ISSN 2466-135X
Vol.10, No.1

The 10th International Conference on

BIG DATA APPLICATIONS AND SERVICES (BIGDAS2022)

PROCEEDING

November 24-26, 2022
Jeju Island, Korea

Hosted by
Korea Big Data Service Society

THE KOREA BIG DATA SERVICE SOCIETY
한국빅데이터서비스학회
충북대학교 빅데이터연구소
JEJU 1952 제주대학교 JEJU NATIONAL UNIVERSITY
SW융합교육센터 Software Convergence Education Center

Table of Contents

A Solar Power Generation Prediction Model Based on CNN-LSTM

Woo-chan Park[1], Dong-kyu Yun[1], Ki-yong Park[1], Sang-hyun Choi[2*]

[1]Dept. Bigdata, Chungbuk National University, Cheongju, South Korea
[2]Dept. Management Information System, Chungbuk National University, Cheongju, South Korea
{2022278010, dongkyu.yun, pky3489, chois}@cbnu.ac.kr

Abstract. The renewable energy is considered to economic growth and key factor for the solution of energy crisis all over the world, and in particular solar energy is spotlighted as an alternative energy that can be easily obtained anywhere. However, solar energy lacks stability because electricity generated fully depend on the meteorological factors. Therefore, accurate forecasting of solar power generation is significantly important. In this work, we report on a solar power generation model based on CNN-LSTM with accuracy of 92%, and shows that compare to previous work.

Keywords: Solar Power, Renewable energy, Deep Learning, Time Series.

1 Introduction

The renewable energy sources are getting attention to meet the energy needs and regarded as promising solutions to solve the energy crisis, greenhouse gas and global warming [3].

Among the renewable energy resources, solar energy is expected the more acceptable and promising source because it is produced with no direct pollution or depletion of resources [2]. However, the accurate forecasting of solar power generation is considerably difficult because the generation of solar power fully depends on the uncertain meteorological factors, such as solar irradiance, humidity, and wind direction [1], [4]. In reason of this, better understanding and accurate forecasting of solar power generation has become an essential topic of research.

An accurate forecasting of solar power generation can reduce the impact of power uncertainty, improve system reliability and it will induce several benefits including environmental and economic benefit. Therefore, this study aims to predict the amount of solar power generation over time based on deep learning.

.

* Corresponding author

2 Data preprocessing

In this study, solar power and weather data were collected and used for about 4 years from 2016-12-13 to 2020-12-31. Each data has data collected every day from 0:00 to 23:00, and the data consists of a total of 35,487 rows.

Weather data, which is important to forecasting the performance of solar power generation, was collected using the Meteorological Administration API (Application Programming Interface). However, it was difficult to collect weather forecast data from the past, so the actual weather observation data was adjusted and used as weather forecast data.

To define correlation of each variables, including power generation, temperature, humidity, dew point temperature, insolation, and total amount of cloud, we use Pearson correlation coefficient. In this procedure, we removed some sections of solar power data which power generation value is recorded as '0' due to a system error and network communication error. Furthermore, when it was recorded as a '0' value at the time between 10 and 14, it was also classified as an outlier and replaced with an average value of power generation per hour. The missing values of weather observation data was calculated by interpolation.

After that, we split the whole data 70% for the training set, 20% for the validation set, and 10% for the training set.

3 Data Analysis

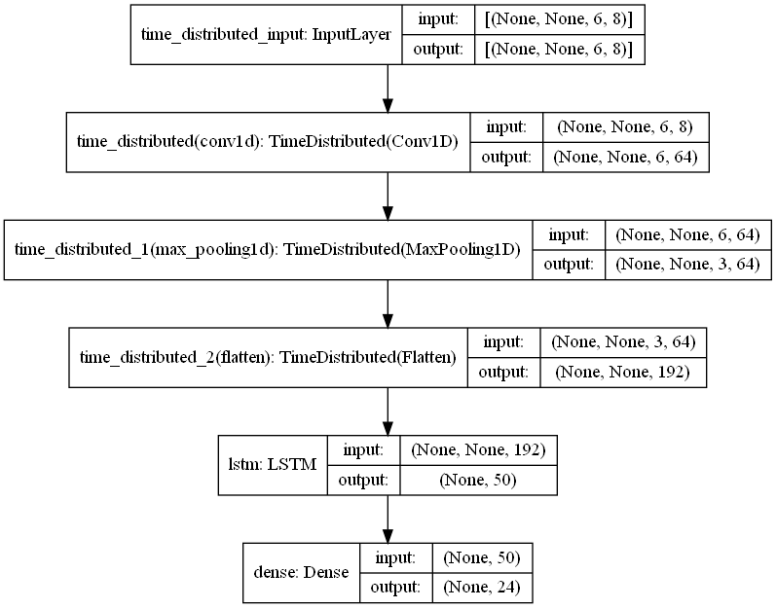


Figure 1 CNN-LSTM Model structure

The main purpose of this study is to predict the amount of solar power generation and enhance its performance by using deep learning algorithm. We designed a deep learning algorithm of a long short-term memory (LSTM) and convolutional neural network (CNN) model to improve the accuracy of solar power generation forecasting. The first step, we used CNN for extracting features from a time-series data. And, we used LSTM to predict solar power generation with relationship learning between past and future data obtained from CNN. <Fig.1> shows the structure of a CNN-LSTM used in this study.

