AST-Net: An Attribute-based Siamese Temporal Network for Real-Time Emotion Recognition

Shu-Hui Wang
chch80703@gmail.com
Chiou-Ting Hsu
cthsu@cs.nthu.edu.tw

Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan

Abstract

Predicting continuous facial emotions is essential to many applications in human-computer interaction. In this paper, we focus on predicting the two dimensional emotions: valence and arousal, to interpret the dynamically yet subtly changed facial emotions. We propose an Attribute-based Siamese Temporal Network (AST-Net), which includes a discrete emotion CNN model and a Stacked-LSTM, to incorporate both the spatial facial attributes and the long-term dynamics into the prediction. The discrete emotion CNN model aims to extract attribute-related but pose- and identity-invariant features; and the Stacked-LSTM is used to characterize the dynamic dependency along the temporal domain. Furthermore, in order to stabilize the training procedure and also to derive a smoother and reliable long-term prediction, we propose to jointly learn the model from two temporally-shifted videos under the Siamese network architecture. Experimental results on AVEC2012 dataset show that the proposed AST-Net not only processes in real time (40.1 frames per second) but also achieves the state-of-the-art performance even when using the vision modality alone.

1 Introduction

Emotion recognition from human faces is an active research area and plays a vital role in many applications. Especially, because human faces are considered as important social signals in human communication, a long-term prediction of emotions will benefit various applications involving human-computer interaction, such as health care and driver assistance systems. Existing methods on automatic facial emotion recognition can be classified into two categories. One is discrete emotion recognition, which usually models the recognition of six or seven universal expressions (such as anger, contempt, disgust, fear, happiness, sadness, and surprise) as a classification problem. Datasets such as Cohn-Kanade [14, 31] and MMI [21, 33] are popularly used to evaluate the performance of discreet emotion recognition. In these datasets, the subjects are asked to pose a specific discrete emotion [14, 21, 31, 33], and each sequence usually has one peak frame and shares similar variation (e.g., neural-to-peak followed by peak-to-neural) along the time domain. The other category of emotion recognition is to analyze the four dimensional emotions, including Valence, Arousal, Expectation and Power. In this paper, we focus on predicting the two widely used dimensions: valence and arousal, where valence measures the degree of emotion toward positive or negative state, and arousal measures the degree of emotional stimulation. We conduct our proposed method on two dimensional emotion datasets AVEC2012 [28] and RECOLA [24, 25]. AVEC2012 dataset [28] was collected by recording the conversations between humans and artificially intelligent agents. This recording scenario is called Sensitive Artificial Listener (SAL) technique [15] and can stimulate richer and natural emotion changes of humans through the conversations. The other dataset, RECOLA [24, 25], was collected by recording a number of participants collaboratively completing a task. In comparison with discreet emotion datasets, dimensional emotion datasets contain more subtle, complex and long-term affective behavior and can better reflect spontaneous human emotions.

Nowadays, even though discrete emotion recognition has achieved very good performance, prediction of dimensional emotions is still far from satisfactory. One of the major challenges comes from the insufficient training data as well as the unreliable labels in the datasets. For example, the training set of AVEC2012 dataset [28] contains only 400k frames of seven subjects, and has very few frames labeled with strong numerical values. Hence, it is extremely difficult to learn a regression model from the small-scaled data without suffering the over-fitting problem. As to the reliability of labels, because the dimensional emotions are labeled in real values, it is by no means easy for a single marker to provide consistent labels, let alone reach a consensus with other markers. Furthermore, the issue of time-delayed label, which results from the temporal delay between the video frame and its labels, also raises serious concerns when training the prediction model. In [19], the authors proposed to estimate the delay probability by assuming that those features which are more relevant to the prediction should be more correlated to the undelayed labels; once determining the average delay, they shifted all the labels with this constant delay. However, because different markers may induce different delays at different instants, it seems impractical to shift all the labels with a constant delay.

