Deep spatio-temporal features for multimodal emotion recognition

Dung Nguyen, Kien Nguyen, Sridha Sridharan, Afsane Ghasemi, David Dean and Clinton Fookes
Speech, Audio, Image and Video Technology (SAIVT) Laboratory
Queensland University of Technology, Brisbane, Australia
{d.nguyentien, k.nguyenthanh, s.sridharan, afsaneh.ghasemighalehbahmani, d.dean, c.fookes}
@qut.edu.au

Abstract

Automatic emotion recognition has attracted great interest and numerous solutions have been proposed, most of which focus either individually on facial expression or acoustic information. While more recent research has considered multimodal approaches, individual modalities are often combined only by simple fusion at the feature and/or decision-level. In this paper, we introduce a novel approach using 3-dimensional convolutional neural networks (C3Ds) to model the spatio-temporal information, cascaded with multimodal deep-belief networks (DBNs) that can represent the audio and video streams. Experiments conducted on the eNTERFACE multimodal emotion database demonstrate that this approach leads to improved multimodal emotion recognition performance and significantly outperforms recent state-of-the-art proposals.

1. Introduction

The emotions of humans manifest in their facial expressions, voice, gestures, and posture. A system that could automatically recognize human emotions using one or more of these modalities would play an important role in a wide range of applications such as video games, human computer interaction, robots, educational software, animations, automobile safety, and affective computing. The development of a robust real-time emotion recognitifon system, therefore, is timely and its applications ought to be thoroughly investigated. For example, the design of more intelligent robots possessing the ability to understand human emotions could be realized by such a system.

Automatic emotion recognition has attracted considerable research and many solutions have been proposed. Most of these emotion recognition systems emphasise the importance of facial expression [18]. However, facial expression cannot always provide adequate information for emotion recognition. Some changes in emotion manifest only as very subtle changes in the facial expression and may go undetected, therefore reducing the accuracy of the emotion detection [42]. More recent research into the area of emotion recognition, has therefore exploited additional sources of information such as audio and gestures, with the aim to improve the reliability of the recognition process. In this approach all single modalities are integrated by multimodal fusion (feature-level and/or decision-level) [18], consequently leading to a some improvement in their performance [42, 37].

Deep learning techniques have shown great performance in a wide range of computer vision and natural language processing applications due to their built-in hierarchical representation learned directly from the data rather than human assumption [19]. Recently, deep learning features have also been shown to outperform hand-crafted features for the emotion recognition task [19]. Motivated by recent success of the deep learning techniques, we propose a novel approach to model spatio-temporal information for the emotion recognition task by a cascaded combination of 3 dimensional convolutional neural networks (C3Ds) and deep belief networks (DBNs) for both video and audio streams. The deep spatio-temporal features learned by our approach effectively model the discriminative features of both visual appearance and audio. Later, a score-level fusion approach is employed to determine the emotion.

The contribution of our paper is threefold:

1. We introduce a novel approach utilizing C3Ds for extracting the spatio-temporal features in both audio and video streams, which are represented as training samples for multimodal deep-belief networks. To the best of our knowledge, employing C3Ds [35] which are learnt in cascade manner with DBNs [2, 11] has not yet been investigated for audio-visual emotion detection.

2. We develop 4 different emotion recognition models as baseline systems, including audio C3D, video C3D, audio C3D + DBN, video C3D + DBN in order to validate our proposed methodology (audio-visual C3D +

978-1-5090-4822-9/17 $31.00 © 2017 IEEE
DOI 10.1109/WACV.2017.140

DBN).

3. We confirm, though experimental results, that our audio-visual C3D + DBN approach has significant potential to detect emotions from spatio-temporal information.

The remainder of this paper is organised as follows: Section 2 presents related research; Section 3 describes our proposed approach; Section 4 illustrates our experimental results; and Section 5 concludes the paper.

2. Related research

2.1. Multimodal emotion recognition

In spite of their recent promising achievements, the development of computer systems possessing the ability of understanding the emotions of human beings is still a challenging goal. Multi-modal data emotional state recognition based on facial expression, speech, and bodily movements could address this challenge, and enable human-machine-interaction systems to perceive emotional expression in real environments [29]. A number of analyses on multimodal human emotion recognition systems have emerged recently. They have shown that the integration of the information between multiple modalities could improve the performance of emotional recognition models [43]. Busso *et al*. [3] proposed two potential approaches: feature-level fusion and decision level fusion to fuse the facial expression and acoustic information, which are compared and analysed. In the first approach, a single classifier with features from both modalities is applied. A technique of sequential backward feature extraction, which helps optimize performance of the classification is used to extract the features of both modalities.

