer is connected to the output layer that has a dimension of 11, which represents the occurrence of 11 AUs labeled in[28].

4.3. Comparison methods

We compare our work with some recently developed methods concerning emotion recognition. The comparison is conducted focusing on following aspects: 1) Comparison between visual features: we compare different types of visual features by testing the classification accuracies using TAGM model. 2) Comparison between temporal models: we compare the results of different types of temporal models using the best feature from previous experiment 3) Comparison between cross-emotion and single-emotion training. As the amount of data is limited, we want to explore if we can use data from other emotions to enhancing the training progress. 4) Comparison of model combination: this is to test if the combination of different features and models can help improving the classification result.

5. Results and Discussion

5.1. Visual features

Table 1. Visual features with TAGM model results, note than HOG outperforms the CNN models

Features	Accuracies
Expression-CNN	0.517
AU-CNN	0.517
Facial Landmarks	0.600
HOG	**0.633**

In Table 1, we list the classification accuracies of several types of visual representation features, including CNN, AU, Facial Landmarks and HOG. We can note that HOG performs better than CNN, which is reasonable cause the CNN model is fine-tuned from emotion classification model trained on FER dataset. As to AU feature, AU detection is still a difficult task compared to facial landmark detection, so it makes sense that AU feature got the lowest result. Meanwhile, facial landmarks also show a promising result.

5.2. Temporal modeling

In Table 3, we list the classification accuracies of HOG feature with different temporal models. We can see that TAGM model outperforms all the other models. Note that RNN works better than GRU and LSTM, as we only select 10 time-steps from a video. An interesting result is that an SVM with concatenated time step features also performs better than LSTM, which show that SVM is still promising when the amount of data is limited.

5.3. Cross-emotion training

In Table 4, we list the results of cross-emotion or per-emotion model. We choose to use HOG with TAGM model,

Table 2. Recognition accuracies of each specific emotion categories. The proposed combination model works well on most emotions, while the others only work well on some emotions.

Features	Angry	Contempt	Disgusted	Happy	Sad	Surprised	Mean
HOG + TAGM	**0.90**	0.50	0.60	0.60	0.60	0.60	0.63
Facial Landmarks + eLSTM	0.60	0.50	0.60	0.50	**0.70**	**0.70**	0.60
HOG-Landmarks + TAGM	0.50	**0.60**	0.50	0.70	0.30	0.60	0.53
Proposed Model	0.70	0.50	**0.80**	**0.80**	**0.70**	0.60	**0.68**

Table 3. Temporal models with HOG feature results

Model	Accuracies
MLP	0.550
LSTM	0.550
GRU	0.550
RNN	0.567
SVM	0.600
TAGM	**0.633**

Table 4. Cross emotion result of HOG + TAGM, model trained for specific emotion works better

Model	Accuracies
TAGM per category	**0.633**
TAGM cross category	0.500
TAGM pretrained cross category	0.617

which is the best among the previous experiments (The model is trained per emotion category). From the results, we can see that it is hard to use one single model to distinguish fake emotions among all six categories (TAGM cross category). Even using other 5 emotions for pre-training (TAGM pretrained cross category) would decrease the classification ability.

5.4. Combining Model

In Table 2, we list the results of model combination. We try to combine two types of features at different levels. First we try to concatenate HOG and geometric features and feed it to TAGM model. Result show that this type of combination does not help. Then, we try concatenate the hidden states of HOG+TAGM and geometic+eLSTM, and use the combined hidden states for binary sigmoid classification. The result shows that this combined model outperforms the TAGM and eLSTM, on most specific emotion categories and the whole validation set. Thus, we submitted this combined model for final testing of the ChaLearn Fake-vs-true emotion challenge, and got an accuracy of 61.7% on the test set [14].

6. Conclusion

The paper explores a new type of automatic emotion recognition task - distinguishing fake and true emotions from video clips. We introduce temporal attention model for this emotion recognition task to recognize the importance of each frame. We utilize facial geometric features by encoding feature sequence unsupervisedly. Combining the facial texture and geometric feature, we gain promising testing result.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 4

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015. 5

[3] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Continuous conditional neural fields for structured regression. In *European Conference on Computer Vision*, pages 593–608. Springer, 2014. 2

[4] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016. 4, 5

[5] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 2

[6] 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*, pages 467–474. ACM, 2015. 2

[7] P. Ekman. Cross-cultural studies of facial expression. *Darwin and facial expression: A century of research in review*, 169222, 1973. 1, 2

[8] P. Ekman and W. V. Friesen. Facial action coding system. 1977. 1

[9] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Fa-*

cial Action Coding System (FACS). Oxford University Press, USA, 1997. 4

[10] 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*, pages 445–450. ACM, 2016. 2

[11] 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*, pages 117–124. Springer, 2013. 5

[12] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–5. IEEE, 2015. 5

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[14] W. Jun, E. Sergio, B. Xavier, J. E. Hugo, G. Isabelle, M. Meysam, A. Jury, G. Jelena, and A. Gholamreza. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV*, 2017. 4, 6

[15] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013. 2

[16] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014. 2

[17] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen. Partial least squares regression on grassmannian manifold for emotion recognition. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 525–530. ACM, 2013. 2

[18] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014. 2

[19] B. Mahasseni and S. Todorovic. Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3054–3062, 2016. 1, 2, 3

[20] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari. Automatic recognition of deceptive facial expressions of emotion. *arXiv preprint arXiv:1707.04061*, 2017. 1, 2, 4

[21] W. Pei, T. Baltrušaitis, D. M. Tax, and L.-P. Morency. Temporal attention-gated model for robust sequence classification. *arXiv preprint arXiv:1612.00385*, 2016. 1, 2, 3

[22] J. Schmidhuber. A local learning algorithm for dynamic feedforward and recurrent networks. *Connection Science*, 1(4):403–412, 1989. 2

[23] B. Sun, L. Li, G. Zhou, and J. He. Facial expression recognition in the wild based on multimodal texture features. *Journal of Electronic Imaging*, 25(6):061407–061407, 2016. 5

[24] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu. Lstm for dynamic emotion and group emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 451–457. ACM, 2016. 2

[25] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–8. IEEE, 2015. 4

[26] 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*, pages 451–458. ACM, 2015. 2

[27] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 4

[28] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 5