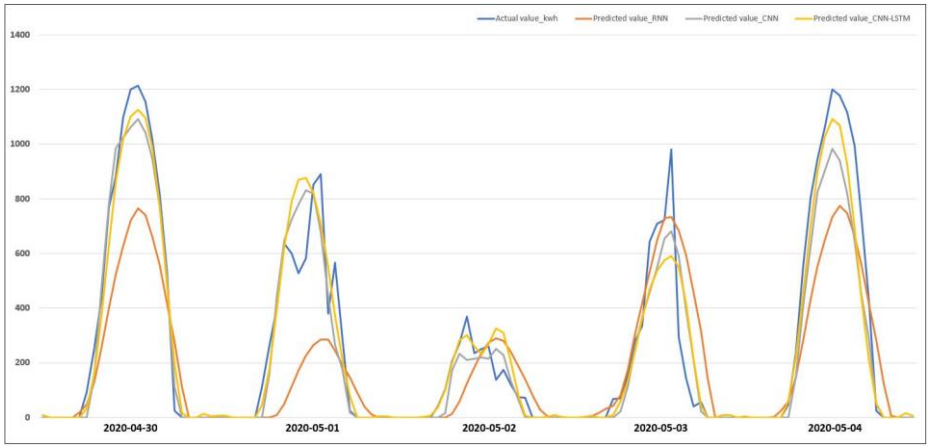


Figure 2 Comparison the performance of results from this work and previous work. (Blue: Actual value, Orange: RNN, Gray: CNN, Yellow: CNN-LSTM)

In previous work, we presented the deep learning algorithms such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) to predict solar power generation [6]. Fig 2 showed comparing the actual power generation and predicted power generation of previous work and CNN-LSTM algorithms on randomly extracted 5-day data from the results. As shown in <Fig 2>, prediction model based on CNN-LSTM is similar to actual value of power generation, and it showed high accuracy of 92.6%. Also, the results showed a further improvement in performance compared to the previous work [6].

	Mean Forecast Error	Mean Forecast Accuracy
RNN (previous work)	15.80 %	84.20 %
CNN (previous work)	12.03 %	87.97 %
CNN-LSTM	7.35 %	92.65 %

Table 1 Compare accuracy with previous work and present work

4 Conclusion

In this study, we discussed the solar power generation and necessity of solar power forecasting. An accurate forecast can enhance the operation and overcome instability of solar power systems according to weather condition.

In this study, we designed a deep learning algorithm based on CNN-LSTM for predict the amount of solar power generation. Also, we confirmed it has shown better performance with accuracy of 92.6% through compared with our previous work.

References

1. Alfredo Nespoli, Emanuele Ogliari, Sonia Leva, Alessandro Massi Pavan, Adel Mellit, Vanni Lughi and Alberto Dolara: Day-Ahead Photovoltaic Forecasting: A Comparison of the Most Effective Techniques. Energies, 12, 1621-1637(2019)
2. Utpal Kumar Das, Kok Soon Tey, Mehdi Seyedmahmoudian, Saad Mekhilef, Moh Yamani Idna Idris, Willem Van Deventer, Bend Horan, Alex Stojcevski.: Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Review, 81, 912-928 (2018)
3. Tao Ma, Hongxing Yang, Lin Lu: Solar photovoltaic system modeling and performance prediction, Renewable and Sustainable Energy Review, 36, 304-315 (2014)
4. J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de-Pison, F. Antonanzas-Torres.: Review of photovoltaic power forecasting, Solar Energy, 136, 78-111 (2016)
5. Su-Chang Lim, Jun-Ho Huh, Seok-Hoon Hong, Chul-Young Park and Jong-Chan Kim: Solar power Forecasting Using CNN-LSTM Hybrid Model. Energies, 15, 8233-8251 (2022)
6. Dong-kyu Yun, Ju-yeon Lee, Woo-seok Choi, Aziz Nasridinov, Sang-hyun Choi, Prediction of Solar Power Generation Based on CNN & RNN

Classification of Data Category from Public Metadata Database using Machine Learning

Woosuk Shin and Nakhoon Baek,

School of Computer Science and Engineering, Kyungpook National University,
41566 Daegu, Republic of Korea
w.shin@knu.ac.kr, nbaek@knu.ac.kr

Abstract. As government became an important provider of large scale public bigdata, management of such data is an important issue. Therefore, Korea government launched "central metadata managing system". However, the system lacks detailed description of the data, including weather it is structured or unstructured data. Considering current circumstances, sorting out structured data and unstructured data from public metadata database seems feasible. In this paper, we label, preprocess public metadata and present machine learning based prediction of unstructured data by its category from public metadata database.

Keywords: unstructured data, document classification, metadata, machine learning

1 Introduction

In an aspect of government, collecting database schema of its agency (administrative agency or public organization) is an important task. Collected database schema can be referred by another agency to minimize effort of re-building same existing database. Also, as government is becoming a great provider of large-scale public bigdata [1] for various applications, managing collected data and standardizing its metadata has become a significant issue [2] to provide high accessibility of the data.

Therefore, Korea government has revised guideline for standardizing database of public organizations [3]. According to the revised guideline, a government should develop "central metadata managing system (public metadata database)" and agencies should provide every metadata of their systems to central metadata managing system. Thanks to tremendous effort of building the system, the system currently has over 12 million metadata record of different agencies in Korea.

However, due to limitations of the system, system records include metadata for both unstructured data (i.e. scientific data, or series of files) and structured data (i.e. database scheme) as a same format. Since public metadata system stores metadata as a database format, [3] provides database fields which should be provided to the system. Provided fields contains no record which distinguishes between structured data and unstructured data. Therefore, though Korea government is preparing to distribute more unstructured data to public [4], it is abstruse to manage them in a useful manner, resulting in various public data providing platforms [5].

Though [6] presented noble method to manage metadata for such case, it requires major changes to currently operating system which is impossible. Furthermore, presented has abstract schema of applications, processes and users, which is not applicable for public metadata system. Thus, considering circumstances, distinguishing metadata of unstructured data or structured data from public metadata database is feasible method in a point of view to managing and providing easily accessible and reusable public data.

2 Data Labeling and Preprocessing

For machine learning, selecting test and train dataset is a important task. Since there is no categorization about unstructured data for original public metadata, we had to label metadata. According to [7], main interest category of unstructured data for Korea government is image data, video data, document data, text data, voice data, 3D data, spatial data, sensor data. Among over 12 million metadata of public metadata system, we labeled total 7 categories from [7], except 3D data's metadata. Also, since we have to consider both structured data and unstructured data, we assume that non-labeled metadata record is structured data's metadata, thus the classification should classify among total 8 category. 3D data category is excluded because it has too few labelable data. Table 1 shows labeled unstructured data category and corresponding number of metadata record for public metadata system.