In the second approach, some criteria were adopted to fuse the posterior probabilities of the mono-modal systems at the decision level, including the following criteria: maximum, in which the emotion with greatest posterior probability in both modalities is chosen; average, in which the posterior probabilities of each modality are equally weighted and the maximum is selected; product, in which the posterior probabilities are multiplied and the maximum is selected; and, weight, in which different weights are used for the different unimodal systems. However, the accuracy of each type of emotion was shown to be dissimilar when the confusion matrices of the two classifiers were analyzed [3].

Another bimodal emotion recognition system is proposed by Emerich *et al*. [7] for combining facial expressions and speech signals. The models obtained from a bimodal corpus with six acted emotions and ten subjects were trained and evaluated by applying support vector machine, naive Bayes and k-nearest classifiers. Then, the feature level fusion and score level fusion are implemented for fusing visual and acoustic information. The results of experiments on the FEEDTUM database [38] show that the performance and robustness of emotion recognition systems is enhanced by adopting such fusion-based techniques.

Recently, a multimodal emotion recognition system has been proposed by Kessous *et al*. [16], in which emotion related features are extracted using expressive face, acoustic analysis of speech, and gesture. Unimodal data, bimodal data and tri-modal data are automatically classified by applying a Bayesian based model. In the multimodal cases, the different modalities are combined using fusion of the modalities at the feature level. Combining the multimodal data results in a substantial rise in the recognition accuracy in comparison with the unimodal systems: the accuracy of multimodal systems increased by greater than 10% compared to the most competitive unimodal system. They also have revealed that the bimodal system pairing gesture-speech achieved the best recognition accuracy compared to the systems pairing face-gesture or face-speech. Combining all three modalities resulted in an improvement of classifications by 3.3% over the best bimodal results.

An asynchronous feature level fusion approach is proposed by Mansoorizahed *et al*. [21] for generating a unified hybrid feature space out of the individual signal measurements. The multimedia content is more effectively clustered and classified by the use of the unified hybrid feature space. The proposed approach is representatively applied in the recognition of basic affective states from speech prosody and facial expressions. Experimental results tested on two audio-visual emotion databases including 42 and 12 subjects indicated a higher performance of the proposed system than the unimodal face based and speech based systems, as well as the synchronous feature level and decision level fusion approaches.

More recent works have focussed on the use of deep learning techniques for multi-modal emotion recognition. Kim *et al*. [17] use a convolutional neural network based model for a hierarchical feature representation in the audio-visual domain to recognise spontaneous emotions. They show that improvement of recognition accuracy is achieved when hierarchical features and multimodal information are adopted. In another effort proposed by Liu *et al*. [20], models are constructed from multiple physiological signals collected from sensors placed on the human body by a adopting multimodal deep learning approach so as to improve their performance and lessen the cost of acquiring physiological signals for real-world applications.

2.2. Deep belief networks

Deep belief networks have also been broadly explored with respect to multimodal emotion recognition. Yelin Kim *et al*. [17] revealed the effectiveness of deep belief networks (DBNs) for multimodal emotion recognition system. Ran-