To tackle the above-mentioned difficulties, our goal is to build a temporal prediction model and to deal with the insufficient training data, unreliable and non-constant time-delayed labels. Instead of determining the label delay as in [19], we propose to learn a temporal model by characterizing the dependency between predictions of consecutive frames. As to the issue of insufficient training data, we propose to learn an attribute-related feature representation by leveraging the rich information in existing discrete emotion datasets. Moreover, because there exist various facial variations that are unrelated to the emotional changes (e.g. individual characteristics, ethnic, illumination changes, and poses), the learned representation should capture the emotional attributes but be invariant to other irrelevant variations. Instead of learning appearance and shape features [1, 4, 19, 20, 27, 30], we propose to learn the feature extractor by training on a discrete emotion dataset that contains different subjects with ethnic and posture variations. Note that, although recent methods [1, 19, 20, 27, 30] usually fuse predictions of multiple modality, e.g. vision, audio and text, to achieve better performance, this paper does not focus on model fusion techniques but instead aims to investigate a vision-only model which can efficiently predict an accurate and temporally smooth result.

To sum up, we propose a dimensional emotion prediction method via an Attribute-based Siamese Temporal Network (AST-Net). AST-Net consists of three major parts. The first one is to extract Attribute-related emotional features using a discrete emotional CNN model. The second one is to learn the temporal dependency between frames using the Stacked-LSTM. Finally, we use the Siamese Network to include the relationship between two temporally shifted sequences by minimizing a new loss function.

Our contributions are summarized as follows:

- We propose to learn the feature representation from the discrete emotion dataset so as to capture the emotion related attributes and to circumvent the limitations of AVEC2012 dataset.
- Through the Stacked-LSTM, we reduce the impact of short-term label-delay by referring to the long-term temporal information.
- With the twin networks in the Siamese network, we propose a new loss function to stabilize the training procedure and also to derive a much smoother and reliable long-term prediction.
- The proposed AST-Net not only processes in real time but also achieves the state-of-the-art performance even when only using the vision modality.

2 Related Work

Representation of facial appearance and its subtle change is crucial to spontaneous and dimensional emotion recognition. The idea of learning spatio-temporal features have been popularly studied in video recognition. A number of research has achieved great success in video recognition [6, 7, 11, 13, 17, 29, 32] based on deep learning techniques. In [11, 32], 3D spatio-temporal filters are used to learn the spatio-temporal features; and in [13, 17], various temporal sampling and pooling methods are studied to combine information from different temporal durations. However, because of the large amount of parameters, it is very difficult to learn the 3D filters from longer video clips. As to temporal pooling, even though the slow fusion model in [13] can preserve more global temporal information, it is still insufficient to capture the subtle changes in time domain. In [29, 34], two-stream architecture has been used to fuse spatio-temporal information for action recognition. Nevertheless, unlike general actions, spontaneous emotions often contain subtle movements. We need a better strategy to characterize the locally and subtly moved facial appearance.

On the other hand, other methods attempt to capture the temporal information through temporal models. For example, the models, e.g., TDNN, HMM, CCRF, have been used to learn the temporal relationships in [1, 16, 20]. In addition, Long Short Term Memory (LSTM) networks [9], which is one type of recurrent neural network with the capability of modeling long-range temporal relationships, has been successfully adopted in video recognition [6, 10]. Motivated by the success of LSTM, we will adopt the Stacked LSTM in the proposed model.

3 Proposed Method

Figure 1 shows the flowchart of the Attribute-based Siamese Temporal Network. We first off-line train a CNN model on a discrete emotion dataset so as to better exploit the much larger data and their more reliable discrete labels. Next, we use the CNN model to extract a feature vector for each frame. After obtaining the features of each frame, we use the Stacked-LSTM to learn the temporal dependency along the time domain. Because the training data in dimensional emotion dataset is of very small scale and with unreliable labels, even with Stacked-LSTM, we often obtain very diverse and noisy predictions along a short period. We thus take advantage of the Siamese network architecture and propose to involve two temporally shifted videos into the learning stage. Finally, we combine the predictions of the two videos and define a new loss function to derive the final prediction.