Table 1. Labelled unstructured data category and its number of records in public metadata system.

Unstructured data category	Number of metadata records
Image	99,788
Video	43,190
Document	97,069
Text	30,723
Voice	1,332
Spatial	47,970
Sensor	8,322
Total	328,394

After labeling data, we made a metadata records in public metadata database as a single document by concatenating record's all attributes in to a single string, and regards them as a single text. Now, we can define metadata record classification problem as a document classification problem.

A common model used for document classification is vectorized model. For document vectorization, selecting good relevant vocabulary is important. Thus, we divided public metadata database's document into bag of morphs with help of open Korean text (OKT) morph analyzer [10]. Additionally, we researched several documents or reports that describes well about the specific unstructured data category, and extracted morphs from them. We assume morphs set of both public metadata database and documents as a dictionary for vectorization.

For the vectorization, since metadata records could contain meaningless repeated morphs, term frequency-inverse document frequency (TF-IDF) model is utilized to evaluate score of each morph in document. Additionally, as some machine leaning algorithms considers relative between its morph and other morphs padding it, we implemented vector as a linked list data structure form for better optimization.

3 Prediction results

For a prediction, as we vectorized single metadata entry as a document, we can utilize multiple machine learning technique to perform prediction. The method includes conventional classification algorithms like K-th nearest neighbor algorithm (KNN) and support vector machine (SVM). We also experimented deep learning-based machine learning algorithms such as recurrent neural network (RNN), long-short term memory (LSTM), bidirectional LSTM [8], and gated recurrent unit (GRU) [9].

Preprocessed dataset was divided into train and test dataset by randomly choosing 80% of data and 20% of data. Table 2 shows average accuracy of 10 steps of metadata categorization for different machine learning algorithms. The accuracy measure counts if the algorithm predicted correct category among 7 unstructured data categories as we showed in Section 2, and also includes if the algorithm has classified a records as non-categorizable, structured data.

Table 2. Accuracy measure for experimented machine learning (ML) algorithm. High accuracy value is better result.

ML Algorithm	Accuracy
GRU	0.286
KNN	0.289
RNN	0.354
LSTM	0.363
LSTM (bidirectional)	0.392
LSTM (w/time distributed layer)	0.424
SVM	0.907

Experiment result depicts that, although we expected more sophisticated and latest algorithm will perform better for classifying, classic machine learning algorithm SVM performed best. This result shows that for metadata classification, single morph used in metadata record has high importance than other documents.

4 Conclusion

In this paper, we show labeling and preprocessing process of unstructured data's metadata by its category. After that, we evaluated different machine learning algorithms by means of accuracy. For metadata classification of unstructured data, support vector machine, a conventional machine learning algorithm showed the

highest accuracy of 0.907. Although multi-layered deep learning algorithm showed highest accuracy among another deep learning algorithms, it showed accuracy of 0.424, which is half of support vector machine. We also want to underline that our proposed categorization model utilizes records in a database converted as a document, by concatenating its attributes and showed reasonable results.

Whilst dealing with government open data, we encountered many situations such as data is not accessible or it is not standardized, which leads to high cost in mining data. In the future, we hope to adopt our prediction method to public metadata system. By classifying metadata's category in real-time, the system will provide better accessibility to the unstructured data for the public.

Acknowledgments. This work has supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grand No.NRF-2019R1I1A3A01061310).

References

1. OECD, OECD OUR (Open, Useful and Re-usable) Data Index:2019, https://www.oecd.org/gov/digital-government/ourdata-index-korea.pdf
2. Rousidis, D., Garoufallou, E., Balatsoukas, P., Sicilia, M.E.: Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories. In: Information Services and Use. ACM, vol. 34, no.3-4, pp. 179--286 (2014). doi: 10.5555/3183921.3183935
3. Ministry of the Interior and Safety, National Information Society Agency: Management Manual for Standardizing Public Database. Jinhan M&B, Seoul (2022). doi: 10.979.11290/28884
4. Ministry of the Interior and Safety, Open Public Data 2.0 : Current status of Major Tasks, https://www.odsc.go.kr/files/boards/2848/2022_06_20_18533300937_2.pdf
5. Park, S., Ko, Y. M.: A Study on Metadata Interoperability between the National Research Data Platform and the Bio Research Data Platform (in Korean). In: Journal of the Korean Society for Information Management (JKOSIM). Vol. 39, no.2. pp. 159--202 (2022). doi: 10.3743/KOSIM.2022.39.2.159
6. Ravat, F., Zhao, Y.: Metadata Management for Data Lakes. In: New Trends in Databases and Information Systems. ADBIS 2019. Communications in Computer and Information Science, vol 1064. Springer, Cham (2019). doi:10.1007/978-3-030-30278-8_5
7. Definition of unstructured data and its category (in Korean). (2020). https://www.data.go.kr/bbs/ntc/selectNotice.do?originId=NOTICE_0000000001762
8. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016, August).: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics. Vol. 2: Short papers, pp. 207--212 (2016). doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00100
9. Dey, R., Salem, F.M.: Gate-variants of Gated Recurrent Unit (GRU) neural networks. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). Pp. 1597--1600 (2017). doi: 10.1109/MWSCAS.2017.8053243.
10. An Open Korean Text Processor. https://github.com/open-korean-text/open-korean-text

KERC2022: Drama Script-based Korean Emotion Recognition Challenge

Eun-Chae Lim[1], Sudarshan Pant[1], Hyung-Jeong Yang[1,*], Soo-Hyung Kim[1], Guee-Sang Lee[1], Aera Kim[1]

[1] Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
{218354, sudarshan, hjyang, shkim, gslee, arkim}@jnu.ac.kr

Abstract. The 4th Korean Emotion Recognition Challenge (KERC2022) focuses on Korean emotion recognition using the textual data from a Korean drama scripts. The KERC2022 aims to develop an emotion recognition model to classify three emotions: euphoria, neutral, and dysphoria. We labeled the conversation data based on collective labels by 64 Korean annotators who evaluated the emotion labels using our annotation software. In the challenge, 105 teams participated, and top 7 teams were awarded the prize based on the ranking of classification F1-score. This paper summarizes the dataset, baseline model, and results of the challenge.

