JAIST Repository
https://dspace.jaist.ac.jp/

Title	Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition
Author(s)	Atmaja, Bagus Tris; Akagi, Masato
Citation	2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 4482-4486
Issue Date	2020-05
Type	Conference Paper
Text version	author
URL	http://hdl.handle.net/10119/16248
Rights	This is the author's version of the work. Copyright (C) 2020 IEEE. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp.4482-4486. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

MULTITASK LEARNING AND MULTISTAGE FUSION FOR DIMENSIONAL AUDIOVISUAL EMOTION RECOGNITION

Bagus Tris Atmaja[1,2]*, Masato Akagi*[1]

[1]School of Information Science, Japan Advanced Institute of Science and Technology, Japan
[2]Department of Engineering Physics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
{bagus,akagi}@jaist.ac.jp

ABSTRACT

Due to its ability to accurately predict emotional state using multimodal features, audiovisual emotion recognition has recently gained more interest from researchers. This paper proposes two methods to predict emotional attributes from audio and visual data using a multitask learning and a fusion strategy. First, multitask learning is employed by adjusting three parameters for each attribute to improve recognition rate. Second, a multistage fusion is proposed to combine results from various modalities final prediction. Our approach used multitask learning, employed at unimodal and early fusion methods, shows improvement over single-task learning with an average CCC score of 0.431 compared to 0.297. Multistage method, employed at late fusion approach, significantly improved the agreement score between true and predicted values on the development set of data (from [0.537, 0.565, 0.083] to [0.68, 0.656, 0.443]) for arousal, valence, and liking.

Index Terms— multitask learning, multistage fusion, audiovisual emotion recognition, dimensional emotion

1. INTRODUCTION

Automatic emotion recognition has been approached using two perspectives: the categorical view and the dimensional view. While most researchers attempted to categorize human emotion within different categories (e.g. happiness, anger, etc.), dimensional emotion recognition is the more challenging task as it seeks to label the emotions as degrees rather than as categories. From dimensional perspective, emotion is described in 2 or 3 attributes [1]. Valence (pleasantness) and arousal (emotion intensity) are the two most common dimensions in 2D emotion models. In 3D models, either dominance (degree of control) or liking is used. Another model, such as expectancy, can be added as a 4th dimension (4D) [2].

In this paper, we evaluated three emotional dimensions/attributes: arousal, valence, and liking, which have been obtained from the dataset in [3]. The task is to obtain the most accurate prediction on a specific metric. As a regression task, the most common metric is the error between true value and predicted emotion degree. However, recent researchers [3] introduced correlation measurement to determine the agreement between true value and predicted emotion degree.

Two approaches are commonly used to minimize the loss of learning process functionality and to obtain the best model to predict emotion dimension, i.e., single-task learning (STL) and multitask learning (MTL). Single-task learning minimizes single loss function only in multiple output learning. For example, when learning to predict arousal, valence, and liking in dimensional emotion, only arousal is minimized. The other dimensions, valence and liking, are ignored in the learning process. By minimizing the error of arousal, the result of the learning process can be used to predict one dimension (arousal) or all three dimensions (arousal, valence, and liking).

The problem with single-task learning is that, when it is used to predict multiple outputs, three scores are predicted using single loss function. A high score in one dimension usually resulted in a lower score on the other dimensions. To address this issue, we introduced the use of multitask learning when minimizing error between true emotion and predicted emotion degree for all emotion dimensions.

The common approach in multitask learning is that the same weighting factors are used for each loss function in learning process. Therefore, the total is the sum of three each loss functions from each emotion dimensions. The method we propose in this paper is intended to obtain a balanced score by assigning a different weighting factor to each loss function for each emotion dimension.

As emotion comes from many modalities, the creation of a fusion strategy to accommodate those modalities is also a challenge. The standard method is by combining the features among different modalities in either the same or different networks. This is called an early fusion strategy. Two or more feature sets are then trained to map those inputs onto labels. Another strategy is the use of late fusion. In this strategy, each modality is trained in its network using its label. The recognition results for each modality are then grouped to find

the highest probability corresponding to the labels. The results from early fusion and late fusion also can be fused by combining those results in support vector regression (SVR). The result from this last step can be repeated in a multistage direction to improve the recognition rate.