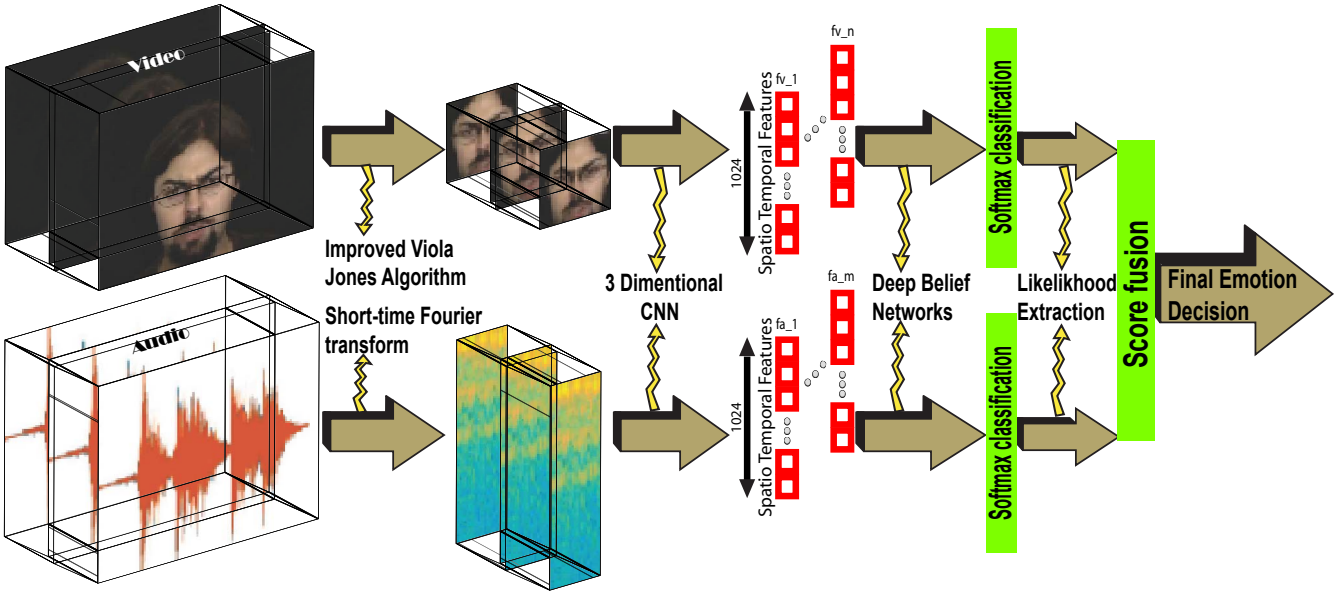

Figure 1. illustrates our proposed methodology. In the video stream, we initially apply the algorithm proposed in [36] to detect face regions, which are subsequently fed into C3D [35] to extract spatio-temporal features as input for DBNs [11]. In the audio stream, the spectrograms of the speech signals are extracted by adopting the short-time fast Fourier transform (FFT) [44], which are then fed into C3D for spatio-temporal extraction, and DBNs for classification respectively. We then extract every sample likelihood ratio test from the video trained-DBNs and the audio trained-DBNs simultaneously. Finally, we exploit a score-level fusion approach based on those likelihoods to determine the final emotion score.

ganathan *et al*. [27] demonstrate four DBN models, which could extract robust multimodal features for emotion classification in an unsupervised manner along with the proposal of convolutional deep belief network (CDBN) models that learn salient multimodal features of expressions of emotions. These models are validated on their emoFBVP database of multimodal (face, body gesture, voice and physiological signals) revealing better recognition accuracy in comparison with the recent novel emotion recognition systems [27]. However, both of these approaches are trained with multiple hand-crafted audio and video features.

Yan *et al*. [34] have also shown that most of the existing approaches are faced with either detecting individual Action Units (AUs) statically or only taking advantage of the temporal evolution of each AU, while ignoring the dynamic relationships between AUs. Nevertheless, psychologists who analyse human behaviour state that the dynamic features of facial actions are a key contributing factor to assessing naturalistic human behaviour. Furthermore, there exists semantic relationships among AUs such as the co-occurrence and mutually exclusive relationships. It is these spatiotemporal relationships among AUs that produce a meaningful facial expression. To address this issue, systematically modelling of the dynamic characteristics of facial actions including both the temporal movement patterns of each AU and the dynamic dependencies among AUs in a spontaneous facial display are proposed [34].

2.3. Three dimensional convolutional neural networks (C3D)

C3Ds have been shown to have the ability to learn spatio-temporal features effectively and their use has been explored for human action recognition. Baccouche *et al*. [1] a proposed two-step scheme with the purpose of classifying human actions. The spatio-temporal features are initially automatically learned with C3D. This can generate the temporal evolution of the learned features, which help discriminate among actions. These features are then treated as input to a recurrent neural network. Shuiwang *et al*. [13] has shown the limitation of CNNs regarding resolving 2D inputs. They, therefore, develop a novel C3D model for action recognition with the capability of extracting features from both the spatial and the temporal dimensions by operating C3D, consequently capturing the motion information encoded in multiple contiguous frames. The experimental results show that the C3D model substantially outperforms the frame-based 2D CNN for most tasks. Motived by this improvement, Tran *et al*. [35] utilized learning spatio-temporal with C3D for action recognition. They also indicated that C3D is well-suited for spatiotemporal feature learning and possesses the potential of modelling temporal information better than 2D CNN as convolution and pooling are spatiotemporally performed in C3D, consequently preserving the temporal information