Keywords: Korean Emotion Recognition; Context-aware Sentiment Analysis; Affective Computing; Challenge; Summary Paper

1 Introduction

People recognize each other's current situation and have an appropriate conversation when they talk. Text sentiment analysis is one of the most actively researched studies in natural language processing. In the past, emotions were analyzed by focusing on the words in the text. Recent research analyzes the emotions of the text by considering the situation, the interlocutors' relationship, the order of the sentences, etc., in addition to the text. Even with the same word, emotions can be analyzed differently depending on who is speaking and in what situation.

Drama is the medium where you can most easily see people talking with implied situational elements in daily life. In the drama, scenes are constructed due to people in various environments and their relationships, and they create situations by talking to each other. And since the drama script can check the situation through the eyes of a third person, we can extract various elements, and it is suitable for use in emotion recognition research.

The 4th Korean Emotion Recognition Challenge (KERC2022) focuses on developing a Korean natural language processing artificial intelligence model that predicts Korean emotions by analyzing conversational text data. We use a text dataset that can consider contextual elements in a drama script, not simply the emotion of a

*Corresponding author.

sentence. This dataset was constructed by the institute of Artificial Intelligence Convergence at Chonnam National University. In addition, through this competition, the participants were encouraged to contribute to context-aware sentiment analysis for Korean language corpus.

2 Related Work

2.1 Emotion Challenge

We compare KERC2022 with other emotion challenges such as KERC, EmotiW, MuSe as shown in Table 1. KERC is an emotion recognition challenge focusing every year on Korean emotion recognition. KERC began in 2019, and the number of participants continues to increase by performing tasks with new datasets yearly. KERC2019 [1] and KERC2020 [2] developed a model for recognizing the emotions of Koreans in video with SADVAW [6] datasets. KERC2021 [3] built a dataset by collecting bio-signals and personal characteristics of Koreans and developed an Arousal-Valence-based multi-modal emotion recognition model.

Several competitions for emotion recognition have been organized in the past, for instance, EmotiW challenge was one of the popular annual emotion recognition challenges. In the latest edition, EmotiW2020 [4], there were four sub-tasks as driver gaze prediction, audio-visual group-level emotion recognition, engagement prediction in the wild, and physiological signal-based emotion recognition. Similarly, MuSe is an active emotion recognition challenge, and MuSe2022 [5] conducted three sub-challenges: Humor Detection (MuSe-Humor), Emotional Reactions (MuSe-Reaction), Emotional Stress (MuSe-Stress).

Table 1. Comparison of the KERC2022 with existing emotion challenges

Challenges	Related Task with KERC2022	Emotions
KERC2019 [1]	Video-based Emotion Classification	Anger, Disgust, Fear, Happiness, Neutrality, Sadness, Surprise
KERC2020 [2]	Multimodal Emotion Recognition	Arousal, Valence, Stress
KERC2021 [3]	Bio-signal based Emotion Recognition	HAHV,HALV,LALV,LAHV
EmotiW2020 [4]	audio-visual group-level emotion recognition;	Positive, Neutral, Negative;
	physiological signal-based emotion recognition	Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise
MuSe2022 [5]	Emotional Reactions (MuSe-Reaction);	Adoration, Amusement, Anxiety, Disgust, Empathic Pain, Fear, Surprise;
	Emotional Stress (MuSe-Stress)	Arousal, Valence
KERC2022	**Text based Emotion Classification**	**Euphoria, Neutral, Dysphoria**

2.2 Drama based Emotion Dataset

We compare KERC2022 with other drama based emotion dataset are shown in Table 2. Korean emotional speech dataset [7] was recorded with Korean utterances from two females and two males from professional actors. They considered the sound quality of the data by removing any background noise, and all speech data were recorded in a professional studio using Korean scripts from dramas and movies. The dataset consisted of four emotions: anger, happiness, neutrality, and sadness, and 120 scenes per emotion which 30 seconds of long conversations between a male and a female per emotional scene.

Figure 1: Korean emotional speech dataset [7] recorded environment. There are two actors during the recording, using a conversation between a male and a female.

Korean facial emotion dataset [8] used the Korean TV series 'Misaeng' which is about various stories of a company. They constructed the facial emotion dataset by pairing an image and text, as shown in Figure 2. They cropped facial images and annotated each image with a textual description of the character's action. In addition, they used seven emotion classes (Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise) for annotation. They hypothesized that when it is difficult to accurately classify a character's facial emotions using only image data, the character's action descriptions(text) can be used to improve emotion recognition model performance. The results suggested that text descriptions of the characters' actions significantly enhance recognition performance.

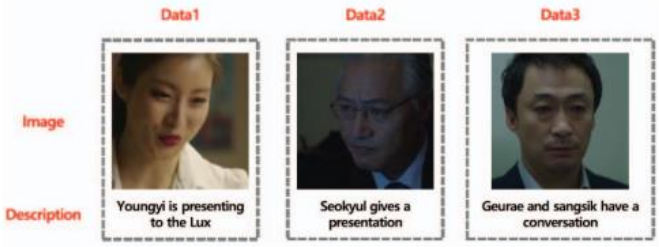


Figure 2: Korean facial emotion dataset [8] used the Korean TV series 'Misaeng' and they constructed pairing a facial image and action description(text).

Emotional RelAtionship of inTeractiOn (ERATO) is based on select dramas and movies with different genres to ensure data diversity. ERATO [9] extracted the pairwise interactive video clips using a total of 487 dramas belonging to 8 genres, with 1161 episodes. Figure 3 shows a sample of this dataset. We can guess the woman seems sick, and the man comforts her. The overall emotional relationship can show through interactions such as dialogue, facial images, actions, and background music. They suggested identifying the emotional relationships for these interactions can help recognize how character relationships develop in video content.