Our contributions of this paper can be summarized as follows, (1) the use of multitask learning to minimize the loss function using three parameters for three emotion attributes from audiovisual features; (2) the fusion strategy by analyzing unimodal and bimodal features on early fusion and late fusion, and combining early-late fusion using multistage SVR to improve audiovisual emotion recognition rate.

2. RELATED WORK

Multitask learning. One of the problems in machine learning is to obtain the appropriate cost function or loss function to model the data. Most problems in regression analysis use error calculation between the true value and predicted value the loss function. The choice of the loss function is frequently determined by the metric used for evaluation. In the case of dimensional emotion recognition, Ringeval et al. of proposed the use of a concordance correlation coefficient (CCC) to score the performance of predicted emotion attributes [3].

Parthasarathy and Busso used multitask learning to minimize mean squared error (MSE) in dimensional emotion recognition [4]. The authors used two parameters to weigh loss function of three emotion attributes: arousal, valence, and dominance. Despite the weighting factor for both arousal and valence being determined, the weighting factor for dominance is obtained by subtracting 1 from the weighting factors of arousal and valence. All weighting factors lie in a range of 0-1 with a 33.3% possibility that one value is zero. It was also found that the best parameters are 0.7 and 0.3 for arousal and valence. In this case, dominance is ignored in learning process, which can be viewed as two-task learning which is similar to single task learning.

Using two approaches to multitask learning, such as shared layer and independent layer, the authors also achieved an improvement of CCC score compared to baseline single-task learning [4]. As the system learned better on the larger network than on the smaller one, the larger the network used, the greater improvement obtained,

Chen et al. also used multitask learning with MSE as the loss function [5]. Although the improvement of CCC score from the given baseline is achieved, the performance comparison to single-task learning is not specified. This potentially leads to a difficulty in determining whether the improvement came from multitask learning or other used strategies.

Multimodal Fusion. As emotion can be recognized from many modalities, e.g., speech, facial image, movement, and linguistic information, the use of multimodal technique to accommodate many features is often considered in such systems. The dataset described in [3] includes multimodal emotion features from audio and visual. Busso et al. provided an emotion dataset from speech and gesture, including facial expressions and hand movements [6]. The improved version of that dataset provided an affective database with audiovisual information which promoting naturalness within the (acted) recording [7].

To deal with various features extracted from multimodal datasets, several categories of feature fusion have been developed by researchers [5, 8, 9, 10, 11]. Most strategies can be divided into early fusion and late fusion. In an early fusion method, also known as feature level fusion, features from different modalities are combined before performing classification. In late fusion method, also known as decision level fusion, the final decision probabilities are given by each unimodal model results by such methods like SVR.

Ringeval et al. provides baseline fusion method for late fusion strategy from SEWA dataset [3, 12]. The results from each modality or feature set can be combined using a static regressor, i.e., SVR to make the final decision of predicted emotion attribute scores from given results of several modalities.

3. DATA AND FEATURE SETS

The SEWA dataset [12] provided in [3] is used in this research. The dataset contains audiovisual recordings from: Chinese, English, German, Greek, Hungarian, and Serbian, but only German (DE) and Hungarian (HU) are used in this work as they did not provide test label in other languages. Three attributes provided to represent emotional states i.e.: arousal, valence, and liking. The scores of those attributes are obtained from annotation of several native speakers: six Germans and five Hungarians. From 96 subjects, 68 subjects (34 each) are used in training, and the rest 28 subjects (14 each) are used for validation/development.

In addition to the dataset, the authors of paper [3] also provided baseline features which are shown in Table 1. Instead of generating new a feature set, we applied the multitask learning and multimodal audiovisual fusion to those feature sets.

