sensors

MDPI

Article

Multi-Modal Residual Perceptron Network for Audio–Video Emotion Recognition

Xin Chang * and **Władysław Skarbek**

Institute of Radioelectronics and Multimedia Technology, Warsaw University of Technology, 00-665 Warsaw, Poland; wladyslaw.skarbek@pw.edu.pl
* Correspondence: xin.chang.dokt@pw.edu.pl

Abstract: Emotion recognition is an important research field for human–computer interaction. Audio–video emotion recognition is now attacked with deep neural network modeling tools. In published papers, as a rule, the authors show only cases of the superiority in multi-modality over audio-only or video-only modality. However, there are cases of superiority in uni-modality that can be found. In our research, we hypothesize that for fuzzy categories of emotional events, the within-modal and inter-modal noisy information represented indirectly in the parameters of the modeling neural network impedes better performance in the existing late fusion and end-to-end multi-modal network training strategies. To take advantage of and overcome the deficiencies in both solutions, we define a multi-modal residual perceptron network which performs end-to-end learning from multi-modal network branches, generalizing better multi-modal feature representation. For the proposed multi-modal residual perceptron network and the novel time augmentation for streaming digital movies, the state-of-the-art average recognition rate was improved to 91.4% for the Ryerson Audio–Visual Database of Emotional Speech and Song dataset and to 83.15% for the Crowd-Sourced Emotional Multi Modal Actors dataset. Moreover, the multi-modal residual perceptron network concept shows its potential for multi-modal applications dealing with signal sources not only of optical and acoustical types.

Keywords: emotion recognition; deep neural network; multi-modal classifier; deep features fusion; audio sensor; video sensor

check for updates

Citation: Chang, X.; Skarbek, W. Multi-Modal Residual Perceptron Network for Audio–Video Emotion classification. *Sensors* **2021**, *21*, 5452. https://doi.org/10.3390/s21165452

Academic Editor: Stefano Berretti

Received: 7 July 2021
Accepted: 10 August 2021
Published: 12 August 2021

1. Introduction

This paper presents a novel end-to-end Deep Neural Network (DNN) based framework, as Figure 1 illustrates, addressing the Audio–Video Emotion Recognition (AVER) problem. Just like human beings understand emotional expressions in our daily social activates through multi-senses (e.g., visual, vocal, textually meaningful), neural computing units as part of intelligent artificial sensors are now playing important roles in emotion recognition tasks. Specialized sensors in Human–Computer Interaction (HCI) capture information responsible for understanding visual and vocal information just like we—the human beings—understand emotional expressions through our multi-senses.

1.1. Emotion Recognition from Face Expression and Voice Timbre

Intuitively, functionalities of intelligent artificial neurons are assigned with concepts similar to our brain cells, processing information from the very raw senses independently and appropriately to their types.

Visually, information captured by the camera is distributed to several frames as Figure 2 shows. Discrete information in a single frame is firstly delivered to pattern extracting intelligent sensors for features such as Fisherfaces and Eigenfaces [1] or the deep features from Convolutional Neural Networks (CNNs) [2]. To fully preserve the information from the discrete signals, some Sequence Aggregation Component (SAC), e.g., Long Short Term Memory (LSTM) [3] or Transformer [4], is then needed to further process the extracted features. Finally, a classifier such as Support Vector Machine (SVM) or some neural dense layer takes the integrated features for the classification.

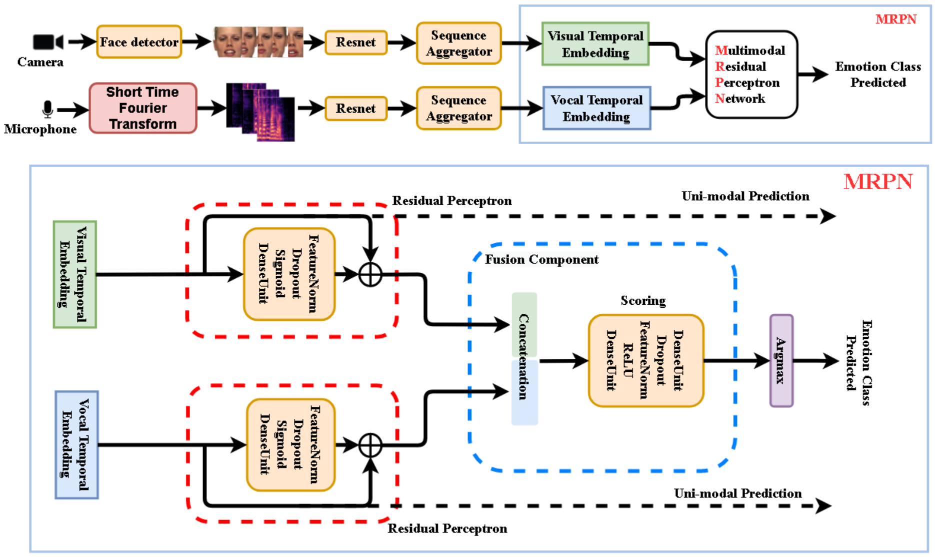

Figure 1. The proposed multi-modal emotion recognition system using Deep Neural Network (DNN) approach. **Upper part:** Video frames and audio spectral segments get independent temporal embeddings to be fused by our multi-modal Residual Perceptron Network (MRPN). **Lower part:** MRPN performs in each modality normalizations via the proposed residual perceptrons and then scores their concatenated outputs in the fusion component. The uni-modal prediction branches are only active in training mode.

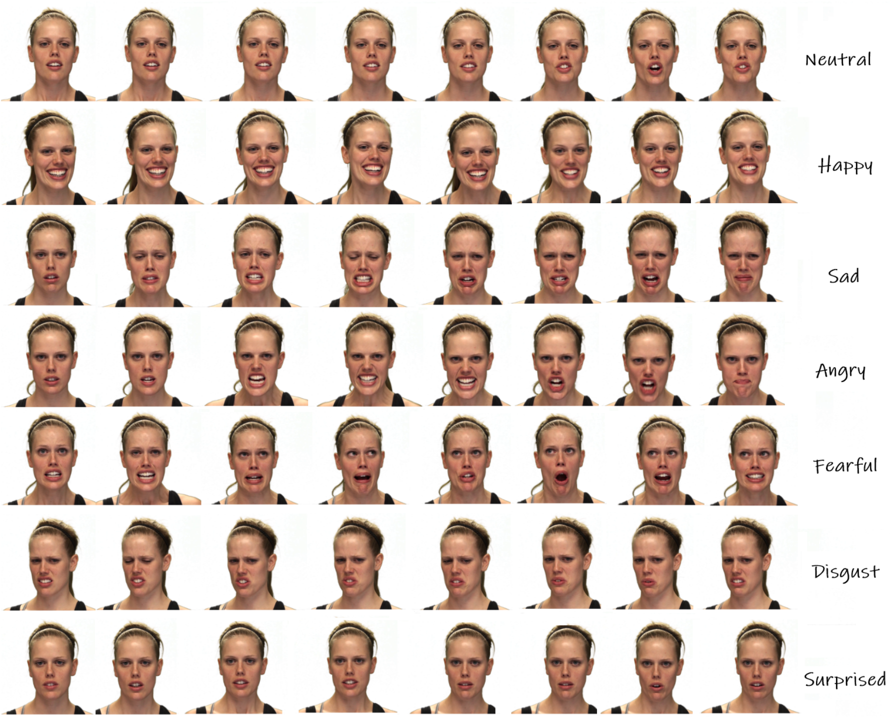

Figure 2. Video frames of visual facial expressions selected from RAVDESS (Ryerson Audio–Visual Database of Emotional Speech and Song) dataset.