They divided the emotion categories into negative, positive, and neutral and divided negative and positive into two. Negatives were classified into Tense and Hostile, and Positives were classified as Mild and Intimate. Looking at the distribution of the entire dataset, it can be seen that the negative is 24%, the positive is 17%, and the neutral is 59%, so the Neutral occupies a lot of distribution. Although, many emotion recognition challenges are organized internationally every year, there are not enough competitions focusing on emotion recognition in the Korean context. With KERC's annual challenges, we aim to contribute to the affective computing domain in Korean Context. Moreover, KERC2022 involved behavioral emotion labels with importance in psychological studies.

Figure 3: ERATO [9] is built with audio, video, and subtitle modalities. They used dialogue and background music as audio, used detection of facial expressions in video clips, and subtitles as text data.

Table 2. Comparison of the KERC2022 with existing drama based emotion dataset

Dataset	Description	Multimodalities
Korean emotional speech dataset[7]	Recorded with actors using Korean drama and movie scrips	Speech, Text
Korean facial emotion dataset[8]	pairing a facial image and action description(text)	Image, Text
ERATO[9]	emotional relationship with interactions such as dialogue, facial images, actions, and background music	Audio, Video, Subtitles
KERC2022	**scene transcripts with realistic conversation data**	**Text**

3 KERC2022 Dataset

The KERC2022 dataset is constructed using the structure of a Korean drama script and consists of the script's scene number, speaker, scene description, and dialogue. Table 3 shows the sample of the KERC2022. In addition, for emotional labeling of the dataset, euphoria, neutral, and dysphoria were used as emotion classification by adding psychological emotion classification in collaboration with the Department of Psychology. Euphoria is a state of intense excitement and happiness, and dysphoria is a state of unease or generalized dissatisfaction with life. And emotions that do not include these two concepts are neutral.

Table 3. Sample of the KERC2022 dataset

Sentence id	Person	Sentence	Scene	Context	Emotion
1068	Yeon-hee	How about this? Try it on.	S0445	Yeon-hee choosing clothes.	Euphoria
1069	Hyeon-chal	How about it?	S0445	Hyeon-chal choosing clothes from the other side	Euphoria
1070	Yeon-hee	You always look so good like a model.	S0445	Yeonhee picks it up and goes	Euphoria

64 Korean participants were recruited and annotated with the emotion corresponding to the dialogue text using an annotation tool developed. The overall annotation progress of the participants is shown in Figure 1. Through the pre-meeting, we explained to participants the data annotation process, and they were divided into two groups to perform the annotation according to the timetable.

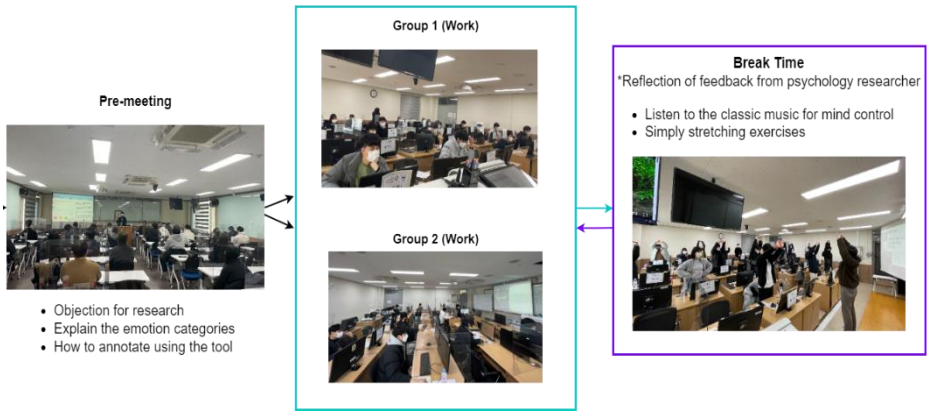


Figure 4: Overall annotation progress of the participants (Pre-meeting→ Annotation work→ Break time).

During the break, we played music so the participants could calm their minds and maintain their physical condition. Music that can relax emotionally and physically was selected, and music with lyrics was selected as classical music without lyrics because it may affect emotions. Then, stretching was performed twice a day during breaks to allow the participants to ventilate.

The characteristics of the dataset are shown in Table 4. The columns of the dataset are sentence id, person, sentence, scene, and context, which refer to the drama script's composition. The dataset consists of 12,289 samples, split into 7,339 train samples, 2,566 validation samples, and 2,384 test samples, available in a tab-separated values(tsv) format.

Participants annotated the emotions of 12,289 sentences, and the emotions of each sentence were determined by the majority vote of the participants' annotations. Table 5 shows the distribution of the emotion categories in the KERC2022 dataset with 7,482, 1,464, and 3,343 samples in dysphoria, neutral, and euphoria respectively. Dysphoria occupies a large distribution because it consists of the drama includes several episodes with events related to family conflicts.

Table 4. Attributes of KERC2022 dataset

Attribute	Description
Emotion Labels	Euphoria (27.2%), Neutral (11.9%), Dysphoria (60.9%)
Features	Sentence id, Person, Sentence, Scene, Context
Sample Size	12,289 (Train 7,339, Validation 2,566, Test 2,384)
Format	tsv (tab-separated values)

Table 5. Sample statistics per category

Category	Definition	Number	Ratio
Euphoria	state of intense excitement and happiness	3,343	27.2%
Neutral	emotions that do not include Euphoria and Dysphoria	1,464	11.9%
Dysphoria	state of unease or generalized dissatisfaction with life	7,482	60.9%

4 KERC2022 Baseline Model

We developed a baseline model for the KERC2022. The overall architecture of the KERC2022 baseline model is shown in Figure 2. The model performs the task of classification of the socio-behavioral emotional state (euphoria, dysphoria, or neutral) of the speaker for each spoken sentence spoken in a conversation. The proposed model consists of a scene-level context module, a speaker-level context module, and the cross-attention-based fusion method. The evaluation metric of the model is the F1 score, and the performance of the KERC2022 baseline model is 0.6372.