For both audio and visual features, the same processing blocks are used, i.e., 4.0 s of window length and 100 ms of hop size, where the label is also given for each 0.1 s. The longest 1768 sequences (label numbers) is then used for all subjects by padding zeros for other sequences below this number. For bimodal feature fusion, audio and visual features are concatenated before they are fed into the classifier.

4. PROPOSED METHOD

4.1. Multitask learning based on CCC loss

CCC is the common metric in dimensional emotion recognition to measure the agreement between true emotion dimen-

Table 1. Audio and visual feature sets evaluated in this research. Feature sets highlighted in bold are used for bimodal/multimodal emotion recognition.

audio	eGeMAPS [13, 14], **Bag-of-Audio-Word eGeMAPS (BoAW-e)** [15], **eGeMAPS functional**, **DeepSpectrum** (DS) [16], MFCCs, **BoAW MFCCs (BoAW-M)** [15], MFCCs functionals.
visual	Facial Activation Units (FAUs) [17], **FAUs functionals**, **ResNet** [18], VGG [19].

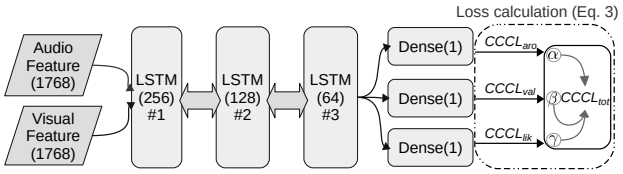

Fig. 1. The architecture of Deep Neural Network (DNN) with a multitask learning approach for minimizing loss function from three dense layers. The number inside the bracket represents units.

sion with predicted emotion degree. The CCC is formulated

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (1)$$

where ρ_{xy} is the Pearson coefficient correlation between x and y, σ is standard deviation, and μ is a mean value. This CCC is based on Lin's calculation [20]. The range of CCC is from -1 (perfect disagreement) to 1 (perfect agreement). Therefore, the CCC loss function (CCCL) to maximize the agreement between true value and prediction emotion can be defined as

$$CCCL = 1 - CCC \quad (2)$$

In single-task learning, the loss function is one of the loss functions from arousal ($CCCL_{aro}$), valence ($CCCL_{val}$), or liking ($CCCL_{lik}$). In multitask learning, when CCC loss is used as a single metric for all arousal, valence, and liking, the $CCCL_{total}$ is a combination of those three CCC loss functions:

$$CCCL_{tot} = \alpha\, CCCL_{aro} + \beta\, CCCL_{val} + \gamma\, CCCL_{lik}, \quad (3)$$

where α, β, and γ are the weighting factors for each emotion dimension loss function. In a common approach, α, β, and γ are set to be 1, while in [4], γ is set to be $1 - (\alpha + \beta)$ to minimize MSE. In that approach, all weighting factors are in range 0-1.

In this paper, we use all three parameters, and the sum of those weighting factors is not limited to only 0-1. As the goal is to strengthen CCC, CCC loss is used instead of MSE.

As shown in Fig. 1, the audiovisual emotion recognition system consists of 3 LSTM layers with 256, 128, and 64 units.

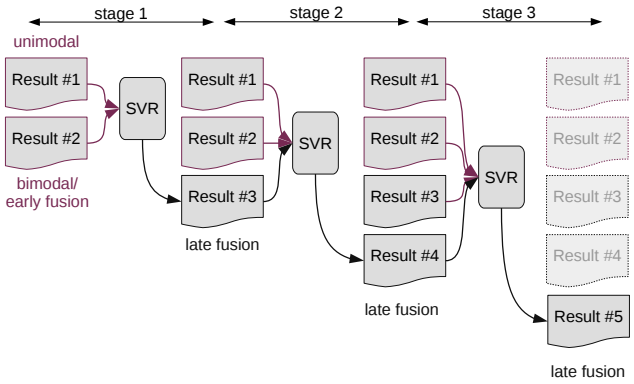

Fig. 2. The flow of multistage SVR to combine early and late fusion results.

A dropout layer with a factor of 0.4 is added after each LSTM layer. A RMSprop optimizer is used with a learning rate of 0.0005 and 34 batch size for 50 epochs in one experiment. To compensate for the delay when making an annotation, the label is shifted 0.1 to the front in the training process and shifted back in writing the prediction.