Raw vocal inputs are usually with 10,000 to 44,100 samples per second, while the visual frame rate is about 25–30 image frames per second. While raw digital signals in the time domain approximate the original signal precisely, their spectral representation, e.g., spectrogram frame, mel-spectrogram coefficients, or log mel-spectrogram frame, appeared more effectively for sound recognition. The spectral converted vocal signals have shown significant improvements in many classification problems, in spite of some limitations. Expression events do not last at the same time, thus the width of the spectrogram frames changes, which is not desirable for CNN pattern extractors. Therefore, the extracted features also need further processing from SAC which outputs integrated features. Figure 3 shows the expression events from different categories and with different time duration.

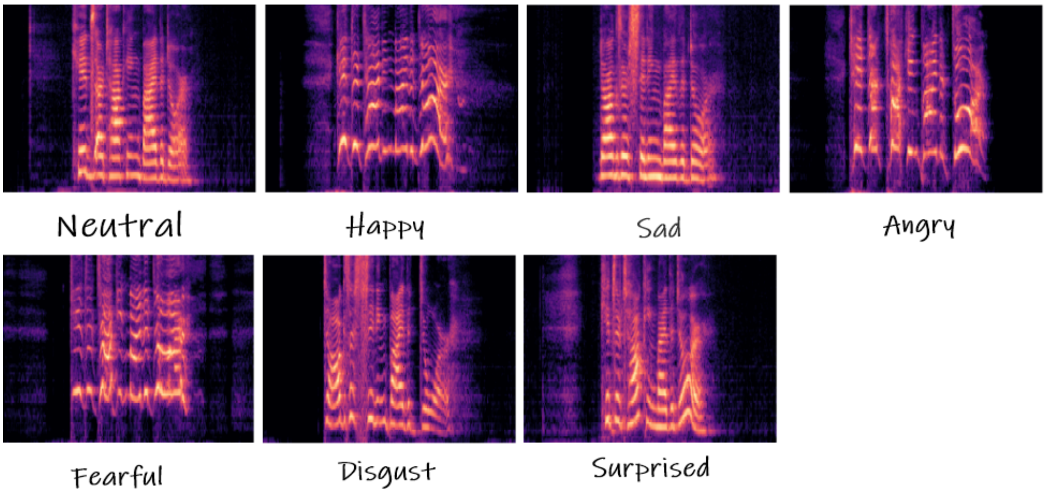

Figure 3. Mel spectrograms of vocal timbres selected from RAVDESS (Ryerson Audio–Visual Database of Emotional Speech and Song) dataset.

1.2. Multi-Modal Emotion Recognition

AVER solution also follows the sensation of human beings; people claim they hear the sound when looking at the sheet music, smell the odor when recalling the memory from a photo or see the sea sight from the smell of the air. The multi-sensation information is processed by different areas of our cerebral cortex, movement, hearing, seeing, etc., then highly correlated by some other brain areas. Thus, the decision made depends not just on the recognition of uni-modal sensations independently but also jointly.

The learning process of neural sensors should also mimic our learning process. The neurons shape their weights just like our cerebral cortex changes from the stimulation of the environments and looks for any correlation of them during the learning process in the supervised neural network sensors training. However, we claim the existing late fusion and end-to-end training strategies have their own advantages but also deficiencies.

1.3. Paper Contribution and Structure

This paper shows by experiments the deficiencies of two training strategies: Late fusion and end-to-end. The late fusion strategy takes trained static uni-modal networks and trains their fusing network components. The end-to-end strategy trains all the multi-modal and uni-modal components together. A novel architecture is proposed to take advantage of both solutions and avoid their side effects, respectively. We demonstrate superiority in the novel end-to-end mechanism and architecture comparing with the naive fusion mechanism in either late fusion or end-to-end training. The proposed DNN framework, data augmentation procedures, and network optimization strategy are discussed. A detailed analysis and discussion are presented by computing experiments on The Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS) and Crowd-Sourced Emotional Multimodal Actors Dataset (Crema-d) datasets. Our major contributions are as follows:

1. Multi-modal framework: We propose a novel within-modality Residual Perceptrons (RP) for efficient gradient blending for the neural network optimization using multi-term loss function in MRPN. The sub-networks and target loss functions produce superior parameterized multi-modal features, preserving the original knowledge of uni-modalities, which impedes inter-modal learning. The within-modality RP components reduce the side effects brought from such multi-term loss functions. As the result, we got significantly better performance over direct strategies including late fusion and end-to-end without MRPN.
2. Time Augmentation of input frames: We demonstrate data augmentation in time involving randomly slicing over input frame sequences from both modalities improved the recognition performance to the state-of-art, even without MRPN. We show also the results that time augmentation does not solve the cases where uni-modal solutions are better than multi-modal solutions, yet solved by MRPN.

2. Related Work

2.1. Superiority in Multi-Modal Approach

Many have shown significant improvement of multi-modal solutions. N. Neverova et al. [5] suggest gradual fusion involving the random dropping of separate channels, and this method was adopted by V. Vielzeuf et al. [6] in the AVER solution for their best result.

Fusion at the early or late stage is discussed by others. R. Beard et al. [7] proposed multi-modal feature fusion at the late stage, while E. Ghaleb et al. [8] try to project features to a shared space in the early stage and provided external loss functions to minimize the distance of features from different modalities. A. Zadeh et al. [9] proposed Multi-view Gated Memory to gate the multi-modal knowledge from LSTM in the time series. E. Mansouri-Benssassi and J. Ye [10] archive early fusion by creating distinct multi-modal neuron groups.

S. Zhang et al. [11] take features from CNN and 3D-CNN models for vocal and visual sources then make global averaging as video features. NC. Ristea et al. [12] take the features extracted by CNNs from both modalities and exploit the fused features for classification purposes. E. Tzinis et al. [13] take cross-modal and self-attention modules. Y. Wu et al. [14] localize events crossing modalities. E. Ghaleb et al. [15] suggest multi-modal emotion recognition metric learning to create a robust representation for both modalities.

2.2. Potential Failures in the Existing Solutions

Let us consider the human brain learning process again. Say the information is wrong in some of the sensory stimulation. A child learned an animal looks just like a dog but having the sound of the cat from the manipulated movies and this kid has never learned the dog and cat in a real-life environment. He will either see a dog and tell it is a cat or hear the cat sound and tell it is a dog. The situation can go even worse if the stimulation he learned from is also fuzzy within their own sensation.

His recognition is still intact to some extent that he can sometimes correctly recognize the visual or acoustic information pattern. However, the recognized information is distorted, along with the correlation of the inter-modal information. This made the distorted uni-modal knowledge he learned having also a negative impact on the other. The same concept we address to the current multi-modal neural network solutions. The within-modal and inter-modal noisiness of the learned pattern both contribute to the wrong recognition. Despite many advantages from the multi-modal solutions which boost the recognition performance of emotion recognition tasks, we hypothesize the uncontrolled fusion strategy, adopted by [6,7,9,11,14,16–18] could lead to potential deficiencies in either late fusion or end-to-end training strategy.

Though many have shown superior performance of late fusion strategy [19–21], for instance, for audio events detection in video material, W. Wang et al. [22] illustrate the results of naive fusion from multi-modal features can be worse than the best uni-modal approach. They propose blending the gradient flow by multi-task loss functions, which is referred to as multi-term loss function by us, from uni-modalities and multi-modality, which help better parameterization of the whole system in many other research areas. Though they suggest benefits from blending the gradient flows, multi-tasking could make the features hard to be optimized serving both uni-modal and multi-modal purposes suggested by many researchers [23–25]. We demonstrate how this proposal can still fail in some inferior cases but is solved by the within-modal RP component in MRPN.

3. Hypothesis

In this section, we discuss our hypothesis where fuzzy information from the uni-modalities can cause chaos in not just the uni-modal neurons but also the correlation neurons, namely the fusion component.