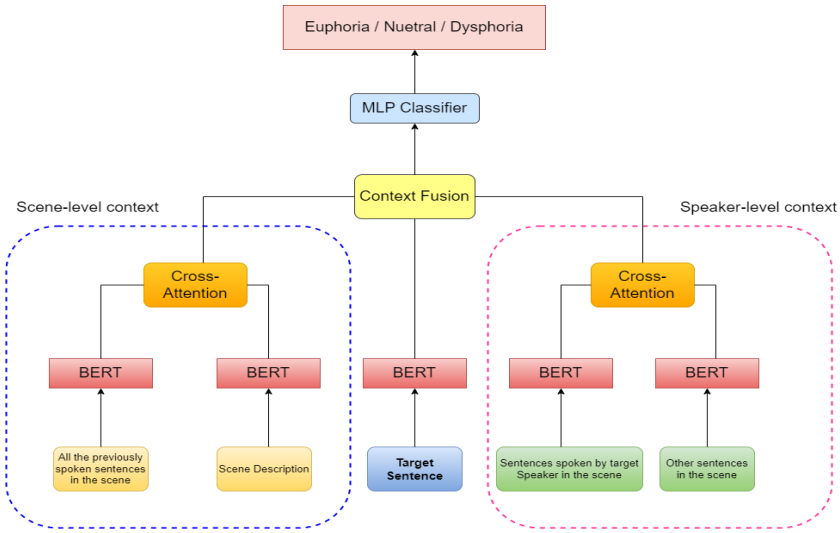


Figure 5: Overall architectures of the KERC2022 Baseline Model. Input: Scene-level context, Speaker-level context, Target sentence; Output: Classify the emotion corresponding to the input context fusion module.

5 Challenge Outcome

Looking at the results of the Top 3 in Table 6, all of them showed more than 0.15 performance than the baseline model. The top 3 teams designed the model based on Transformers. The team ranked 1 collected sentences for each scene and proceeded with the token classification task, and they improved the model performance using Focal Loss as the loss function. The team at the 2nd position improved the model performance by using the Korean benchmarking dataset model, KLUE-RoBERTa-Large, and they weaved all dialogues in the same scene to train. The team at the 3rd position used KLUE-RoBERTa-Large models and used ensemble method to improve the model performance through hard and soft voting methods.

Table 6. Top 3 on KERC2022

Rank	Team Name	Winner's approach	F1-score
1	Ju-hyuk Lee	Transformers+Focal Loss	0.8058
2	Hyun-Je Lee	Transformers+KLUE-RoBERTa	0.7995
3	AromaJewel	Transformers+Hard and Soft ensemble	0.7903
-	Baseline	cross-attention-based fusion	0.6372

6 Conclusion

The 4th Korean Emotion Recognition Challenge was held from August 22, 2022, to October 23, 2022. We provided a baseline model code as a starting model for participants in the KERC2022. Participants suggested various models based on the baseline. Overall, 105 teams from 7 nations and 33 universities registered in response to the call for participation. In addition, 44 out of 105 participant teams submitted their validation results, and 26 out of 105 participant teams submitted their test results. The top three teams reported f1 scores of 0.8058, 0.7995, and 0.7903, outperforming the baseline score of 0.6372 by a huge margin.

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-2020R1A4A1019191)

References

1. Inha, S, Vo, Thi T, V, Aran, O, Guee-Sang, L, Hyung-Jeong, Y, Soo-Hyung, K.: KERC 2019: The 1st Korean Emotion Recognition Challenge. In: SMA 2020, pp. 177--180. (2021)
2. Songa, K, Van, Thong, H, Dung, Tran, T, Aran, O, Guee-Sang, L, Hyung-Jeong, Y, Soo-Hyung, K.: The 2nd Korean Emotion Recognition Challenge: Methods and Results. In: IW-FCV 2021, pp. 176--183. CCIS. (2021)
3. Eunchae, L, Sudarshan, P, Hyung-Jeong, Y, Soo-Hyung, K, Guee-Sang, L.: Bio-signal and Personality based Korean Emotion Recognition Challenge: KERC 2021. In: KCEC2022, pp. 159--162. (2022)
4. Abhinav, D, Garima, S, Roland, G, Tom, G.: EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges. In: ICMI '20, pp. 784—789. (2020)
5. Shahin, A, Lukas, C, Andreas K, Eva-Maria, M, Alan, C, Erik, C, Bjorn W, S.: MuSe2022 Challenge: Multimodal Humour, Emotional Reactions, and Stress. In: Proceedings of the 30th ACM International Conference on Multimedia, pp.738--7391. MM'22. (2022)
6. Thi-Dung, T, Junghee, K, Ngoc-Huynh, H, Hyung-Jeong, Y, Sudarshan, P, Soo-Hyung, K, Guee-Sang, L.: Stress Analysis with Dimensions of Valence and Arousal in the Wild. Applied sciences. 11, (2021)
7. Sung-Woo, B, Ju-Hee, K, Seok-Pil, L.: Multi-Modal Emotion Recognition Using Speech Features and Text-Embedding. In: Applied sciences, 11, (2021)
8. Jung-Hoon, L, Hyun-Ju, K, Yun-Gyung, C.: A Multi-modal Approach for Emotion Recognition of TV Drama Characters Using Image and Text. In: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 42--424, (2020)
9. Xun, G, Yin, Z, Jie, Z, Longjun, C.: Pairwise Emotional Relationship Recognition in Drama Videos: Dataset and Benchmark. In: MM '21, pp. 3380--3389, (2021)

Separation Loss for Crowd Behavior Classification of Surveillance Videos

Jong-Hyeok Choi[1], Aziz Nasridinov[1,2,*], Yoo-Sung Kim[3,*]

[1]Bigdata Research Institute, Chungbuk National University, Cheongju, South Korea
[2]Department of Computer Science, Chungbuk National University, Cheongju, South Korea
[3]Department of Artificial Intelligence, Inha University, Incheon, South Korea

{leopard, aziz}@chungbuk.ac.kr, yskim@inha.ac.kr

Abstract. Crowd behavior classification from surveillance videos brings various studies because crowd behavior classification can make an effective video surveillance system that monitors various situations, such as clashes between crowds. For this reason, various studies have been proposed to perform such classification using 3D Convolutional Neural Networks (CNNs). However, it was difficult to classify the various crowd behaviors in the real world because they have similar but complex movements. To solve this problem, in this paper, we propose a *separation loss*, a new loss function to classify crowd behavior more clearly. The proposed loss function maximizes the difference in prediction scores between the target and other classes using individual predictions for each crowd behavior class, and it allows the crowd behavior classification model to focus on the features needed to classify crowd behaviors more clearly. To this end, we will explain the *separation loss* and then show the effectiveness of our loss through comparative experiments.