4.2. Multistage Fusion using SVR

In Fig. 1, the system produces a prediction of arousal, valence, and liking degree from bimodal audio and visual feature sets. This result can be combined with the other results from the unimodal or bimodal (early) fusion using SVR (from different feature set), and the resulting prediction from SVR also can be input to the same SVR system (implemented using scikit-learn tool [21]). In Figure 2, this combination of early fusion and late fusion is illustrated in three stages. First, the result from unimodal, named as result #1, and multimodal (bimodal), named as result #2, or unimodal and unimodal are trained using SVR method. This learning process results in a new result (namely, result #3 in that Figure). The result #3 from late fusion is fed again to SVR method results in result #4. Result #4 is fed again to SVR method results in result #5. This multistage fusion can be performed n-times to gain improvement of CCC score.

5. MULTITASK LEARNING RESULTS

To evaluate the effectiveness of the proposed MTL method versus STL and previous MTL methods, we compared CCC scores among those methods. Table 2 shows CCC scores for different attributes with its average. Our proposed MTL2 outperforms STL and previous proposed MTL1. To find the optimum parameter of α, β, and γ, we performed random search for those parameters in range 0-1. The parameters used in Table 2 are the optimum ones i.e. 0.7, 0.2, and 1.0 for α, β, and γ, respectively.

Table 2. CCC score of development set from FAUs feature set comparing STL and MTL. STL is performed by setting a weighting factor to 1 for the related attribute (Eq. (3)).

Loss	Arousal	Valence	Liking	Average
STL1 (Aro)	0.511	0.235	0.107	0.284
STL2 (Val)	0.255	0.558	0.077	0.297
STL3 (Lik)	0.255	0.32	0.191	0.244
MTL1 [4]	0.476	0.524	0.009	0.336
MTL2 (ours)	**0.522**	**0.578**	**0.194**	**0.431**

Our proposed MTL learning with three parameters outperforms STL and previous MTL [4]. For STL approaches, both arousal and valence obtained the highest CCC score when its attribute is optimized. Although the liking is optimized in STL3, it remains the most difficult to estimate. This problem should be addressed in future research.

6. MULTISTAGE FUSION RESULTS

To obtain multistage fusion results, the following steps are performed,

1. Unimodal emotion recognition: This step is performed to investigate the importance feature set for bimodal or multimodal fusion.
2. Bimodal fusion: This step is performed by concatenating two feature sets, from different or same modality.
3. Multimodal fusion: While the first two steps are performed using DNN, this third step is performed using SVR by combining results from unimodal or bimodal emotion recognition.
4. Multistage fusion: Output from multimodal SVR can be combined using the same SVR to improve the recognition rate of emotion recognition.

We run experiments on unimodal feature sets by inputting one feature set into a system to find which feature sets gives better performance. For this purpose, we use small networks with previously explained LSTM layers (implemented in Keras [22]). From 12 feature sets, we choose 7 feature sets by highest average CCC scores. The combination of seven feature sets resulted in 21 pairs of bimodal feature sets. Note that the definition of bimodal here is not audio and visual modalities but a pair of two feature sets. From unimodal and bimodal results, we choose the 11 highest CCC scores and input those 11 results to SVR to perform multimodal audiovisual emotion recognition by late fusion.

This last multimodal fusion using SVR can be regarded as 1-stage feature fusion. By inputting the result from SVR to the same SVR system, a 2-stage multimodal fusion can be performed. We limited this multistage multimodal fusion to 5 repetitions. The result of CCC scores for arousal, valence, and liking from 1 to 5 stages is shown in Fig. 3. That figure shows that CCC scores improved as the number of stages increased.