3.1. Within-Modal Information Can Be Missing or Fuzzy

The missing or fuzzy information can be noticed in either visual or vocal modality emotion recognition solution, and then the success rate of recognition can not be increased noticeably. Missing information refers to feature data where emotion categories are confused with neutral categories in the uni-modality. Fuzzy information stands for feature data where one emotion category cannot be distinguished from another one in the uni-modality.

Namely, the visual modality results from the challenge FER-2013 [26] for single image facial recognition have only improved about 4% to 76.8% over the past eight years by W. Wang et al. [27].

Moreover, for video frames, HW. Ng et al. [28] got 47.3% validation accuracy and 53.8% testing accuracy on EmotiW dataset [29] using transfer learning and averaged temporal deep features. Similarly, for vocal solutions, the results for Interactive Emotional Dyadic Motion Capture dataset (IEMOCAP) [30] with the raw inputs are reported around 76% by S. Kwon [31] and 64.93% by S. Latif et al. [32]. The recognition rate for these cases is far from optimal.

Apart from the design, functionality, and training of the neural network, the human voting for those datasets draws our concerns. As the teacher in supervised learning, almost all datasets related to emotion recognition have unsatisfactory knowledge. The ones who understand human emotions the best, the human beings themselves, cannot make a majority of the agreement to the author's labeling. On average, the human rate of the emotional categories is 72% from IEMOCAP, human accuracy on FER-2013 [26] is 65 ± 5%, Crema-d [33] holds the accuracy of 63.6% and RAVDESS [34] has the results of 72.3%.

All the reports point out that in every uni-modality, the information of data is never crystal, thus the learned knowledge of a uni-modality in emotion recognition, can be corrupted and uncontrolled by the network. We cannot identify or agree on which samples are wrong because the boundaries of the clusters are quite subjective.

3.2. End-to-End Modeling for Multi-Modal Data Can Be Distorted

Multi-modal solutions seem to find more generalized patterns via the extension of parameters. However, the fused features have left a backdoor for distorted pattern learning, the side effects are concealed by its benefits.

Figure 4 shows us the different clustering from shuffled train/validation sub-datasets in each uni-modal solution and multi-modal solution, respectively. The actors in the validation set are different than they are in the training set. Due to the missing and fuzzy information from the uni-modal data, the clustering of the same uni-modality shown in the top and bottom panels differs. It can be seen from the figure, for the same modality solutions, either from uni-modal or multi-modal, missing information is causing the overlapping clustering for the neutral category. The cases appear similarly for the overlapping of emotional categories caused by the fuzzy information. This suggests that patterns within uni-modality are hard to be generalized, aligned with human voting results mentioned.

Under such circumstances, we do not know which training sample is fuzzy in which modality, not just it causing the fuzzy direction of within-modal learning, but also inter-modal learning in end-to-end training, i.e., information in modality A is fuzzy while crystal in modality B, can results in correct learning for modality B yet fuzzy learning for modality A. In the end, the distribution of the wrong direction learned knowledge is unknown.

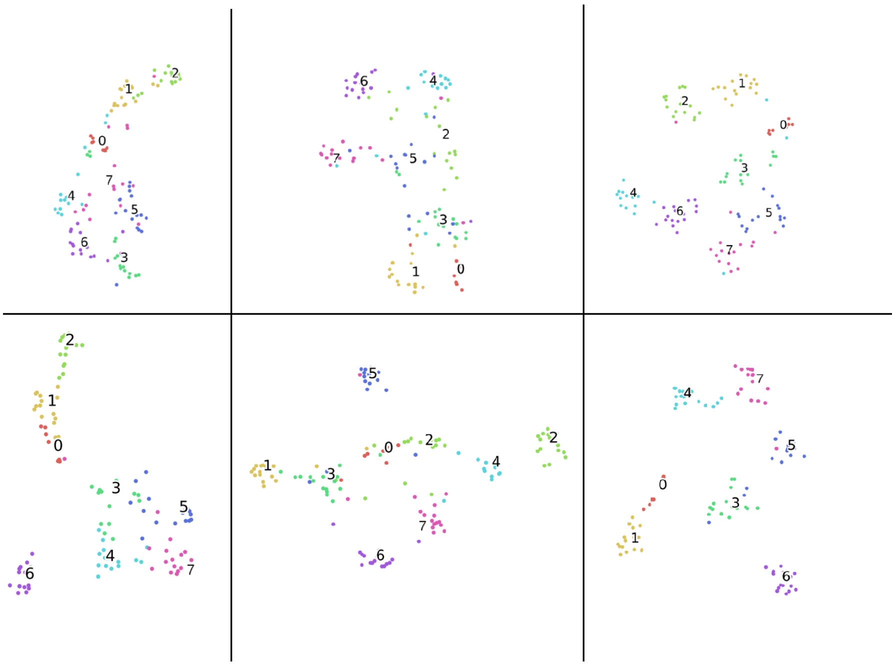

Figure 4. Visualization (t-SNE algorithm) of deep features clustering from two different setups where the train/validation sets are shuffled. The clustering with respect to emotion classes are listed. 0: Neutral; 1: Calm; 2: Happy; 3: Sad; 4: Angry; 5: Fearful; 6: Disgust; 7: Surprised. **Top part**: Clustering results from one setup of uni-modalities and multi-modality. Left part: Only image modality. Middle part: Only audio modality. Right part: Multi-modality. **Down part**: Clustering results from the other setup where train/validation sets are shuffled.

Figure 5 illustrates the source of deficiency in the architecture during the gradient backpropagation, wherein the blue frame denotes the fusion component. Namely, the concatenation unit of the features from different modalities can backpropagate the gradients modifying jointly weights in each modality, potentially distort the knowledge in some modality.

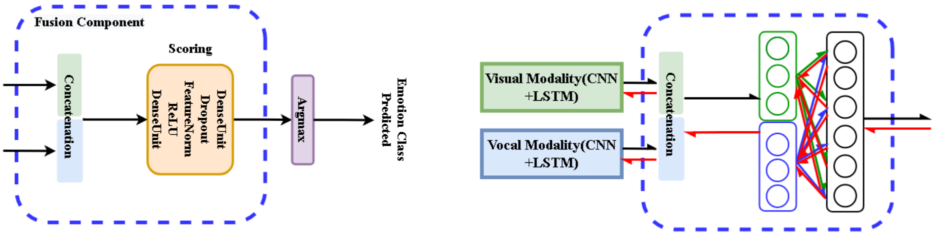

Figure 5. Distorted gradients backpropagation in some modality since the gradients from fused layer makes impact on gradients flow into neural weights of both modalities.

3.3. Late Fusion Modeling for Multi-Modal Data Can Be Insufficient

The late fusion seems to prevent the inter-modal learning of the system, however, not just the distribution of the fuzzy information to each modality is unknown, but also the clean data which holds the highly correlated information between modalities are. If the samples contain clean information in all modalities, then the frozen parameters of the shadow layers cannot make proper adjustments to learn inter-modal information from the joint gradient flow.

4. Proposed Methods

Addressing the mentioned issues, we proposed a novel MRPN along with a multi-term loss function for the better parameterization of the whole network taking advantage of both late fusion and end-to-end strategies while avoiding their deficiencies. MRPN can eliminate the problems without assuming the data is noisy or clean.

4.1. Functional Description of Analyzed Networks

The functional descriptions of the analyzed deep networks are presented for their training mode (see Figure 6). They are based on the selected functionalities of neural units and components. We use index m for inputs of any modality. In our experiments $m = v$ or $m = a$.