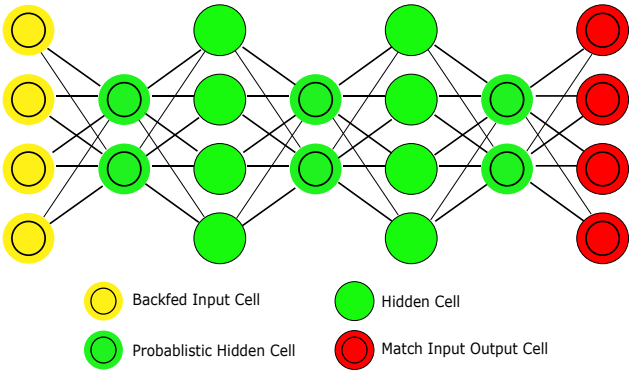

Figure 2. A simple graphical illustration of a deep belief network with 2 hidden layers

of the input signals, whereas they are only spatially operated in 2D CNN, thereby losing temporal information of the input signal right after each convolution step. All of the above-mentioned proposals outperform dramatically the novel baseline benchmarks for human action recognition. In other proposals, convolutional gated restricted Boltzmann machines have also been trained with spatio-temporal features for action recognition [32]. While the use of C3D have been demonstrated successfully for action recognition, its ability to capture the spatio temporal information for multimodal emotion recognition has not yet been explored.

3. Proposed methodology

Our proposed methodology is inspired by the success of the deep belief network for multimodal emotion classification and the success of C3D in learning effective spatio temporal features for action recognition. Based on the success of these techniques, we propose a novel approach to combine DBN and C3D for multimodal emotion recognition. In this approach we effectively learn the spatio-temporal features with C3D, which are then used as input for deep belief networks for training and classifying phases. Our experimental results show that the proposed system outperforms significantly the existing state-of-the-art approaches for audio-visual emotion recognition validated on the same database in the same manner. To the best of our knowledge, the proposed approach for the multimodal deep leaning model, in which DBNs are trained with audio and visual spatio-temporal features, which are extracted from trained audio and video C3D models, has not been investigated in the past for multimodal emotion recognition.

3.1. Pre-processing

3.1.1 Video stream

All frames are extracted from video for further steps. Since many face regions are missed or inaccurately detected using the Viola-Jones algorithm, consequently affecting the accuracy of recognition system, we propose the following simple algorithm to address this problem:

1. We apply the Viola-Jones algorithm to detect all bounding boxes containing face regions in each frame.

2. Based on these detected boxes, we detect the face region.

3. In some challenging frames, the Viola-Jones algorithm fails to detect faces. Based on the observation that the positions of the face in contiguous frames do not significantly change, we perform the detection in the current frame by cropping the position of the face region in the previous frame.

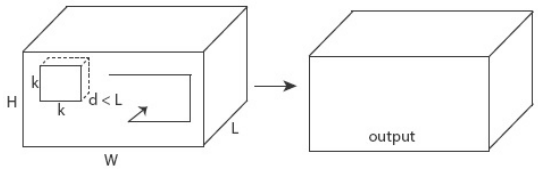

Figure 3. Utilizing 3D convolution on a video volume leads to another volume and preserving temporal information of the input signal [35].

3.1.2 Audio stream

We convert audio sampled at 48 kHz into frames by an overlapping 25-ms window at every 20-ms. The DFT of each overlapping Hamming windowed frame is then computed by applying a short-time fast Fourier transformation with 512 points before transforming the spectrogram segment into a log-power spectra feature with 257×72 dimensions [44] as input for C3D (see Figure 1).

3.2. C3D feature extraction

We apply the leave-one-subject-out evaluation protocol. We create the training set by sliding an overlapped window of 16-frames (one clip) with 5-frame steps and each clip is labelled with the corresponding ground truth label. If the same label occurs in each clip more than 8 frames, it is assigned this label. Two C3D networks, one for video and one for audio, are trained. The network settings and its parameters are explained in more detail in Section 4. To extract audio and video spatio-temporal features, all clips in test sets are passed to the trained-C3D to extract fc7 activations [35]. After this, each clip is converted into a 1024 dimensional vector capturing the spatio temporal information in 16 contiguous frames in a video. Figure 3 illustrates how C3D operates on a clip.

3.3. Deep belief networks

Deep belief networks (DBNs) are graphical models, formed by stacking and training restricted Boltzmann machines (RBMs) in a greedy manner. The joint distribution between the observed vector and the hidden layers are modelled as follows [11, 2],

$$P(x, h^1, \ldots, h^\ell) = \left(\prod_{k=0}^{\ell-2} P(h^k | h^{k+1}) \right) P(h^{\ell-1}, h^\ell), \tag{1}$$

where, $x = h^0, P(h^{k-1}|h^k)$ is a conditional distribution for the visible units conditioned on the hidden units of the RBM at level k, and $P(h^{\ell-1}, h^\ell)$ is the visible-hidden joint distribution in the top-level RBM.