3.1 Pre-Processing

Before learning the models, we first detect and crop face regions from input videos, and then align all the faces using the locations of eyes. Next, in order to learn the long-term

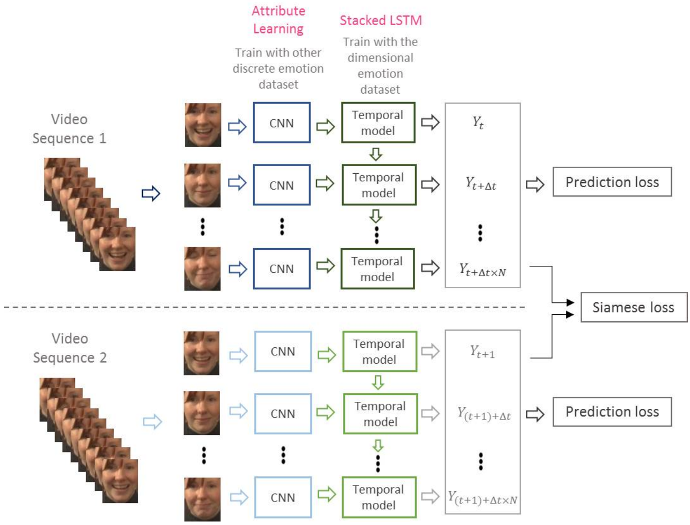

Figure 1: **AST-Net: An Attribute-based Siamese Temporal Network.**

temporal information, we construct a number of video clips of fixed length (i.e., N = 50 frames in our experiments) by sampling the input videos every 10 frames (i.e., $\Delta t = 10$ in Fig. 1). Therefore, each video clip of 50 frames covers the temporal duration of 500 frames in the original video. From our experiments, training on the sampled video clips indeed improves the overall performance than training on the original videos. The reason is that there is usually very little emotional change over a short period, and the model can hardly learn much temporal information from video clips covering a very short duration.

3.2 Discrete emotion CNN model

We use VGG-16 as our CNN model, and pre-train the model on VGG-Deepface [22]. We then fine-tune the CNN model on the discrete emotion dataset, Fer2013 [8]. As noted before, in comparison with AVEC2012 dataset and its real-valued labels, the discrete emotion dataset is much easier to collect and to assign its discrete labels. Therefore, the CNN model is more likely to converge without suffering the over-fitting problem when training on the discrete emotion dataset. In addition, although discrete emotion and dimensional emotion indicate different characteristics, the six discrete labels in Fer2013 are not irrelevant to the continuous emotion dimensions. From the theory of psychology [23], e.g., "Each emotion can be understood as a linear combination of these two dimensions, or as varying degrees of both valence and arousal.", we argue that, both the discrete and dimensional emotions are implicitly correlated and do rely on some common but latent facial attributes. Thus, we can view the features extracted from the fully connected layer (fc6) of the CNN model as the attribute-related features, which capture the degree of tendency toward a certain discrete expression, e.g., the degree of anger or smile.

Furthermore, we observe that, in the AVEC2012 dataset, facial appearances reflect not only the emotional changes but also individual characteristics, ethnic, and environmental variations (e.g., viewpoint changes, illumination changes, and occlusions). If we simply use the AVEC2012 dataset to train the CNN model, the learned feature may mostly capture the noisy variations unrelated to the emotional changes.

3.3 Stacked-LSTM

Next, after extracting the features of each video clip, we use a stack of LSTM to learn the dynamic and long-short term temporal dependency between frames. With the Stacked-LSTM, we also alleviate the problem of time-delayed labels; because when referring to long-term trend, we can largely reduce the impact caused by the short-time delay.