1. F_m: Feature extractor for input temporal sequence x_m of modality m, e.g., F_v for video frames x_v, F_a for audio segments x_a.
2. A_m: Aggregation component SAC for temporal feature sequence leading to temporal feature vector f_m, e.g., A_v, A_a for video and audio features, respectively.

$$f_m \doteq A_m(F_m(x_m)) \longrightarrow f_v \doteq A_v(F_v(x_v)),\; f_a \doteq A_a(F_a(x_a)) \tag{1}$$

3. Standard computing units: `DenseUnit`—affine (a.k.a. dense, full connection); `Dropout`—random elements dropping for model regularizing; `FeatureNorm`—normalization for learned data regularizing (batch norm is adopted in the current implementation); and `Concatenate`—joining feature maps; `ReLU, Sigmoid`—activation units.
4. `Scoring`—component mapping feature vectors to class scores vector, usually composing the following operations:

$$\rightarrow DenseUnit \rightarrow ReLU \rightarrow FeatureNorm \rightarrow DenseUnit \tag{2}$$

$$\begin{aligned} &m \in \{v, a\},\; \hat{f}_m \doteq FeatureNorm(f_m) \longrightarrow s_m \doteq Scoring(\hat{f}_m) \\ &g_{va} \doteq FeatureNorm(Concatenate(g_v, g_a)) \longrightarrow s_{va} \doteq Scoring(g_{va}) \end{aligned} \tag{3}$$

5. `FusionComponent`—concatenates its inputs g_v, g_a, then makes the statistical normalization, and finally produces the vector of class scores:

$$\begin{aligned} &s_{va} \doteq FusionComponent(g_v, g_a) \longrightarrow \\ &g_v, g_a \rightarrow Concatenate \rightarrow Scoring \rightarrow s_{va} \end{aligned} \tag{4}$$

   In our networks g_v, g_a are statistically normalized multi-modal features $(\hat{f}_v, \hat{f}_a)$ or their residually updated form (f'_v, f'_a)—cf. those symbols in Figure 6.
6. `SoftMax`—computing unit for normalization of class scores to class probabilities:

$$\begin{aligned} &m \in \{v, a\} \longrightarrow p_m \doteq Softmax(s_m) \\ &p_{va} \doteq Softmax(s_{va}) \end{aligned}$$

7. `CrossEntropy`—a divergence of probability distributions used as loss function. Let p is the target probability distribution. Then the following loss functions are defined:

$$\begin{aligned} &m \in \{v, a\},\; p_m \doteq Softmax(s_m) \longrightarrow \mathcal{L}_m \doteq CrossEntropy(p, p_m) \\ &p_{va} \doteq Softmax(s_{va}) \longrightarrow \mathcal{L}_{va} = CrossEntropy(p, p_{va}) \\ &\mathcal{L} \doteq \mathcal{L}_v + \mathcal{L}_a + \mathcal{L}_{va} \end{aligned} \tag{5}$$

   where $\mathcal{L}$ is multi-term loss function implying the gradient blending in the backpropagation stage.

8. `ResPerceptron` (Residual Perceptron)—component performing statistical normalization for the dense unit (perceptron) computing residuals for normalized data. In our solution it transforms a modal feature vector f_m into f'_m, as follows:

$$\begin{aligned} &\hat{f}_m \doteq FeatureNorm(f_m) \longrightarrow f'_m \doteq ResPerceptron(\hat{f}_m) \longrightarrow \\ &f'_m \doteq \hat{f}_m + FeatureNorm(Sigmoid(DenseUnit(\hat{f}_m))) \end{aligned} \tag{6}$$

Three networks $\mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2$ are defined for further analysis:

1. Network $\mathcal{N}_0(f_v, f_a; p)$ with fusion component and loss function $\mathcal{L}_{va}$:

$$\begin{aligned} &\hat{f}_v \doteq FeatureNorm(f_v),\ \hat{f}_a \doteq FeatureNorm(f_a) \\ &s_{va} \doteq FusionComponent(\hat{f}_v, \hat{f}_a) \\ &p_{va} \doteq SoftMax(s_{va}) \longrightarrow \mathcal{L}_{va} \doteq CrossEntropy(p, p_{va}) \end{aligned} \tag{7}$$

2. Network $\mathcal{N}_1(f_v, f_a; p)$ with fusion component and fused loss function $\mathcal{L} \doteq \mathcal{L}_v + \mathcal{L}_a + \mathcal{L}_{va}$:

$$\begin{aligned} &\hat{f}_v \doteq FeatureNorm(f_v),\ \hat{f}_a \doteq FeatureNorm(f_a) \\ &s_v \doteq DenseUnit(\hat{f}_v),\ s_a \doteq DenseUnit(\hat{f}_a),\ s_{va} \doteq FusionComponent(\hat{f}_v, \hat{f}_a) \\ &p_v \doteq SoftMax(s_v) \longrightarrow \mathcal{L}_v \doteq CrossEntropy(p, p_v) \\ &p_a \doteq SoftMax(s_a) \longrightarrow \mathcal{L}_a \doteq CrossEntropy(p, p_a) \\ &p_{va} \doteq SoftMax(s_{va}) \longrightarrow \mathcal{L}_{va} \doteq CrossEntropy(p, p_{va}) \end{aligned} \tag{8}$$

3. Network $\mathcal{N}_2(f_v, f_a; p)$ with normalized residual perceptron, fusion component and fused loss function $\mathcal{L} \doteq \mathcal{L}_v + \mathcal{L}_a + \mathcal{L}_{va}$:

$$\begin{aligned} &\hat{f}_v \doteq FeatureNorm(f_v),\ \hat{f}_a \doteq FeatureNorm(f_a) \\ &f'_v \doteq ResPerceptron(\hat{f}_v),\ f'_a \doteq ResPerceptron(\hat{f}_a) \\ &s_v \doteq DenseUnit(\hat{f}_v),\ s_a \doteq DenseUnit(\hat{f}_a),\ s_{va} \doteq FusionComponent(f'_v, f'_a) \\ &p_v \doteq SoftMax(s_v) \longrightarrow \mathcal{L}_v \doteq CrossEntropy(p, p_v) \\ &p_a \doteq SoftMax(s_a) \longrightarrow \mathcal{L}_a \doteq CrossEntropy(p, p_a) \\ &p_{va} \doteq SoftMax(s_{va}) \longrightarrow \mathcal{L}_{va} \doteq CrossEntropy(p, p_{va}) \end{aligned} \tag{9}$$

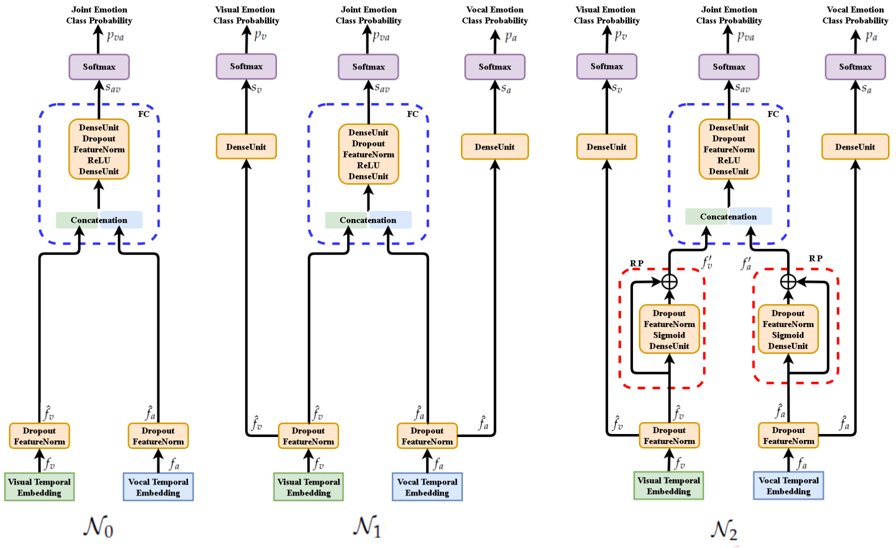

Figure 6. Evolution of network design for multi-modal fusion (presented for training mode). $\mathcal{N}_0$: Fusion component (FC) only. $\mathcal{N}_1$ ([22]): Beside FC, independent scoring of each modality is considered. $\mathcal{N}_2$: Extending $\mathcal{N}_1$ network by Residual Perceptrons (RP) in each modality branch.