A greedy layer-wise unsupervised learning algorithm for DBNs is proposed by Hinton *et al*. [11, 2]. Their experiments prove the hypothesis that this introduced training algorithm seemingly helps the optimization by initializing weights in a region near a good local minimum, leading to rising high-level abstractions of the input, providing better generalization. The principle of greedy layer-wise unsupervised training includes several processing steps [11, 2] as follows:

1. The first layer is trained as an RBM modelling the raw input $h^{(0)}$ as its visible layer.

2. That first layer is used to obtain a representation of the input as data for the second layer. The mean activations $p(h^{(1)} = 1|h^{(0)})$ or samples of $p(h^{(1)}|h^{(0)})$ are chosen for this presentation.

3. The second layer is trained as an RBM, taking the samples or mean activations (training examples) for the visible layer of that RBM.

4. Step 2 and Step 3 are iterated to obtain the desired number of layers, each time propagating upward either samples or mean values.

5. All the parameters of these deep networks are fine-tuned regarding a proxy for the DBN log-likelihood, or regarding a supervised training criterion.

In order to classify emotional state, we treat this DBN as multiple layer perception, in which its hidden layers are built as RBMs with the principle as follows: The first layer RBM is treated as the input of the network. The hidden layer of the RBM i^{th} represent as the input of the RBM $(i+1)^{th}$, and the last hidden layer RBM becomes the output of the network. A logistic regression classification is added on the top of the network to predict the input x based on the output of the last hidden layer $h^{(l)}$ of the DBN. Performing fine-tuning is conducted by supervised gradient descent of the negative log-likelihood cost function.

3.4. Score level fusion approach

The fusion of multimodal data can be executed by various fusion approaches, i.e at different levels of abstraction. Moreover, applying appropriate methods of fusion, such as at low level (early fusion or fusion at the signal level), intermediate level, or high level (semantic, late fusion, or fusion at the decision level) in order to achieve the best accuracy relies in part on the specific projects [6]. Among these fusion approaches, the fusion at the high level has been broadly utilized to fuse video, audio signal, and other streams to determine emotional state [6].

A large variety of research [24, 41, 31, 30, 4, 39, 10, 8, 9] has taken advantage of the feature level approach, whereas the decision level approach has also been utilized in [23, 15]. The experimental results indicated that the models with fused data outperformed the others. In other works [12, 25, 21, 26, 3, 40], the authors conducted experiments by adopting both early and late fusion approaches and they have shown that the overall accuracy of their proposals when combining audio and visual channels at the score level is higher than that when evaluating them with low level methods.

Feature-level fusion has been considered as a popular and straightforward technique in order to concatenate all recorded observation channels into a single high dimensional feature vector for the classifier. As a large amount of meaningful information is added into a combined feature representation, a rise in classification accuracy can be expected and the adoption of this fusion approach yields promising results [37].

Nevertheless, this fusion approach still has some shortcomings, including the generation of high dimensional features on small datasets, leading to stress on the computational resources with respect to training and evaluating the classification model [37]. For these reasons, in our proposed methodology we propose an algorithm to fuse audio and visual channels utilizing the score level approach instead of feature level fusion. We initially extract all likelihoods of every sample test from the trained DBN and find the corresponding weight, then we compute all likelihoods of every video based on these extracted likelihoods to determine the final emotion score. We compute this likelihood as follows,

$$L_{av}[i] = \sum_{j=1}^{n} L_a[j] * W_a[j] + \sum_{k=1}^{m} L_v[k] * W_v[k], \tag{2}$$

where n, m: the number of audio clip, video clip in video i^{th} respectively; $L_{av}[i], L_a[j], L_v[k]$: likelihood of video i^{th}, likelihood of audio clip j^{th}, likelihood of video clip k^{th}, respectively; $W_a[j]$, $W_v[k]$: weights of corresponding $L_a[j], L_v[k]$, respectively; Here we set $W_a[1]$=...=$W_a[n] =$

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.
Prosody, MFCC, LPC + SVM, NN [25]	-	-	-	-	-	-	25.00
MFCC + HMM [6]	48.06	38.46	91.72	69.34	41.48	44.68	55.90
Pitch Features + SVM [21]	43.00	21.00	30.00	44.00	30.00	30.00	33.00
Pitch, energy, formants, LPC, MFCC [24]	-	-	-	-	-	-	-
Acoustic Features + SVM [5]	53.00	-	-	55.00	49.00	-	52.33
MFCC + SVM [28]	66.55	69.2	62.43	68.12	70.25	67.84	67.40
The proposed method (Audio C3D)	76.42	81.67	81.94	85.21	24.78	86.90	72.82
The proposed method (Audio C3D + DBN)	79.92	83.68	84.25	83.93	81.82	83.38	**82.83**

Table 1. Results of our proposed method for audio emotion recognition in comparison with other audio state-of-the-art proposals (- means the results are not shown in the mentioned systems).