3.4 Siamese Temporal Network and Loss Function

Finally, we observe that the predictions of Stacked-LSTM are still very diverse in a short-term period. In order to learn a smoother prediction along the temporal domain, we propose to use the Siamese network to jointly learn from two temporally-shifted clips, whose time stamps are $[T_t, T_{t+\Delta t}, ..., T_{t+\Delta t \times N}]$ and $[T_{t+1}, T_{(t+1)+\Delta t}, ..., T_{(t+1)+\Delta t \times N}]$, respectively. ($\Delta t = 10$ in the experimental setting)

The Siamese network consists of two identical networks; each network processes an input sequence, and the two networks are then joined by a loss function to aggregate their outputs. We define the loss function as follows,

$$Loss = Prediction\ loss + \lambda \times Siamese\ loss. \tag{1}$$

Prediction loss. The prediction loss is a linear combination of 4 loss terms, defined by:

$$Prediction\ loss = label\ loss + w_1 \times trend\ loss + w_2 \times local\ loss + w_3 \times global\ loss. \tag{2}$$

1. The label loss term measures the Euclidean distance between the predictions and the ground truth labels of all the frames in the video clip.
2. The trend loss term is designed to measure the overall trend of emotional changes. Even though the exact values of ground truth labels are highly subjective, these labels usually reach a consensus on how the dimensional values ascend (or descend) in the temporal domain. We thus measure the distance of the label changes between adjacently sampled frames to constrain the overall trend of the prediction.

$$\begin{gathered} trend\ loss = \frac{1}{2N}\sum_{i=1}^{N} \|\hat{g}_i - \hat{y}_i\|^2,\ N: the\ number\ of\ frames\ in\ a\ video\ clip, \\ where\ \hat{g}_i = gt_i - gt_{i-\Delta t},\ \hat{y}_i = y_i - y_{i-\Delta t}, \\ gt\ is\ the\ ground\ truth\ label\ and\ y\ is\ the\ prediction \end{gathered} \tag{3}$$

3. The local loss term is used to constrain the intensity changes in a local interval. We divide the video clip into M intervals and define the term by:

$$\begin{gathered} local\ loss = \frac{1}{2M}\sum_{k=1}^{M} \|\hat{g}_k - \hat{y}_k\|^2,\ M: the\ number\ of\ intervals, \\ where\ \hat{g}_k = \max_{i \in interval_k} gt_i - \min_{i \in interval_k} gt_i,\ and\ \hat{y}_k = \max_{i \in interval_k} y_i - \min_{i \in interval_k} y_i \end{gathered} \tag{4}$$

4. In addition to the local loss, we design the global loss to further constrain the range of prediction of the whole chip. Because the exact values of dimensional labels are difficult to predict, our earlier experiments show that the prediction tends to fall into a small range. We thus define this term to minimize the distance of global ranges by:

$$\begin{gathered} global\ loss = \|\hat{g} - \hat{y}\|^2, \\ where\ \hat{g} = \max gt - \min gt,\ and\ \hat{y} = \max y - \min y \end{gathered} \tag{5}$$

Siamese loss. The Siamese loss term is to minimize the distance between predictions of the two clips. With this term, we expect to derive a temporally smoothing and reliable prediction.

$$\begin{gathered} Siamese\ loss = \frac{1}{2N}\sum_{i=1}^{N}\|\hat{g}_i - \hat{y}_i\|^2,\ N: the\ number\ of\ frames, \\ where\ \hat{g}_i = gt_i - gt_{i-1},\ and\ \hat{y}_i = y_i - y_{i-1} \end{gathered} \tag{6}$$

4 Experiments

4.1 Datasets

AVEC2012 Dataset. The goal of the 2nd Audio-Visual Emotion recognition Challenge and Workshop (AVEC 2012) is to recognize four continuously valued affective dimensions: arousal, expectancy, power, and valence. This challenge uses the SEMAINE corpus as the source of data, which consists of a large number of emotional interactions between human participants and sensitive artificial listener agents. The dataset contains 95 video clips, which are split into 31 training sessions, 32 development sessions, and 32 test sessions. Each video is recorded at frame rate of 49.479 frames/s and with resolution of 780x580 pixels. Although the dataset also contains audio modality, we only use the visual modality to conduct experiment and evaluate the performance. More details about the dataset can be found in [28].