For the networks $\mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2$ detailed in Figure 6, we can observe:

1. All instances of FeatureNorm unit are implemented as batch normalization units.
2. In testing mode only the central branch of networks $\mathcal{N}_1, \mathcal{N}_2$ are active while the side branches are inactive as they are used only to compute the extra terms of the extended loss function.
3. The above facts make network architectures $\mathcal{N}_0, \mathcal{N}_1$ equivalent in the testing mode. However, the models trained for those architectures are not the same, as weights are optimized for different loss functions.
4. In the testing mode all Dropout units are not active, as well.
5. The architecture of FusionComponent is identical for all three networks. The difference between models of $\mathcal{N}_0$ and $\mathcal{N}_1$ networks follows from the different loss functions while the difference between models of $\mathcal{N}_1$ and $\mathcal{N}_2$ networks is implied by using ResPerceptron (RP) components in $\mathcal{N}_2$ network.
6. To control the range of affine combinations computed by Residual Perceptron (RP) component, we use Sigmoid activations instead of the ReLU activations exploited in other components. The experiments confirm the advantage of this design decision.
7. The Residual Perceptron (RP) was introduced in the network $\mathcal{N}_2$ to implement better parameterization of within-modal features before their fusion.

4.2. MRPN Components' Role in Multi-Term Optimization

1. As we discussed in the hypothesis section, the late fusion strategy has the advantage of preserving the best information in each uni-modality since the uni-modality extracts generalized deep features which suffer little from the outliers of their own modality, i.e., a small amount of wrongly labeled data in uni-modal solutions will not contribute to the generalized feature patterns; they are "filtered out" by the uni-modal neural network. Thus the additional term of loss functions implies the blended gradient in the shallow layers of each uni-modality, and helped for better parameterization of the features before fusion, preserving the knowledge as uni-modalities are trained respectively. The above facts make the end-to-end strategy suffer less inter-modal information as late fusion does.
2. However, the multi-term optimization can result in extracting inferior uni-modal features as the input to the fusion component. This problem was mentioned in the literature [23–25]. RP is introduced to make modified uni-modal features, instead of storing all knowledge for uni-modal and multi-modal purposes in one unit, causing the clash of loss converging from two directions, the uni-modal and multi-modal knowledge can be stored in the original uni-modal features and modified multi-modal features, creating a new path for the gradient flow. RP can preserve the best of the uni-modal solution while the modified features from the short-cut can still fulfill the purpose of integrating new multi-modal features.

The mentioned two novel properties make MRPN free from side-effects of late fusion and end-to-end strategy while preserving their own advantages.

4.3. MRPN in General Multi-Modal Applications

We suggest that MRPN can be adopted in any multi-modal application that involves many multi-modal inputs and one target function or many multi-modal inputs and many multi-modal target functions as Figure 7 shows. In both cases MRPN benefits from many terms of loss functions as the numbers of uni-modalities, updating the whole system together while avoiding learning from inter-modal fuzzy information. MRPN is general to be compatible with any other proposed mechanism.

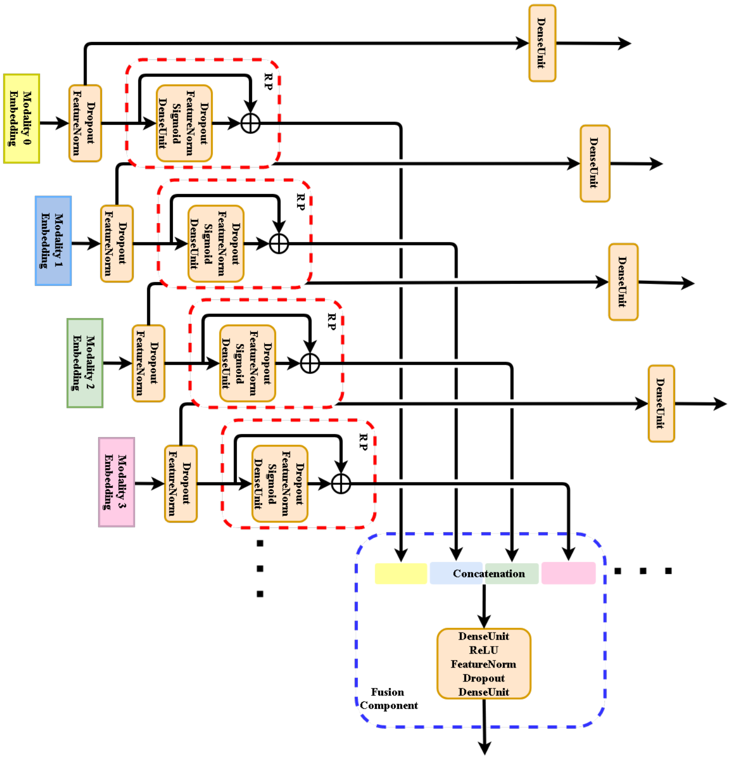

Figure 7. Generalization of our MRPN fusion approach to many modalities. It could be used for either regression or classification applications.

4.4. Pre-Processing

Our data pre-processing includes the procedures for both modality inputs, namely spatial and time-dependent augmentation are applied.

1. Spatial data augmentation for visual frames:
The facial area for the visual input frames is cropped using a CNN solution from Dlib library [35]. Once the facial area is cropped, spatial video augmentation is applied during the training phase. The same random augmentation parameters are applied for all frames of a video source illustrated in Figure 8.
2. Time dependent data augmentation for visual frames:
Obviously, expressions from the same category do not last the same duration. To make our system robust to the inconsistent duration of the emotion events, we perform data augmentation in time by randomly slicing the original frames as Figure 9 illustrates. Such operation should also avoid too few input frames missing information of the expression events. Thus the training segments are selected to have at least one-second duration unless the original duration of the file is less than that.
3. Spatial data augmentation for vocal frames: Raw audio inputs are resampled at 16 kHz and standardized by their mean and standard deviation without any denoising or cutting, to remove influence from the distance of the speaker to the microphone, or subjective base volume of the speaker. The standardized wave is then divided into one-second segments and converted to spectrograms by a Hann windowing function of size 512 and hop size of 64.
The above facts specify the size of the spectrogram at 256 × 250, approaching the required input shape of Resnet-18—the CNN extractor used in our experiment. The chunk size using such inputs for Resnet-18 [36] is close to its desired performance utilizing the advantage in the middle deep features.
4. Time dependent augmentation for vocal frames: Similar to the time augmentation in visual inputs, raw audio inputs are also randomly sliced. The raw data is further

over-sampled in both the training and testing mode by a hopping window. A window of 0.2 s, 1/5 duration of the input segments to the CNN extractor is specified. The oversampling further improved our results by the increasing number of deep features from the output deep feature sequences of CNN to the SAC. The mechanism grants the opportunity for the SAC to investigate more details of the temporal information in the deep feature vectors.

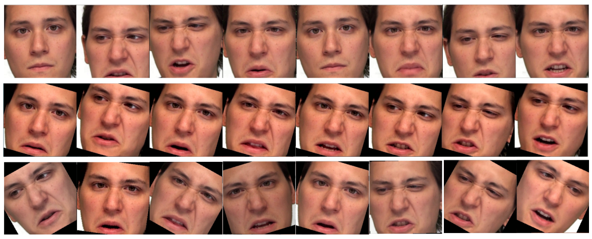

Figure 8. Visual comparison of augmentation procedure for cropped video frames. **Top part**: Original video frames. **Middle part**: Applying random augmentation parameters—same for all frames. **Bottom part**: Applying random augmentation parameters—different for each frame.

