Aural-visual two-stream 기반의 아기 울음소리 식별

박철*, 이종욱**, 오스만*, 박대희**, 정용화**
*고려대학교 컴퓨터정보학과, **고려대학교 컴퓨터융합소프트웨어학과
e-mail: ku_bozhao@korea.ac.kr

Aural-visual two-stream based infant cry recognition

Zhao Bo*, Jonguk Lee**, Othmane Atif*, Daihee Park**, Yongwha Chung**
*Dept. of Computer Information Science, Korea University
**Dept. of Computer Convergence Software, Korea University

Abstract

Infants communicate their feelings and needs to the outside world through non-verbal methods such as crying and displaying diverse facial expressions. However, inexperienced parents tend to decode these non-verbal messages incorrectly and take inappropriate actions, which might affect the bonding they build with their babies and the cognitive development of the newborns. In this paper, we propose an aural-visual two-stream based infant cry recognition system to help parents comprehend the feelings and needs of crying babies. The proposed system first extracts the features from the pre-processed audio and video data by using the VGGish model and 3D-CNN model respectively, fuses the extracted features using a fully connected layer, and finally applies a SoftMax function to classify the fused features and recognize the corresponding type of cry. The experimental results show that the proposed system classification exceeds 0.92 in F1-score, which is 0.08 and 0.10 higher than the single-stream aural model and single-stream visual model.

1. INTRODUCTION

Due to the newborns' inability to speak, they are unable to communicate with their parents through normal human speech. Instead, infants express their feelings and needs to the outside world through non-verbal communication methods such as crying, displaying facial expressions, and body movement. While infant experts, like pediatric nurses and maternity matrons, are trained to understand the non-verbal messages and the reasons behind an infant's cry, for first-time parents who lack experience, interpreting the messages that their babies send them remains a challenging task [1]. The parents' failure to make sense of their babies' cries and take appropriate actions quickly can have negative effects on the cognitive and motor development of the newborns who are going through a period of rapid growth [2]. On the other hand, parents who recognize their infants' cries correctly and fast have higher chances of strengthening the parental bond between them and their babies, which can also promote the babies' social development. Therefore, building a system that can help parents understand the meaning behind the messages that their babies send them is necessary. In this study, we propose an infant cry recognition system that leverages the infants' facial expression and crying sound information to help first-time and inexperienced parents recognize the reason why their babies are crying in order to better respond to their needs. The proposed system is non-invasive and harmless since it only relies on audio and video data.

To distinguish different meanings behind infants' cries automatically, most of the methods that previous work proposed mainly analyzed the acoustic properties of babies' cries to recognize the different types of cries through their acoustic features. In these proposed methods, sound features, such as Mel-Frequency Cepstral Coefficient (MFCC) or spectrogram, were first extracted from the audio data to be used as an input. Then, some works used statistical methods, such as Bayesian or Gaussian Mixture Model (GMM), which had a big success in speech recognition, to classify the sound features [3-4]. With the prevalence of deep learning, other recent works mainly used deep learning methods, such as Convolutional Neural Network (CNN), as the classifier [5-7].

Although deep learning methods can significantly improve the recognition accuracy, models based only on audio data are very susceptible to environmental noise and the recognition accuracy is not always satisfactory. Thus, apart from using the crying sound data, some works utilized the infants' facial expression information and baby movement information in some neonatal pain detection and assessment tasks. For example, Sun and Shang [8] first extracted the geometric features, Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP), from the infant's face, then used Support Vector Machine (SVM) to classify those three features to detect infants' pain. To assess the intensity of the neonatal postoperative pain, Salekin et al. [9] used CNN to extract the spatial information from the infant's face, then used the Long Short-Term Memory (LSTM) to extract the temporal information from the frame sequences of the infant pain video. Later, the same authors extended their work to build a multimodal approach by leveraging the body movement and crying sound, excluding the facial expression information, and their results show that the multimodal approach outperforms the unimodal method by far [10].

Unlike baby pain assessment, which is for assessing the intensity of pain newborns are feeling, the infant cry recognition task is for distinguishing the reasons why the babies are crying to help inexperienced parents better take care of their babies. However, most of the existing infant cry recognition systems only use crying sound information, which tends to lack robustness and achieve low accuracy. In this paper, we propose an aural-visual two-stream model based infant cry recognition system that can distinguish different types of infant cry automatically with high accuracy. The proposed system first separates the videos into audio and frame sequences and then extracts the log-mel spectrogram from the audio and crops the infants' faces from the frame sequences. VGGish [11], which was specifically built for audio classification, and 3D-CNN [12], which is used for video classification, are then used to extract the audio and video features, respectively. Finally, a fully connected layer is used to fuse the extracted features and a SoftMax layer classifies the data into one of the four classes.