Fer2013 Dataset. This dataset contains over 20k subjects with various ethnics and poses. There are 35887 images, including 4953 "Anger" images, 547 "Disgust" images, 5121 "Fear" images, 8989 "Happiness" images, 6077 "Sadness" images, 4002 "Surprise" images, and 6198 "Neutral" images. The resolution of each image is 48x48.

RECOLA Dataset. The RECOLA database focuses on the research of spontaneous collaborative and affective interactions. This dataset consists of multimodal data, i.e. audio, visual, and physiological (electrocardiogram, and electrodermal activity) recordings of online interactions between participants, who were solving a task in collaboration [24, 25]. We use 9 training videos for training and 9 development videos for testing. Each video is recorded at frame rate of 25 frames/s and with resolution of 1280x720 pixels. Although this dataset contains audio and biosignal modality, we only use the visual modality to conduct experiment and evaluate the performance. Compared to AVEC2012 dataset, the RECOLA dataset is of much smaller scale and is more difficult for learning and analyzing. More details about the dataset can be found in [24, 25].

4.2 Evaluation Scheme

Similar to previous work [16, 19], we use the Correlation Coefficient (CC) to evaluate the performance. CC has been popularly used to measure the trend of emotional changes and is considered as more meaningful than the error measurement on prediction values. In the following experiments, we will show the Max CC, Min CC and Mean CC of the testing videos and will compare with other state-of-the-art methods using Mean CC.

4.3 Implementation Details

In the training stage, we use the training set of AVEC2012 and RECOLA to train their individual temporal model. During the testing stage, we evaluate the performance of AVEC2012 dataset on its development set and test set, and of RECOLA dataset on its development set. Because the size of face images in the pre-trained model VGG-Deepface is 224 x 224, we

Loss Function	Max CC / Min CC	Mean CC
La Loss	0.7903/0.1516	0.4241
T Loss	0.7190/0.0964	0.3965
La + T Loss	0.8182/0.1779	0.4757
La + S Loss	0.8098/0.2121	0.4282
La + T + S Loss	0.7939/0.2253	0.4593
La + T + G Loss	0.7616/0.2436	0.4886
La + T + Lo Loss	0.8157/0.1802	0.4968
La + Lo + G Loss	0.8191/0.1325	0.3843
La + T + Lo + G Loss	0.7931/0.1864	0.5047
La + T + Lo + G + S Loss	**0.8789/0.2337**	**0.5874**

Table 1: Evaluation of loss function on AVEC2012 Valence Development Set (La: Label Loss; T: Trend Loss; Lo: Local Loss; G: Global Loss; S: Siamese Loss)

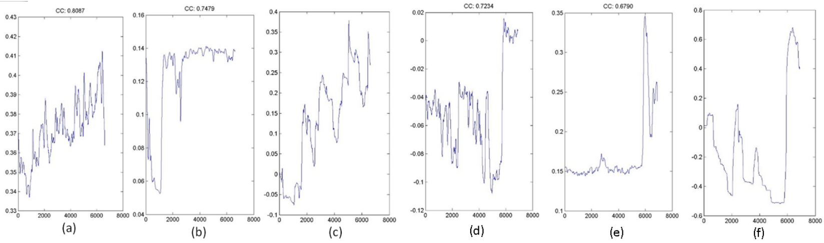

Figure 2: Predictions of valence in Video005 ((a)-(c)) and in Video001((d)-(e)) of AVEC2012 development set (x-axis is the frame number; y-axis is the prediction value). We use the proposed loss function in (a)(c); La+T+G+S loss in (b); and La+T+Lo+S loss in (e). (c) and (f) are the corresponding ground-truth labels.

resize all the face images in FER2013, AVEC2012 and RECOLA into 224 x 224 before fine-tuning and feature extraction. The dimension of extracted features from fc6 of the CNN model is 4096; the number of input frames to the Stacked-LSTM is 50.