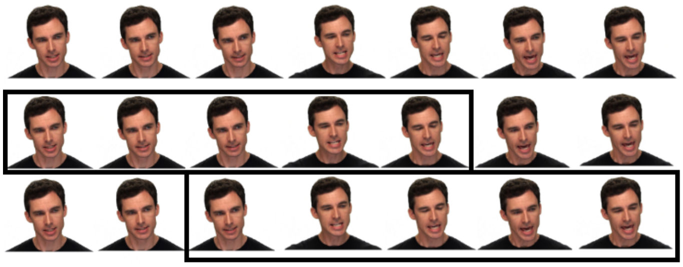

Figure 9. Examples of time dependent augmentation for visual frames. **Top part**: Original frames. **Middle part**: Sliced frames start at beginning of the original frames. **Bottom part**: Sliced frames start at the middle of the original frames.

5. Computational Experiments and Their Discussion

This section presents an evaluation regarding the advantages of our proposed framework and time-dependent augmentation. Two datasets, RAVDESS and Crema-d are employed for this purpose. The improvement of the time-augmentation mechanism is analyzed in the naive fusion model which brought us state-of-art results even without MRPN design. The inferior cases in such common neural multi-modal solutions are detected and discussed in the comparison. Improvement of MRPN is then presented, not just in the detected inferior sub-datasets, but also in general data samples.

5.1. Datasets

RAVDESS and Crema-d differ in numbers of expression categories, total files, identifies, and also video quality.

1. RAVDESS dataset includes both speech and song files. For the speech recognition proposal, we only use the speech files from the dataset. It contains 2880 files, 24 actors (12 female, 12 male), state two lexically identical statements. Speech includes calm, happy, sad, angry, fearful, surprise, and disgusted expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression, in a total of 8 categories. It is the most recent video-audio emotional dataset with the highest video quality in this research area to our best knowledge.
2. Crema-d dataset consists of visual and vocal emotional speech files in a range of basic emotional states (happy, sad, anger, fear, disgust, and neutral). 7442 clips of 91 actors with diverse ethnic backgrounds were rated by multiple raters in three modalities: Audio, visual, and audio-visual.

For both datasets, the training set and testing set are separated using similar concepts as 10-fold inter-validation. Additionally, identities of the actors are also separated in train and val sets to prevent the results from leaning on the actors. Around 10% of the actors are used for validation while the remaining 90% are used for training, male and female actors are balanced in each set. We rotate the split train/validation sub-datasets to get multiple results over the whole dataset. The crema-d dataset has fewer categories for the classification tasks, but from the report of the authors themselves, Crema-d holds the accuracy of 63.6% from human recognition for six categories which are less than RAVDESS at 72.3% for eight categories. The resolution of the video source is verified not the cause of the worse performance. The better results in the RAVDESS dataset, in our opinion, are the more crystal and natural emotion information inside the RAVDESS dataset.

5.2. Model Organization and Computational Setup

The naive fusion model $\mathcal{N}_0$, advanced fusion network $\mathcal{N}_1$, which is equivalent to the Facebook [22] solution and the $\mathcal{N}_2$ (MRPN) have the same CNN extractors at the initial stage of the training. To compare the impact of strategy from features fusion only, CNN extractor architecture is fixed to Resnet-18 [36].

The CNN in visual modality is initialized from a facial image expression recognition task, the challenge FER2013 [26]. As for vocal modality, The CNN is pretrained on the voice recognition task from VoxCeleb dataset [37]. The initialization of the CNN extractors made the whole system much easier to be optimized.

AdaMW optimizer is adopted for the model optimization, with the initial learning rate at 5×10^{-5}, decreased two times if validation loss is not dropping over ten epochs.

5.3. Data Augmentation Cannot Generalize Multi-Modal Feature Patterns

This subsection illustrates the improvement of time-dependent augmentation. The improvement also proves that the inferior case of multi-modal solution does not depend on the with-modal patterns. The single modality solutions in our experiments (shown in Table 1) take pretrained Resnet-18 as extractors and LSTM cells as SACs. The naive multi-modal solution takes twice of the components with an additional fusion layer as Figure 6 illustrates on the left panel. Adopting time-dependent augmentation shows overall performance improvements on either single or multi-modal solutions. The Table notations are presented in the follows:

Table 1. Comparison of single modalities models with $\mathcal{N}_0$ model (RAVDESS cases): VM—Visual Modality only; AM—Audio Modality only; JM—Joint Modalities ($\mathcal{N}_0$ model); T—having time augmentation by signal random slicing; NT—not having time augmentation.

RAVDESS	**A1,2**	**A3,4**	**A5,6**	**A7,8**	**A9,10**	**A11,12**
AM (NT)	70.8%	55.0%	57.5%	74.1%	43.5%	65.8%
AM (T)	71.6%	77.5%	71.6%	90.0%	55.8%	69.1%
VM (NT)	82.5%	70.0%	66.7%	74.1%	80.3%	63.3%
VM (T)	86.6%	75.0%	70.6%	76.6%	87.3%	69.1%
JM (NT)	90.8%	89.1%	85.2%	89.3%	78.5%	85.5%
JM (T)	97.5%	90.3%	87.5%	97.5%	86.5%	87.5%
RAVDESS	**A13,14**	**A15,16**	**A17,18**	**A19,20**	**A21,22**	**A23,24**
AM (NT)	59.8%	57.5%	51.6%	55.5%	55.8%	63.3%
AM (T)	70.0%	69.1%	57.5%	63.3%	68.3%	68.3%
VM (NT)	71.3%	60.0%	63.3%	70.8%	65.8%	70.8%
VM (T)	73.3%	65.0%	64.1%	78.3%	66.6%	74.1%
JM (NT)	77.5%	75.5%	76.3%	85.2%	82.8%	80.0%
JM (T)	82.4%	79.6%	83.2%	89.0%	85.5%	84.2%

In the variational train/val sub-datasets in Table 1, Ax,y stands for the validation files that came from actor x and y, odd number notes for a male actor, and even number for a female actor.

5.4. Discussion on Inferior Multi-Modal Cases

The time augmentation shows overall improvements in either uni-modal or multi-modal approach, yet the inferior case where uni-modal solution better than multi-modal solution still exists, which suggests data augmentation cannot generalize multi-modal features. Only one inferior case is detected in Table 1 of the case A9,10, but we argue such deficiency is common in fuzzy multi-modal data. The pattern learning ability from both modalities is well enough, both solutions have performance over 85% in cases like A7,8 and A1,2. However, the ratio of mismatched learned and target patterns are ranging along with the shuffling of the sub-datasets.

The degeneration of the performance became visible only because the percentage of the pattern mismatched samples has passed some kind of threshold in the training set. If so, by eliminating or reducing such side effects, overall improvements should be expected for any train and testing sub-dataset.

5.5. Improvement of MRPN

This subsection addresses the improvement of MRPN preventing the side-effects in the existing late fusion and end-to-end strategies we hypothesized as Tables 2 and 3 illustrate. The end-to-end strategy of $\mathcal{N}_1$, which takes multi-term loss function helped the better parameterization shows improved average performance over naive end-to-end and late fusion training strategies, yet it can still fail in some cases. Our proposed MRPN on the contrary demonstrates the same performance or most improvement in any circumstance.

Table 2. Comparison for RAVDESS of MRPN approach (network $\mathcal{N}_2$) with late fusion strategy ($\mathcal{N}_0$), end-to-end strategy ($\mathcal{N}_0$), and advanced end-to-end fusion strategy ($\mathcal{N}_1$).