$W_v[1]$=...=$W_v[m] = 1/(m+n)$.

4. Experiments & results

We evaluated our proposed methodology on the eNTERFACE audio-visual database [22], in which five different experiments were carried out by using the Caffe framework [14] for C3D and the Theano framework [33] for DBN, including audio C3D, video C3D, audio C3D + DBN, video C3D + DBN, and audio-visual C3D + DBN. We applied the leave-one-subject-out evaluation protocol for both C3D and DBN. All experiments are conducted in a subject independent manner.

The eNTERFACE audio-visual database [22] includes 44 subjects (1166 video sequences). Each subject was asked to express anger, disgust, fear, happiness, sadness, and surprise and each emotion is simulated in five different reactions. A 720×576 Microsoft AVI format and a DivX 5.0.5 Codec is utilized for the video processing with 25 frames per second and the video compressing respectively. The audio was sampled at 48 KHz, in an uncompressed stereo 16-bit format. Finally, the database includes a total of 1166 video sequences. The number of women and men who participated in recordings are 264 and 902 accounting for (23%) and (77%) respectively [22].

- We set up our models of C3D based on the C3D network settings [35] used for training on UCF101 database. Our networks, therefore, also consists of 5 convolution layers (with appropriate spatial and temporal and stride 1) followed by a max pooling layer with kernel size 2×2×2 (except for the first layer with kernel size 1×2×2) with stride 1, 2 fully-connected layers and a softmax loss layer to predict emotion labels. The number of filters for 5 convolution layers from 1 to 5 is 64, 128, 256, 256, and 256, respectively. Two fully connected layers have 2048 and 1024 outputs respectively. Videos are split into overlapped 16-frame clips (dimensions 3×16×100×100 for video, 3×16×72×257 for audio) taken as input to the networks. There are 6 different emotions (anger, disgust, fear, happiness, sadness, and surprise). The networks are trained using mini-batches of 30 clips, with an initial learning rate of 0.003, which is divided by 10 after every 4 epochs and the training is stopped after 16 epochs.

- Once the C3D model training completes, we extract features from the fully connected layer and then feed them into the deep belief nets for training, including a layer-wise pretraining and a fine-tuning stage [11, 2]. For the first training stage, all the layers of the network are sequentially visited. For the second training stage, the final fine-tuning is conducted by adding a logistic regression layer on top of the network and the whole network is trained by stochastic gradient descent on the cross-entropy regarding the target classification. We set up the networks as follows: 1024 inputs, 2 outputs, 3 hidden layers with 1000 units per layer. An L2 weight decay hyper-parameter is also optimized. The network runs for 100 pre-training epochs with mini-batches of size 10, an unsupervised learning rate of 0.01, a supervised learning rate of 0.1. This is corresponding to performing 500,000 unsupervised parameter updates. Hyper-parameters were computed by minimizing on the validation error. Unsupervised learning rates in $\{10^{-1}$,..., $10^{-5}\}$ and supervised learning rates in $\{10^{-1}$,..., $10^{-4}\}$ are tested.

- In order to determine the final emotion, we combine audio and video streams by applying a score level fusion method. The idea is that we initially extract all likelihoods of every audio and video sample test from the trained audio and trained video DBNs, then we apply Equation (2) to compute the final likelihood of every video, and finally we compare the likelihood of each video with a threshold of 0.5 to determine the corresponding emotion score.

Audio: In Table 1, we show the experimental results of our audio emotion recognition systems. The proposed audio C3D + DBN system is 10% higher than the proposed

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.
Facial points + SVM, NN[25]	-	-	-	-	-	-	33.00
LBP + HMM [6]	18.62	60.28	48.06	53.14	19.7	26.66	37.74
Facial points features + SVM [21]	39.00	32.00	36.00	43.00	38.00	32.00	37.00
Face points[24]	-	-	-	-	-	-	-
Facial expression + SVM [5]	-	-	-	-	-	-	82.00
Spatio-temporal feature + SVM[28]	70.87	76.68	72.23	77.19	69.41	78.34	74.12
The proposed method (Video C3D)	77.12	83.56	83.31	86.40	82.72	85.57	83.11
The proposed method (Video C3D + DBN)	79.53	83.66	84.28	85.74	81.62	85.22	**83.34**

Table 2. Results of our proposed method for video emotion recognition in comparison with other video state-of-the-art proposals (- means the results are not shown in the mentioned systems).