2. PROPOSED METHOD

The overall pipeline of our proposed system is shown in figure 1. The video is first split into audio and images streams. These two streams are pre-processed individually and then fed into each of the sub-modules to extract the two-stream features. The last layers of the two sub-modules are concatenated and fused using a fully connected layer before performing a joint classification of four different types of cry.

2.1 Aural Stream Pre-processing

For aural pre-processing, the audio signal is first extracted from the video and down sampled to 16kHz, and then pre-emphasis and z-score normalization are applied to the resampled audio clip. Finally, the audio data is converted into 96×64 log-mel spectrogram, where 96 is the time steps and 64 is the mel-bins, using the public VGGish spectrogram feature extractor [11].

2.2 Visual Stream Pre-processing

For visual pre-processing, we first apply the MTCNN [13] for every video frame which provides us with a detected face bounding box and five key points including two eyes center points, a nose point, and two mouth corners points. After that, face alignment is performed to rotate and crop the detected face and ensure that the line going through the two eye center points is horizontal. Finally, 8 key face images are selected from the aligned faces with the same interval and resized to 64×64 to represent the video clip.

2.3 Aural-Visual Two-Stream based model

This model includes two submodules and one fusion layer. Each of the submodules can extract the aural features and visual features respectively. The fusion layer first concatenates the extracted features of both sub-modules before fusing the concatenated features with a fully connected layer. The fused features are then fed to a SoftMax activation function to classify them into one of the four types of baby cries.

The aural model is used to analyze the log-mel spectrogram with a VGGish network [11] to extract the features from the log-mel spectrogram. VGGish is derived from the VGG model and is specifically built for audio classification tasks. The model is pre-trained on AudioSet database [14] and can embed the log-mel spectrogram into 128-dimentsional highly represented features.

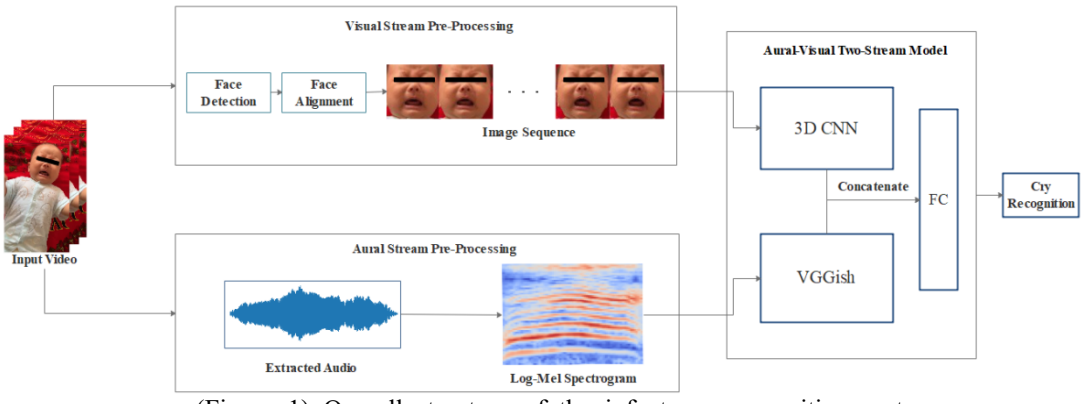
On the other hand, the visual model is used to extract features from the face sequences. At this level, we use a 3D-CNN to extract the features. Unlike the 2D-CNN that can only extract spatial features, 3D-CNN can extract both temporal and spatial features simultaneously and has shown a promising ability on learning spatial-temporal features [15]. Similar to the aural model, the visual model extracts 128-dimensional features from the face sequences.

The features extracted from both sub-models are then concatenated and fed to a fully connected layer with 128 dimensions to fuse the concatenated features. Finally, a 4-dimensional fully connected layer with a SoftMax activation function are used to perform classification into four classes.