Stacked-LSTM. We have tested with different numbers of layers and memory cells, and empirically determine to use two stacked LSTM layers, each with 50 and 40 memory cells.

Loss Function. Different error terms in the loss function have different ranges. For example, the error range of label loss is much larger than that of Siamese loss. Therefore, we need to assign different weights to balance the loss terms. In the experiments, we set the weights in equations 1 and 2 as $w_1 = w_2 = w_3 = 10$, and $\lambda = 1000$. Note that, the reason of different weight setting is because the magnitudes of errors in prediction loss and siamese loss are in very different range (e.g., the prediction loss is about 10^{-1}, siamese loss is around 10^{-4}).

4.4 Results

Evaluation of Loss Function. Table 1 shows the performance using different combinations of loss terms in the loss function.

We first investigate the local (Lo) and global (G) loss terms. In Table 1, from the results of La+T+G Loss and La+T+Lo Loss, the local loss better improves the performance than the global loss, because the global loss is designed to constrain the global range but cannot guarantee to minimize the local loss. Nevertheless, both local and global terms are critical to the overall performance. In Figure 2, we show the results when only removing either Lo or G loss from the loss function. The results show that, without either of them, the overall

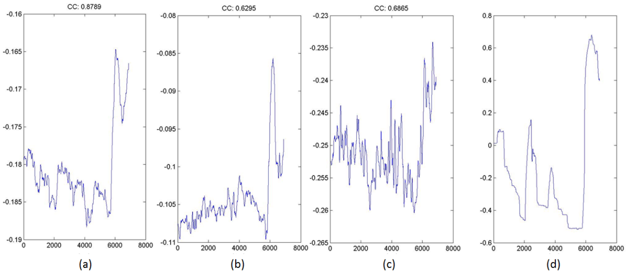

Figure 3: Predictions of valence on Video001 in AVEC2012 development set using (a) the proposed loss function; (b) La+T+S loss; and (c) La+T loss. (d) is the ground-truth label. (x-axis is the frame number; y-axis is the prediction value)

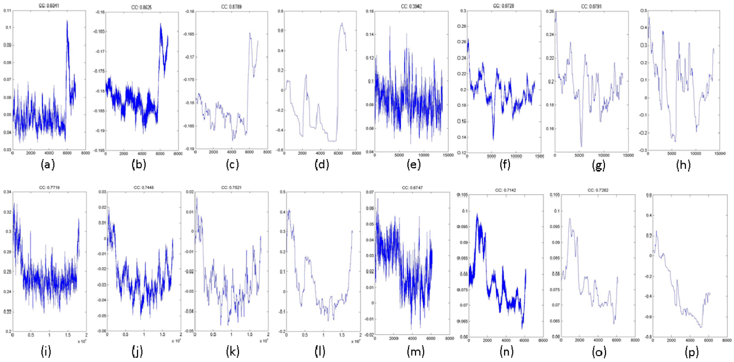

Figure 4: Predictions of valence in: (a)-(d) video001; (e)-(h) video010; (i)-(l) video022; (m)-(p) video013. In each video, we show the results of La+T+Lo+G loss, the proposed loss function, the median filtered prediction, and the ground truth labels, respectively. (x-axis is the frame number; y-axis is the prediction value)

predictions fail to capture the emotional changes in either short or long-term duration.

Next, when we include only label (La) or trend (T) loss in the loss function, the performance is poorer than combining these two terms (La+T). However, once we include the Siamese loss, i.e., (La+T+S), we have a smoother prediction but poorer CC performance. One possible reason is that, when enforcing a smoother prediction, we may also compromise the capability of predicting subtle changes in a very short duration (as shown in Fig.3 (b) and (c)). Therefore, in the proposed loss function, we include the local (Lo) and global (G) loss terms to simultaneously preserve the short-term and global emotional changes. As shown in Fig.3 (a), including Lo and G not only results in a smoother predictions but also increase the CC performance. More visualization results are given in Fig.4, where we only conduct median filtering to better visualize the predictions and still measure the performance using the original predictions. Fig.6 shows the statistics of valence prediction of all the videos in the development set. The results verify that the proposed loss function not only achieves the highest mean CC but also performs the best for most of the videos (for example, over 70 % of videos have CC > 0.5).