Keywords: Crowd Behavior Classification, Video Surveillance, Loss Function

1 Introduction

In the computer vision field, various studies have been conducted to enable crowd behavior classification from surveillance videos because crowd behavior classification can make an effective video surveillance system that monitors various situations automatically, such as clashes between crowds. For this reason, human behavior classification [1, 2] and crowd behavior classification [3-6] receive continuous attention, and various studies have been proposed to perform such classification using 3D Convolutional Neural Networks (CNNs). However, it was difficult to classify them due to the complexity of unpredictable human actions or various interactions between crowds [3]. In particular, crowd behavior classification was challenging to classify the various crowd behaviors in the real world because they have similar but complex movements due to various interactions between crowds.

* Corresponding authors: Aziz Nasridinov (email: aziz@chungbuk.ac.kr) and Yoo-Sung Kim (email: yskim@inha.ac.kr).

In this paper, we propose a new loss function called *separation loss* to classify crowd behaviors more clearly. The proposed *separation loss* maximizes the difference in prediction scores between the target and other classes using individual predictions for each crowd behavior class. So, using this loss function, the crowd behavior classification model can focus on the features needed to classify crowd behaviors more clearly. To this end, we will introduce the *separation loss* in detail. After then, we will conduct comparative experiments with the existing crowd behavior classification model to prove the effectiveness of our *separation loss*.

2 Related Work

2.1 3D Convolutional Neural Networks

Various deep learning-based crowd behavior classification methods currently perform classification through 3D CNNs [3-5]. Moreover, various methods use the Crowd-11 dataset [3], which can define 11 crowd behaviors, including no crowd, and Fig. 1 shows the differences between these crowd behavior classes.

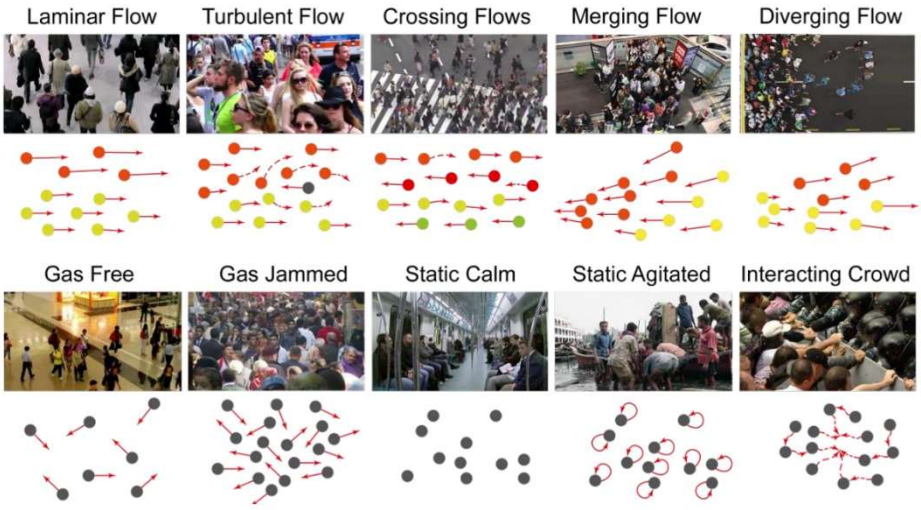


Fig. 1. Characteristics of each class in the Crowd-11 dataset, excluding no crowd [3].

The method proposed with the Crowd-11 dataset used a 3D CNN called Convolutional 3D (C3D) [1], and it showed better performance than the existing classification models based on 2D CNNs. Since, Inflated 3D ConvNet (I3D) [2], which showed good performance in human behavior classification by extending the GoogLeNet, was applied to crowd behavior classification using the Crowd-11 dataset. In particular, I3D significantly improved human behavior classification performance by a two-stream approach, which uses RGB and optical flow frames on different I3D models and merges the inference results. Because a two-stream I3D can make better

results by merging spatial features through RGB and temporal features through optical flow. So, even with the Crowd-11 dataset, the two-stream method performed the most accurate crowd behavior classification. However, in the cases of turbulent flows, crossing flows, merging flows, and diverging flows, which have shown similar behaviors among grouped crowds, are not accurately classified, so the need for a new method that can classify crowd behaviors more accurately is very high [3, 4].

2.2 Loss Functions

Currently, general classification problems deal with multi-class classification problems that seek to find a single class to which current data corresponds from multiple classes, and various loss functions have been proposed and utilized for this purpose. Currently, the most representative loss functions are Mean Squared Error (MSE) loss or Cross-entropy (CE) loss which utilizes classification error [7, 8]. However, the CE loss is most commonly used in deep learning-based classification methods because it is easy to increase the efficiency of the learning process, and it is divided into binary or categorical according to the number of classes. However, in the case of CE, since the same weight is given to all classes even if the number of data belonging to each class is different, models using CE have faced a class imbalance problem that does not pay attention to rare classes. In order to solve this problem, various losses, such as focal loss [8], are continuously being studied, and in this paper, we also intend to propose a new loss function to solve the problem caused by CE.

3 Separation Loss for Crowd Behavior Classification

This section describes the theoretical background of *separation loss*, a new loss function that can further improve the accuracy of crowd behavior classification based on the Crowd-11 dataset.