3. EXPERIMENTAL RESULTS

3.1 Data Collection and Datasets

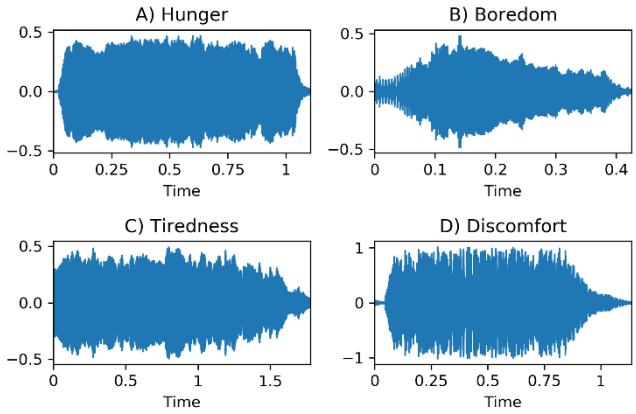
The infants' cries data was collected and annotated by skilled and experienced nurses and baby caregivers using their smartphones at home and in a hospital in Hebei, China. The babies' parents' consent to collect and use data in our study was obtained beforehand. Seven infants, including four boys and three girls, were filmed in total. The age of every recorded infant was between 0 and 3 months. Examples of video frames for different infants are shown in figure 2. Four types of cry, namely hunger, boredom, tiredness and discomfort, were recorded. The hunger cry was collected before the infants feeding time. The boredom cry was collected when a baby was left alone for a long time or trying to get a hug. The tiredness cry was collected before a baby fell asleep, and the discomfort cry was collected when a baby was experiencing colic or getting an injection. We manually trimmed the collected data into 0.3~2.5 seconds video clips that only contain cry signals to be used in our experiments. Furthermore, we removed



(Figure 1) Overall structure of the infant cry recognition system.

images with no baby face in them from the sequences of frames. The number of video clips for each type of cry recorded from each baby is shown in table 1. Examples of signal waveforms of different cries extracted from the short video clips are provided in figure 3.

(Figure 2) Examples of video frames for different infants.



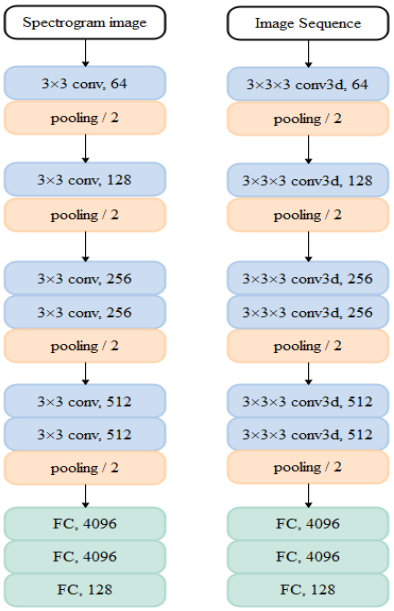
(Figure 3) Waveforms of different types of cry.

<Table 1> Data distribution for different infants

Infant	Hunger	Boredom	Tiredness	Discomfort
1 month boy	25	19	6	17
1 month girl	29	29	41	11
2 months boy	9	14	10	42
2 months girl	3	12	0	8
3 months boy	27	55	45	14
3 months boy	13	5	11	18
3 months girl	26	21	41	20
Total Clips	**132**	**155**	**154**	**130**

3.2 Feature Extraction Model

We used the VGGish model to extract the audio features and 3D-CNN to extract the video features. The architecture of VGGish and 3D-CNN are both provided in figure 4. The VGGish model has four groups of convolution/maxpool layers and 3 fully connected layers. The shape of the log-mel spectrogram it receives as input is 96×64 and the filter size used in each convolution layer and pooling layer is 3×3 and 2×2. On the other hand, the input of the 3D-CNN model is an 8 frames sequence with a size of 64×64 per frame. An architecture similar to the VGGish model was used and the only modification done was changing the filter size in each convolution layer to $3 \times 3 \times 3$ and the filter size in each pooling layer to $2 \times 2 \times 2$. The VGGish model and 3D-CNN model can embed audio and video features to 128-dimensional highly represented features. When training the two-stream model, these two 128-dimensional features are concatenated then fed to a 128-dimensional fully connected layer to fuse the features, which are then fed to a 4-dimension fully connected layer with SoftMax activation to predict the result. Meanwhile, for the single-stream model with only audio data or video data, a 4-dimensional fully connected layer with SoftMax activation is added directly after the 128-dimensional feature to predict the cry reason.



(Figure 4) Feature extraction models. Left: VGGish, right: 3D-CNN.

3.3 Implementation Details

We used Keras library [16] with TensorFlow2.0 as backend to implement our models, and all experiments were conducted on a computer running Windows 10, with an Intel i7-6700K CPU, 32GB of RAM, and a GTX 1080 graphic card.

Both of the single models' training and the two-stream model training share the same hyperparameters, with SGD as the optimizer using 0.001 for learning rate, 0.000001 for decay rate, 0.9 for momentum, ReLU as the activation function, and were trained for 100 epochs. 70% of the total data was used for training and 30% of the data for validation.