Comparisons. Table 2 and Table 3 show the comparisons of our method with existing methods on AVEC2012 Development Set and Test Set, respectively. Our method achieves the state-of-the-art performance in both sets. Note that, most existing work usually achieves

Method	**Mean CC**		
	Arousal	**Valence**	**Average**
Stepwise HMM [20]	0.3964	0.2348	0.3156
Fuzzy System [30]	0.52	0.47	0.495
Dynamic Cues [19]	**0.644**	0.350	0.497
Ours	0.5870	**0.5874**	**0.5872**

Table 2: Comparisons with existing methods on AVEC2012 Development Set.

Method	**Mean CC**		
	Arousal	**Valence**	**Average**
Stepwise HMM [20]	0.3248	0.1825	0.25365
Correlated Spaces [18]	0.46	0.2	0.33
CCRF [1]	0.341	0.326	0.3335
TDNN [16]	0.444	0.308	0.376
Baysian Fusion [27]	0.48	0.35	0.415
Fuzzy System [30]	0.42	0.42	0.42
Dynamic Cues [19]	0.61	0.341	0.4755
3D Model [4]	**0.564**	0.454	0.509
Ours	0.5442	**0.5362**	**0.5403**

Table 3: Comparisons with existing methods on AVEC2012 Test Set.

better performance on arousal prediction than on valence prediction. Especially, because audio modality is crucial to arousal prediction, inclusion of audio modality often favors the arousal prediction over the valence prediction. For example, Dynamic Cues [19], which fused vision and audio models in the method, perform slightly better on arousal than ours. Nevertheless, in terms of average performance, AST-Net outperforms these methods even using only the vision model. Moreover, the proposed model can process the test videos in real time (with frame rate 41 frames/s) and also achieve good performance even when the testing videos contain large pose variations (Fig. 5). We believe the pose-invariant predictions may attribute to two reasons. One is that the FER2013 dataset contains over 20k subjects with different poses and ethnics; and the other is that the Stacked-LSTM refers to longer duration and is less sensitive to short-term noisy features due to pose variations.

Table 4 shows the comparisons of our method with existing methods (using vision modality alone) on the RECOLA Development Set. Because the goal of RECOLA dataset is on research of collaborative and affective interactions, the subjects did not always look at the camera with any emotional change. Furthermore, the RECOLA dataset consists of multi-modal information (including audio, visual, and physiological signal). Because the audio and biosignals features are more related to the emotional changes in this dataset, (e.g. cc = 0.788 on Arousal when applying to audio modality alone in [25]), most existing methods focus on the fusion methods to improve the performance. Therefore, we achieve merely comparable performance with existing methods on the RECOLA dataset.

5 Conclusion

We proposed an Attribute-based Siamese Temporal Network (AST-Net), which includes a discrete emotion CNN model and a Stacked-LSTM, to incorporate the latent facial attributes

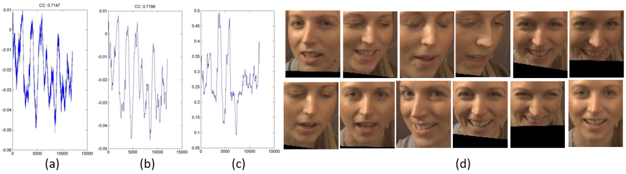

Figure 5: Video021 in AVEC2012 development set. (x-axis is the frame number; y-axis is the prediction value) (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.