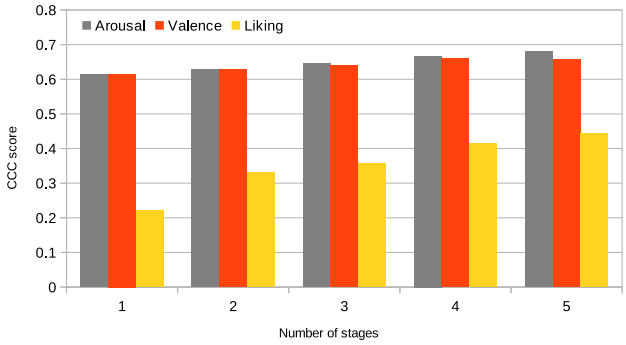

Fig. 3. CCC score among attributes from the different numbers of stages in proposed multistage feature fusion using SVR.

Table 3. CCC score comparison of development set on bilingual dataset by different methods; each row is the highest obtained score among feature sets using the same method.

Method	CCC Development (DE+HU)		
	Arousal	Valence	Liking
Baseline (FAUs) [3]	0.531	0.565	0.083
Unimodal	0.522	0.578	0.194
Bimodal early fusion	0.552	0.557	0.284
Multimodal late fusion	0.627	0.616	0.292
Multistage Fusion	**0.680**	0.**656**	**0.443**

In comparison with unimodal, bimodal, and multimodal fusion (1 stage), multistage fusion gained significant improvements. The proposed multistage fusion could improve CCC score of liking attribute, which is the most challenging attribute in this task, from 0.083 (baseline unimodal) to 0.443. Other two attributes obtained relative improvement over baseline results of 26.63% and 16.11% respectively for arousal and valence.

7. CONCLUSIONS

A multitask learning strategy is proposed to balance the CCC score among arousal, valence, and liking by adjusting parameters for those attributes. The result shows that by using different weighting factors for each emotional dimension, an improvement in terms of CCC scores can be obtained. Using weighting factors of 0.7, 0.2, and 1.0 for arousal, valence, and liking, respectively, for MTL parameters, we achieved an improvement of average CCC score from 0.297 using STL to 0.431 using our MTL. To deal with multimodal fusion of several feature sets, we proposed a multistage fusion using SVR method. This proposed method improves the CCC score on development test significantly for bilingual emotion recognition (DE+HU), especially on liking attribute i.e., [0.680, 0.656, 0.443] with an average CCC score of 0.593.

8. REFERENCES

[1] J Russel, "Three dimensions of emotion," *J Pers Soc Psychol*, vol. 9, no. 39, pp. 1161–1178, 1980.

[2] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[3] Fabien Ringeval et al., "AVEC 2019 workshop and challenge: State-of-mind, depression with AI, and cross-cultural affect recognition," in *Proc. of the 2019 on Audio/Visual Emotion Challenge and Workshop*, 2019.

[4] Srinivas Parthasarathy and Carlos Busso, "Jointly predicting arousal, valence and dominance with multi-task learning.," in *INTERSPEECH*, 2017, pp. 1103–1107.

[5] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proc. of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.

[6] Carlos Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[7] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. on Affect. Comput.*, vol. 8, no. 1, pp. 67–80, 2016.

[8] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," pp. 1103–1114, 2017.

[9] N Majumder, D Hazarika, A Gelbukh, E Cambria, and S Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," 2018.

[10] Jinming Zhao and Shizhe Chen, "Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions," pp. 65–72, 2018.

[11] Bagus Tris Atmaja, Kiyoaki Shirai, and Masato Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," in *APSIPA Assoc. Annual Summit and Conf.*, 2019.

[12] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Bjorn Schuller, Kam Star, et al., "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *arXiv preprint arXiv:1901.02839*, 2019.

[13] Florian Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. on Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[14] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, New York, New York, USA, 2013, pp. 835–838, ACM Press.

[15] Maximilian Schmitt and Björn W. Schuller, "openXBOW - Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, may 2017.

[16] Shahin Amiriparian et al., "Snore sound classification using image-based deep spectrum features.," in *INTERSPEECH*, 2017, pp. 3512–3516.

[17] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conf. on comput. vision and pattern recognit.*, 2016, pp. 770–778.

[20] Lawrence I-Kuei Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[21] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] François Chollet et al., "Keras," `https://keras.io`, 2015.