3.4 Results & Analysis

The experimental results of both single-stream models (visual and aural) are provided in table 2 and table 3, whereas the results of the aural-visual two-stream model are provided in table 4. Precision, Recall, and F1-score are used to measure the performance. As seen from the tables, the average F1-score of the aural stream model and visual stream model are 0.84 and 0.82 respectively while the F1-score of the two-stream model is 0.92, which is 0.08 and 0.10 higher than the single aural and the single visual models. The results confirm that the two-stream model can highly improve the performance compared to the Single-stream model in baby cry recognition. Also, the results are good enough, which indicates that the proposed system can effectively help inexperienced parents distinguish their babies' cries.

<Table 2> Experimental results with aural stream model

Class	Precision	Recall	F1-Score
Hunger	0.76	0.93	0.83
Boredom	0.92	0.87	0.89

Tiredness	0.93	0.91	0.92
Discomfort	0.77	0.67	0.72
Average	**0.84**	**0.84**	**0.84**

<Table 3> Experimental results with visual stream model

Class	Precision	Recall	F1-Score
Hunger	0.73	0.55	0.63
Boredom	0.88	0.87	0.87
Tiredness	0.78	0.93	0.85
Discomfort	0.90	0.97	0.93
Average	**0.82**	**0.83**	**0.82**

<Table 4> Experimental results with aural-visual two-stream model

Class	Precision	Recall	F1-Score
Hunger	0.79	0.93	0.85
Boredom	0.94	0.90	0.92
Tiredness	1.00	0.98	0.99
Discomfort	0.97	0.86	0.91
Average	**0.92**	**0.92**	**0.92**

4. CONCLUSION

To help first-time parents who have no experience in taking care of babies and feel powerless when dealing with babies' cries, in this paper, we propose an aural-visual two-stream based system to identify babies' cries. The results show that the two-stream model achieves a high accuracy and performs better than the single stream model for recognition of infants' cries, making it more reliable in real life application. In the future, we plan to take advantage of the lightweight models so that our proposed two-stream infant cry recognition model can be implemented in mobile equipment and provide better and more practical services for inexperienced parents.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B07044938 and NRF-2020R1I1A3070835) and by BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF)

References

[1] P. High, "The Happiest Baby on the Block: The New Way to Calm Crying and Help Your Newborn Baby Sleep Longer," Journal of Developmental & Behavioral Pediatrics, Vol. 26, No. 1, pp. 68-69, 2005.

[2] R.E. Grunau, M.F. Whitfield, J.P. Thomas, A.R. Synnes, I.L. Cepeda, A. Keidar, M. Rogers, M. MacKay, P.H. Richard, and D. Johannesen, "Neonatal Pain, Parenting Stress and Interaction, in Relation to Cognitive and Motor Development at 8 and 18 Months in Preterm Infants," Pain, Vol. 143, No. 1-2, pp. 138-146, 2009.

[3] H.E. Baeck and M.N. Souza, "A Bayesian Classifier for Baby's Cry in Pain and Non-Pain Contexts," Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 3, pp. 2944-2946, 2003.

[4] I.A. Bănică, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu, "Automatic Methods for Infant Cry Classification," International Conference on Communications, pp. 51-54, 2016.

[5] C.Y. Chang and J.J. Li, "Application of deep learning for recognizing infant cries," IEEE International Conference on Consumer Electronics-Taiwan, pp. 1-2, 2016.

[6] Zhao Bo, Jonguk Lee, Othmane Atif, Daihee Park, and Yongwha Chung, "Infant cry recognition using a deep transfer learning method," 한국정보처리학회 추계학술발표대회, 온라인 개최, 27 권, 제 2 호, pp. 971-974, 2020.

[7] M.A.T. Turan and E. Erzin, "Monitoring Infant's Emotional Cry in Domestic Environments Using the Capsule Network Architecture," Interspeech, pp. 132-136, 2018.

[8] Y. Sun, C.F Shan, T. Tan, X. Long, A. Pourtaherian, S. Zinger and P.H.N. de With, "Video-based discomfort detection for infants," Machine Vision and Applications, Vol. 30, pp. 933–944, 2019.

[9] M.S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho and Y. Sun, "First Investigation Into the Use of Deep Learning for Continuous Assessment of Neonatal Postoperative Pain," arXiv preprint arXiv:2003.10601, 2020.

[10] M.S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho and Y. Sun, "Multimodal Spatio-Temporal Deep Learning Approach for Neonatal Postoperative Pain Assessment," Computers in Biology and Medicine, Vol. 129, 104150, 2020.

[11] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 131-135, 2017.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725-1732, 2014.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE signal Processing Letters, vol. 23, pp. 1499-1503, 2016.

[14] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, and M. Ritter, "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 776-780, 2017.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features With 3D Convolutional Networks," Proceedings of the IEEE International Conference on Computer Vision, pp. 4489-4497, 2015.

[16] Keras, ver.2.4.0, Available Online: https://github.com/keras-team/keras (accessed on 15 September 2020).