Method	Mean CC		
	Arousal	Valence	Average
AVEC 2015 baseline(apperance) [26]	0.183	0.358	0.2705
AVEC 2015 baseline(geometric) [26]	0.361	0.423	0.392
ETS System(apperance) [2]	0.173	0.263	0.218
ETS System(geometric) [2]	0.103	0.389	0.246
LSTM-RNN(LGBP-TOP) [3]	0.535	0.463	0.499
Multimodal-RNN(apperance) [5]	**0.571**	0.496	**0.5335**
Multimodal-RNN(geometric) [5]	0.471	**0.530**	0.5005
Ensemble(apperance) [12]	0.313	0.313	0.313
Ensemble(geometric) [12]	0.172	0.401	0.2865
NN fusion system [25]	0.427	0.431	0.429
Ours	0.4783	0.4445	0.4614

Table 4: Comparisons with existing methods on RECOLA Development Set.

and the long-term dynamics into the prediction. With the Siamese Network, we imposed a new loss function to stabilize the training procedure and also to derive a much smoother and reliable long-term prediction. The discrete emotion CNN model is trained to extract attribute-related emotion features which are also invariant to other unrelated factors. The Stacked-LSTM effectively characterizes the temporal dependency between these attribute-related features. Experiment results show that AST-Net consistently outperforms existing approaches and achieve the state-of-the-art performance in real-time even only using the vision modality. In the future, we will test AST-Net on more data and will also test on fusing multiple modalities to further improve the performance.

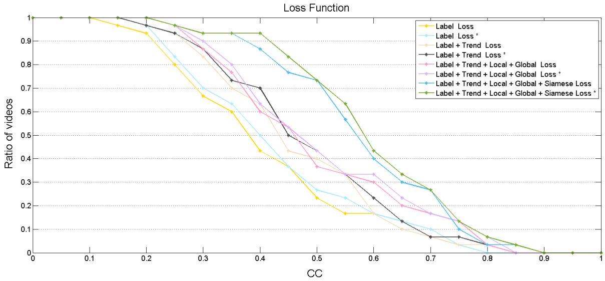

Figure 6: The statistics of valence predictions of all videos in AVEC2012 development set.

References

[1] T. Baltrušaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[2] P. Cardinal, N. Dehak, AL. Koerich, J. Alam, and P. Boucher. Ets system for av+ ec 2015 challenge. In *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*, pages 17–23. ACM, 2015.

[3] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2015.

[4] H. Chen, J. Li, F. Zhang, Y. Li, and H. Wang. 3d model-based continuous emotion recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1836–1845, 2015.

[5] S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56. ACM, 2015.

[6] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conference On Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[8] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, and Y. Zhou. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

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

[10] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[11] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 221–231, 2013.

[12] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels. Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges. In *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*, pages 9–16. ACM, 2015.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101. IEEE, 2010.

[15] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084. IEEE, 2010.

[16] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas. Time-delay neural network for continuous emotional dimension prediction from facial expression sequences. *IEEE transactions on cybernetics*, 46(4):916–929, 2016.

[17] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[18] M. A. Nicolaou, S. Zafeiriou, and M. Pantic. Correlated-spaces regression for learning continuous emotion dimensions. In *Proc. 21st ACM International Conference on Multimedia*, pages 773–776. ACM, 2013.

[19] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proc. 14th ACM International Conference on Multimodal interaction*, pages 501–508. ACM, 2012.

[20] D. Ozkan, S. Scherer, and L. P. Morency. Step-wise emotion recognition using concatenated-hmm. In *Proc. 14th ACM International Conference on Multimodal interaction*, pages 477–484. ACM, 2012.

[21] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. ICME.*, page 5. IEEE, 2005.

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference*, 2015.

[23] J. Posner, JA. Russell, and BS. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005.

[24] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[25] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015.

[26] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proc. 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM, 2015.

[27] A. Savran, H. Cao, A. Nenkova, and R. Verma. Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities. *IEEE Transactions on Cybernetics*, 45(9):1927–1941, 2015.

[28] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proc. 14th ACM International Conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[29] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[30] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In *Proc. 14th ACM International Conference on Multimodal Interaction*, pages 493–500. ACM, 2012.

[31] Y. I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[33] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010.

[34] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.