RAVDESS	**A1,2**	**A3,4**	**A5,6**	**A7,8**	**A9,10**	**A11,12**
$\mathcal{N}_0$ (late fusion)	61.6%	92.1%	87.5%	96.6%	66.6%	87.5%
$\mathcal{N}_0$ (end-to-end)	**97.5**%	90.3%	87.5%	**97.5**%	86.5%	87.5%
$\mathcal{N}_1$ (end-to-end)	**97.5**%	89.1%	88.3%	**97.5**%	90.0%	**90.0**%
$\mathcal{N}_2$ (end-to-end)	**97.5**%	**92.1**%	**90.8**%	**97.5**%	**91.4**%	**90.0**%
RAVDESS	**A13,14**	**A15,16**	**A17,18**	**A19,20**	**A21,22**	**A23,24**
$\mathcal{N}_0$ (late fusion)	80.8%	85.0%	81.6%	87.5%	86.6%	65.8%
$\mathcal{N}_0$ (end-to-end)	82.4%	79.6%	83.2%	89.0%	85.5%	84.2%
$\mathcal{N}_1$ (end-to-end)	77.5%	89.1%	86.6%	92.5%	89.1%	**90.6**%
$\mathcal{N}_2$ (end-to-end)	**84.3**%	**89.7**%	**89.8**%	**93.3**%	**90.6**%	**90.6**%

Table 3. Comparison for Crema-d of MRPN approach (network $\mathcal{N}_2$) with simple fusion strategy ($\mathcal{N}_0$), and advanced fusion strategy ($\mathcal{N}_1$).

Crema-d	**S1**	**S2**	**S3**	**S4**	**S5**
$\mathcal{N}_0$ (late fusion)	76.5%	79.9%	76.6%	62.3%	78.2%
$\mathcal{N}_0$ (end-to-end)	77.3%	81.3%	79.2%	74.8%	78.6%
$\mathcal{N}_1$ (end-to-end)	72.6%	82.3%	77.3%	74.8%	74.2%
$\mathcal{N}_2$ (end-to-end)	**79.5**%	**83.0**%	**83.0**%	**76.8**%	**81.9**%
Crema-d	**S6**	**S7**	**S8**	**S9**	
$\mathcal{N}_0$ (late fusion)	81.8%	78.8%	80.0%	77.5%	
$\mathcal{N}_0$ (end-to-end)	**82.0**%	75.1%	79.5%	77.5%	
$\mathcal{N}_1$ (end-to-end)	**82.0**%	74.8%	79.3%	75.8%	
$\mathcal{N}_2$ (end-to-end)	**82.0**%	**80.0**%	**80.5**%	**78.6**%	

It can been seen from the confusion matrices in Figures 10 and 11 the averaged improvements of $\mathcal{N}_2$ (MRPN) over the late fusion and end-to-end $\mathcal{N}_0$ models. Performance on some specific categories shows a slight decrease for MRPN, especially for the categories of calm and neutral expressions because they are naturally close to each other in the RAVDESS dataset. $\mathcal{N}_1$ does not always perform better than the existing solutions, the almost 6% improvements of $\mathcal{N}_2$ (MRPN) over $\mathcal{N}_1$ suggests the level of data fuzziness can make the end-to-end multi-term optimization even harder without proposed RP components. The overall improvements suggest that multi-modal patterns are more generalized from the solution of $\mathcal{N}_2$ (MRPN).

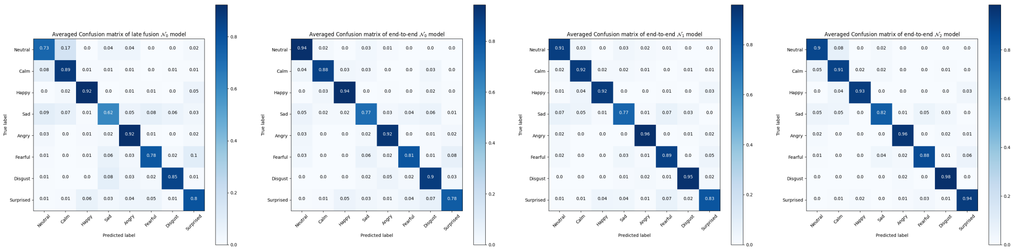

Figure 10. Averaged confusion matrices of tested models for RAVDESS dataset.

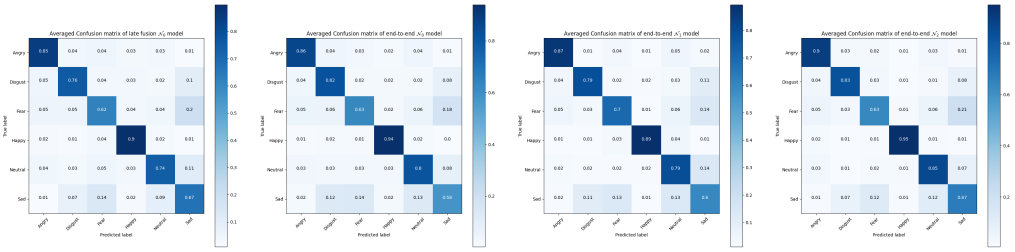

Figure 11. Averaged confusion matrices of tested models for Crema-d dataset.

5.6. Comparing Baseline with SOTA

Our proposed MRPN shows stat-or-art results on both datasets. It has no conflicts with any potential advantages from another novel mechanism. The detailed results are presented in Table 4. Experiments regarding the pretraining of the CNN extractors and the time augmentation have made the network robust to overcome the overfitting issues regarding the small amount of the training and testing data.

Additionally, replacement of LSTM of Bidirectional LSTM and Transformer as aggregator has been conducted but no noticeable differences of them as sequence aggregators can be seen. As for Transformer, the average feature is taken from the decoded outputs, the concept follows from Vision Transformer (VIT) [38].

Table 4. Comparison of our fusion models with others recent solutions. Options used: IA—image augmentation; WO—without audio overlapping; VA—video frames augmentation; and AO—audio overlapping; X symbol—there is no report from authors for the given dataset.

Model (Our)	**RAVDESS**	**Crema-d**
$\mathcal{N}_0$ (end-to-end), Resnet18 + LSTM, IA	83.20%	77.25%
$\mathcal{N}_0$ (end-to-end), Resnet18 + LSTM, VA + WO	85.20%	79.25%
$\mathcal{N}_0$ (late fusion), Resnet18 + LSTM, VA + AO	81.6%	76.84%
$\mathcal{N}_0$ (end-to-end), Resnet18 + LSTM, VA + AO	87.55%	81.30%
$\mathcal{N}_1$ (end-to-end), Resnet18 + LSTM, VA + AO [22]	89.8%	77.0%
MRPN (end-to-end), Resnet18 + LSTM, VA + AO	**90.8%**	**83.00%**
MRPN (end-to-end), Resnet18 + Transformer(avg), VA + AO	**91.4%**	**83.15%**

Table 4. *Cont.*

Model (Others)	RAVDESS	Crema-d
(OpenFace/COVAREP features + LSTM) + Attention [7]	58.33%	65.00%
Dual Attention + LSTM [8]	67.7%	74.00%
Resnet101 + BiLSTM [39]	77.02%	X
custom CNN [12]	X	69.42%
Early Cross-modal + MFCC + MEL spectrogram [10]	83.6%	X
CNN + Fisher vector + Metric learning [15]	X	66.5%
custom CNN + Spectrogram [31]	79.5% (Audio)	X

6. Conclusions

This paper focuses on explaining the potential deficiencies in the existing fusion layer of the multi-modal approach to AVER tasks using late fusion or end-to-end strategy. The proposed MPRN architecture along with the multi-term loss function makes superior fused features from multi-modal sources. We observe the elimination of inferior cases of multi-modal solutions with respect to uni-modal solutions.