Method	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Avg.
Feature and decision fusion approach [25]	-	-	-	-	-	-	39.00
Decision-level fusion [6]	-	-	-	-	-	-	56.27
Hybird fusion approach [21]	73.00	69	69.00	70.00	70.00	73.00	71.00
Feature level [24]	87.00	75.00	60.00	99.00	64.00	41.00	71.00
Score-level bimodal SVM [5]	-	-	-	-	-	-	87.40
Bayes sum rule (BSR) [28]	77.5	80.92	80.19	82.23	78.38	82.44	80.28
The proposed method (A-V C3D + DBN)	90.68	89.32	90.00	90.00	89.47	89.81	**89.39**

Table 3. Results of our proposed method for audio-visual emotion recognition in comparison with other audio-visual state-of-the-art proposals (- means the results are not shown in the mentioned systems).

audio C3D system, however both these systems outperform significantly the state-of-the-art audio emotion recognition proposals, which utilized the hand-crafted features. Our audio C3D + DBN proposal achieves the best recognition accuracy (**82.83%**), which is approximately 3.3 times higher than the recognition rate listed in [25].

Video: The results of our video emotion recognition systems are listed in Table 2. Both our proposed video C3D + DBN and Video C3D systems achieve quite similar recognition accuracy (**83.34%**, **83.11%** respectively), which are slightly higher than the best accuracy of the state-of-the-art video system [5] adopting hand-crafted features (**82%**) and is over 2.5 times higher than the recognition accuracy shown in [25]. This accuracy reveals that our video proposals performs better than our audio ones.

Audio-visual: In Table 3, we compare the recognition accuracy of our proposed methodology (audio-visual C3D + DBN) with the existing state-of-the-art audio-visual approaches. The recognition accuracy of our system (**89.39%**) is 2% higher than the best accuracy of audio-visual system based on hand-crafted features (**87.40%**) [5], is around 2.3 times higher than the audio-visual system [25], and is also considerably higher than the accuracy of audio-visual systems proposed in [6, 21, 24, 28]. When compared to our developed proposals (audio C3D, audio C3D + DBN, video C3D, video C3D + DBN), our proposed methodology also achieves the best recognition accuracy overall (**89.39%**).

5. Conclusion

The major challenge for a quality multimodal audio and video emotion recognition system is effective representation of spatial and temporal information. In this paper, we have proposed new spatio-temporal features based on a cascaded combination of 3 dimensional convolution neural networks (C3Ds) and deep belief networks (DBNs) for the multimodal emotion recognition task. The experimental results validated on eNTERFACE show that our proposed methodology (audio-visual C3D + DBN) achieves the best recognition accuracy (**89.39%**) in comparison with four baseline systems utilizing deep spatio-temporal features (audio C3D, audio C3D + DBN, video C3D, video C3D + DBN), and outperforms the state-of-the-art utilizing hand-crafted based approaches by a large margin.

6. Acknowledgement

This research was supported by an Australian Research Council (ARC) Discovery grant DP140100793.

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. *Sequential Deep Learning for Human Action Recognition*, pages 29–39. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[2] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. pages 153–160, 2007.

[3] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 205–211, New York, NY, USA, 2004. ACM.

[4] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, ICMI '06, pages 146–154, New York, NY, USA, 2006. ACM.

[5] G. Chetty, M. Wagner, and R. Goecke. *A Multilevel Fusion Approach for Audiovisual Emotion Recognition*, pages 437–460. John Wiley & Sons, Inc., 2015.

[6] D. Datcu and L. J. M. Rothkrantz. *Semantic Audiovisual Data Fusion for Automatic Emotion Recognition*, pages 411–435. John Wiley & Sons, Inc., 2015.

[7] S. Emerich, E. Lupu, and A. Apatean. Emotions recognition by speechand facial expressions analysis. In *Signal Processing Conference, 2009 17th European*, pages 1617–1621, Aug 2009.

[8] W. A. Fellenz, J. G. Taylor, R. Cowie, E. Douglas-Cowie, F. Piat, S. Kollias, C. Orovas, and B. Apolloni. On emotion recognition of faces and of speech using neural networks, fuzzy logic and the assess system. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 2, pages 93–98 vol.2, 2000.

[9] H.-J. Go, K.-C. Kwak, D.-J. Lee, and M.-G. Chun. Emotion recognition from the facial image and speech signal. In *SICE 2003 Annual Conference*, volume 3, pages 2890–2895. IEEE, 2003.

[10] M.-J. Han, J.-H. Hsu, K.-T. Song, and F.-Y. Chang. A new information fusion method for bimodal robotic emotion recognition. *JCP*, 3:39–47, 2008.

[11] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[12] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll. Bimodal fusion of emotional data in an automotive environment. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii/1085–ii/1088 Vol. 2, March 2005.

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

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[15] I. Kanluan, M. Grimm, and K. Kroschel. Audio-visual emotion recognition using an emotion space concept. In *Signal Processing Conference, 2008 16th European*, pages 1–5, Aug 2008.

[16] L. Kessous, G. Castellano, and G. Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1-2):33–48, 2010.

[17] Y. Kim, H. Lee, and E. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3687–3691, May 2013.

[18] A. Konar, A. Halder, and A. Chakraborty. *Introduction to Emotion Recognition*, pages 1–45. John Wiley & Sons, Inc., 2015.

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *NATURE*, 521(7553):436–444, 2015.

[20] W. Liu, W. Zheng, and B. Lu. Multimodal emotion recognition using multimodal deep learning. *CoRR*, abs/1602.08225, 2016.

[21] M. Mansoorizadeh and N. Moghaddam Charkari. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, 49(2):277–297, 2010.

[22] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACE'05 audio-visual emotion database. In *Proceedings of the 22Nd International Conference on Data Engineering Workshops*, ICDEW '06, pages 8–, Washington, DC, USA, 2006. IEEE Computer Society.

[23] B.-H. Moon and K.-B. Sim. Emotion recognition method based on multimodal sensor fusion algorithm. *International Journal of Fuzzy Logic and Intelligent Systems*, 8(2):105–110, 2008.

[24] M. Paleari, R. Chellali, and B. Huet. Features for multimodal emotion recognition : An extensive study. In *CIS 2010, IEEE International Conference on Cybernetics and Intelligent Systems, June 28-30, 2010, Singapore*, SINGAPORE, 06 2010.

[25] M. Paleari and B. Huet. Toward emotion indexing of multimedia excerpts. In *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*, pages 425–432, June 2008.

[26] M. Paleari and C. L. Lisetti. Toward multimodal fusion of affective cues. In *Proceedings of the 1st ACM International Workshop on Human-centered Multimedia*, HCM '06, pages 99–108, New York, NY, USA, 2006. ACM.

[27] H. Ranganathan, S. Chakraborty, and S. Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[28] M. Rashid, S. A. R. Abu-Bakar, and M. Mokji. Human emotion recognition from videos using spatio-temporal and audio features. *The Visual Computer*, 29(12):1269–1275, 2013.

[29] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Multimodal approaches for emotion recognition: a survey. volume 5670, pages 56–67, Jan 2005.

[30] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Emotion recognition based on joint visual and audio cues. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 01*, ICPR '06, pages 1136–1139, Washington, DC, USA, 2006. IEEE Computer Society.

[31] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition-a new approach. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04, pages 1020–1025, Washington, DC, USA, 2004. IEEE Computer Society.

[32] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. *Convolutional Learning of Spatio-temporal Features*, pages 140–153. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[33] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[34] Y. Tong and Q. Ji. *Exploiting Dynamic Dependencies Among Action Units for Spontaneous Facial Action Recognition*, pages 47–67. John Wiley & Sons, Inc., 2015.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[36] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[37] J. Wagner, F. Lingenfelser, and E. Andr. *Building a Robust System for Multimodal Emotion Recognition*, pages 379–410. John Wiley & Sons, Inc., 2015.

[38] F. Wallhoff, B. Schuller, M. Hawellek, and G. Rigoll. Efficient recognition of authentic dynamic facial expressions on the feedtum database. In *2006 IEEE International Conference on Multimedia and Expo*, pages 493–496, July 2006.

[39] Y. Wang and L. Guan. Recognizing human emotion from audiovisual information. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–1125. IEEE, 2005.

[40] Y. Wang, R. Zhang, L. Guan, and A. N. Venetsanopoulos. *Kernel Fusion of Audio and Visual Information for Emotion Recognition*, pages 140–150. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[41] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig. Low-level fusion of audio and video feature for multi-modal emotion recognition. In *Proceedings of the Third International Conference on Computer Vision Theory and Applications (VISIGRAPP 2008)*, pages 145–151, 2008.

[42] Q. Yao. Multi-sensory emotion recognition with speech and facial expression, 2014. Copyright - ProQuest, UMI Dissertations Publishing 2014;.

[43] Z. Zeng, M. Pantic, and T. Huang. Emotion recognition based on multimodal information. In J. Tao and T. Tan, editors, *Affective Information Processing*, pages 241–265. Springer London, 2009.

[44] W. Q. Zheng, J. S. Yu, and Y. X. Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 827–831, Sep 2015.