Our results achieve an average accuracy of 91.4% on the RAVDESS dataset and 83.15% on the Crema-d dataset. MRPN solution contributes to a better average recognition rate of approximately 2%. We have observed the maximum improvement of MRPN for a subset to be around 90% from nearly 80%.

The proposed data pre-processing by time augmentation makes general overall rate improvements for both the uni-modal and multi-modal data. It also illustrates data augmentation cannot generalize multi-modal features due to the deficiencies in the existing multi-modal solutions.

Moreover, the MRPN concept shows its potential for multi-modal classifiers dealing with signal sources not only of optical and acoustical type.

Author Contributions: Conceptualization, X.C. and W.S.; Data curation, X.C.; Formal analysis, X.C. and W.S.; Investigation, X.C. and W.S.; Methodology, X.C. and W.S.; Project administration, X.C. and W.S.; Resources, X.C. and W.S.; Software, X.C. and W.S.; Supervision, W.S.; Validation, X.C. and W.S.; Visualization, X.C. and W.S.; Writing—original draft, X.C. and W.S.; Writing—review & editing, X.C. and W.S. All authors have read and agreed to the published version of the manuscript.

Funding: Beside statutory support of Warsaw University of Technology this research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The initialization of visual CNN extractor were sampled from the FER-2013 challenge [26]. This data was made available under the Open Database License and can be found at https://www.kaggle.com/msambare/fer2013 (accessed on 7 July 2021). The initialization of vocal CNN extractor were sampled from VoxCeleb dataset [37]. This data was made available under the Open Database License and can be found at https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox1.html (accessed on 7 July 2021). The evaluation of the MRPN and time augmentation is based on RAVDESS dataset [34] and Crema-d [33] under the Open Database License and can be found at https://zenodo.org/record/1188976#.YOOv3Oj7SUk (accessed on 7 July 2021) and https://github.com/CheyneyComputerScience/CREMA-D (accessed on 7 July 2021) respectively.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AVER	Audio–video emotion recognition
CNN	Convolution Neural Network
Crema-d [33]	Crowd-sourced Emotional multi-modal Actors Dataset
DNN	Deep Neural Network
HCI	Human–Computer Interaction
FC	Fully connected layer
IEMOCAP [30]	Interactive emotional dyadic motion capture dataset
LSTM	Long Short Term Memory
MRPN	multi-modal Residual Perceptron Network
RAVDESS [34]	The Ryerson Audio–Visual Database of Emotional Speech and Song
RP	Residual Perceptron
SAC	Sequence Aggregation Component
SOTA	State of the Art Solution
STFT	Short-term Fourier transformation
SVM	Support Vector Machine
VIT [38]	Vision Transformer

References

1. Belhumeur, P.N.; Hespanha, J.a.P.; Kriegman, D.J. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 711–720. [CrossRef]
2. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
3. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
5. Neverova, N.; Wolf, C.; Taylor, G.W.; Nebout, F. ModDrop: Adaptive multi-modal gesture recognition. *arXiv* **2015**, arXiv:1501.00102.
6. Vielzeuf, V.; Pateux, S.; Jurie, F. Temporal Multimodal Fusion for Video Emotion Classification in the Wild. *arXiv* **2017**, arXiv:1709.07200.
7. Beard, R.; Das, R.; Ng, R.W.M.; Gopalakrishnan, P.G.K.; Eerens, L.; Swietojanski, P.; Miksik, O. Multi-Modal Sequence Fusion via Recursive Attention for Emotion Recognition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 251–259.
8. Ghaleb, E.; Popa, M.; Asteriadis, S. Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3–6 September 2019; pp. 552–558.
9. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory Fusion Network for Multi-view Sequential Learning. *arXiv* **2018**, arXiv:1802.00927.
10. Mansouri-Benssassi, E.; Ye, J. Speech Emotion Recognition With Early Visual Cross-modal Enhancement Using Spiking Neural Networks. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
11. Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Tian, Q. Learning Affective Features with a Hybrid Deep Model for Audio–Visual Emotion Recognition. *IEEE Trans. Cir. Syst. Video Technol.* **2018**, *28*, 3030–3043. [CrossRef]
12. Ristea, N.; Duţu, L.C.; Radoi, A. Emotion Recognition System from Speech and Visual Information based on Convolutional Neural Networks. In Proceedings of the 2019 International Conference on Speech Technology and Human–Computer Dialogue (SpeD), Timisoara, Romania, 10–12 October 2019; pp. 1–6.
13. Tzinis, E.; Wisdom, S.; Remez, T.; Hershey, J.R. Improving On-Screen Sound Separation for Open Domain Videos with Audio–Visual Self-attention. *arXiv* **2021**, arXiv:2106.09669.
14. Wu, Y.; Zhu, L.; Yan, Y.; Yang, Y. Dual Attention Matching for Audio–Visual Event Localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October – 2 Nov 2019.
15. Ghaleb, E.; Popa, M.; Asteriadis, S. Metric Learning-Based Multimodal Audio–Visual Emotion Recognition. *IEEE MultiMedia* **2020**, *27*, 37–48. [CrossRef]
16. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Audio–Visual Emotion Recognition in Video Clips. *IEEE Trans. Affect. Comput.* **2019**, *10*, 60–75. [CrossRef]
17. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]

18. Ma, F.; Zhang, W.; Li, Y.; Huang, S.L.; Zhang, L. Learning Better Representations for Audio–Visual Emotion Recognition with Common Information. *Appl. Sci.* **2020**, *10*, 7239. [CrossRef]
19. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *arXiv* **2016**, arXiv:1608.00859.
20. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. *arXiv* **2014**, arXiv:1406.2199.
21. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *arXiv* **2018**, arXiv:1705.07750.
22. Wang, W.; Tran, D.; Feiszli, M. What Makes Training Multi-Modal Classification Networks Hard? In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 12692–12702.
23. Standley, T.; Zamir, A.R.; Chen, D.; Guibas, L.; Malik, J.; Savarese, S. Which Tasks Should Be Learned Together in Multi-task Learning? *arXiv* **2020**, arXiv:1905.07553.
24. Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
25. Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; Finn, C. Gradient Surgery for Multi-Task Learning. *arXiv* **2020**, arXiv:2001.06782.
26. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. *Neural Netw.* **2015**, *64*, 59–63. [CrossRef]
27. Wang, W.; Fu, Y.; Sun, Q.; Chen, T.; Cao, C.; Zheng, Z.; Xu, G.; Qiu, H.; Jiang, Y.G.; Xue, X. Learning to Augment Expressions for Few-shot Fine-grained Facial Expression Recognition. *arXiv* **2020**, arXiv:2001.06144.
28. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 443–449.
29. Dhall, A.; Kaur, A.; Goecke, R.; Gedeon, T. EmotiW 2018: Audio–Video, Student Engagement and Group-Level Affect Prediction. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 653–656.
30. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower Provost, E.; Kim, S.; Chang, J.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
31. Mustaqeem; Kwon, S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors* **2020**, *20*, 183. [CrossRef] [PubMed]
32. Latif, S.; Rana, R.; Qadir, J.; Epps, J. Variational Autoencoders for Learning Latent Representations of Speech Emotion. *arXiv* **2017**, arXiv:1712.08708.
33. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [CrossRef]
34. Livingstone, S.R.; Russo, F.A. The Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
35. King, D.E. Dlib-Ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
37. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. 2017. Available online: https://arxiv.org/pdf/1706.08612.pdf (accessed on 12 August 2021).
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
39. Mustaqeem; Sajjad, M.; Kwon, S. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875. [CrossRef]