3.1 Solving the Easy/Hard Example Problem Using Focal term

As can be seen from the focal loss [8], CE tends to ignore classes that are difficult to predict and have a small number of data because a small loss is obtained when there are many classes that can be easily and accurately determined. In particular, these problems of CE in the model using the crowd-11 dataset are concentrated on classes such as laminar flow, static calm, interacting crowd, and no crowd with unique features or a large number of data. Conversely, CE makes it challenging to distinguish merging or diverging flows with similar features and a small number of data [4]. Therefore, to improve the crowd behavior classification accuracy, it is necessary to give more weight to the hard examples so that we can focus on the hard examples.

For this, different weights are given to easy or hard examples using the focal term of focal loss in *separation loss*. In the *separation loss*, when the ground truth $y \in [0, 1]$ is given, and the predicted value $p \in [0, 1]$ is obtained from the model , the probability p_t for each class for the focal term is calculated as follows.

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise,} \end{cases} \quad (1)$$

After that, each class's *Focal loss* value FL is calculated similarly to CE by applying the focal term $(1 - p_t)$ to the original CE in the following way using p_t:

$$\mathrm{FL}(p_t) = -(1 - p_t)\log(p_t) \quad (2)$$

The FL obtained in this way can respond to the easy/hard example by generating a low loss weight in the accurately predicted easy example and a high loss weight in the inaccurately predicted hard example. After that, in the *separation loss*, the FL calculated for each class is used to improve the class separation.

3.2 Improved Class Separation through Separation Weight

After calculating the FL for each class, SL, which is a final *separation loss* value, is generated by using FLs. In this step, using the FL obtained for each class, a *separation weight* is calculated to maximize the class separation between the target class and the rest of the classes. Such a *separation weight* is calculated by adding the target class's focal term to the remaining classes' average focal terms. So, this *separation weight* makes it possible to give a small weight to correct predictions and a larger weight to incorrect predictions. Therefore, this weight can improve the performance of the model by making it possible to focus more on problems that are difficult to classify.

Separation weight SW is calculated and utilized as follows when given video clip x with n classes:

$$\mathrm{SW}(x) = \sum_{i=1}^{n} \left(y_i\left(1 - p_{i,t}\right) + \frac{(1 - y_i)\left(1 - p_{i,t}\right)}{n - 1} \right) \quad (3)$$

Then, the final *separation loss* SL using these FL and SW is calculated as follows:

$$\mathrm{SL}(x) = \mathrm{SW}(x) * \sum_{i=1}^{n} \mathrm{FL}\left(p_{i,t}\right) \quad (4)$$

This way, since our *separation loss* uses both focal terms and the *separation weight*, more accurate prediction for each class and more accurate classification of crowd behavior classes with similar features becomes possible.

4 Experiment Result

In this section, we will evaluate the effectiveness of our *separation loss* in crowd behavior classification through experiments. To do this, we will use the original I3D structure with the Crowd-11 dataset to obtain a more objective comparison, and the experimental environment and results will be discussed in the following subsections.

4.1 Experimental Environment

In the case of the original I3D, the inference had performed using 64 RGB or 64 optical flow frames obtained from the same video, and the last classification layer classified as many as classes defined in the dataset. This paper uses this structure identically, except for the last classification layer, to change the number of classes to 11, which is the number of classes in the Crowd-11 dataset.

In the case of training and validation of I3D using the Crowd-11 dataset, the training and validation phase use only 5987 video clips consisting of 64 frames or more from a total of 6146 video clips, excluding videos lost online. Moreover, since it is not splitting into training and validation as official annotations, datasets are split randomly into 80% as training and 20% as validation. In addition, since the performance of optical flow-based classification models greatly depends on the optical flow algorithm, we use the traditional TV-L1 algorithm [9] for all cases. Furthermore, in the training step, 64 frames extracted from video clips were resized to 256×256 and then cropped to 224×224 with random crop and random flip, and in the validation step, only resizing to 224×224 was performed. Furthermore, in all these processes, categorical cross-entropy, the most representative CE used in multi-class classification problems, was selected as a comparison target to show the performance of *separation loss*.

Finally, the model has trained 200 epochs with a batch size of 16 on Titan RTX. And we use the SGD optimizer with a momentum of 0.9 and a base learning rate of 0.1 with a cosine annealing scheduler.

4.2 Experiment Result

Table 1 shows the difference in the classification accuracy of the I3D model trained on the Crowd-11 dataset according to loss functions. Since 64 frames were extracted randomly from the video clip, the validation test was repeated five times using only RGB or optical flow and the two-stream to show more evident results.

Table 1. Classification accuracy of I3D by loss functions for the Crowd-11 dataset.

Loss	Variant	1	2	3	4	5	Average
Categorical Cross-Entropy	RGB	**72.62%**	71.79%	71.95%	71.37%	71.87%	71.92%
	Optical Flow	71.79%	**72.37%**	71.79%	71.20%	71.20%	71.67%
	Two-Stream	77.38%	77.30%	76.96%	**77.80%**	77.46%	77.38%
Separation Loss	RGB	73.46%	73.87%	74.04%	**74.96%**	73.54%	**73.97%**
	Optical Flow	**78.05%**	76.83%	77.05%	76.71%	76.79%	**77.09%**
	Two-Stream	80.05%	80.30%	**81.22%**	80.22%	80.47%	**80.45%**

The experiment confirmed that classification accuracy improvement was made in all cases when using the *separation loss* compared to when using the CE loss. In particular, it highly improved when using the optical flow, and these results affect the average accuracy improvement by up to 3.07% in the case of two-stream. Through this, we can confirm that our *separation loss* is superior to the existing CE loss in crowd behavior classification, which can be said to be due to the focal term and *separation weight*.

5 Conclusion

In this paper, we proposed a new loss function called *separation loss* to make a more accurate crowd behaviors classification using the Crowd-11 dataset. The proposed *separation loss* maximizes the difference in prediction scores between the target and other classes using focal term and *separation weight*. So, using our loss function, the crowd behavior classification model can focus on the features needed to classify crowd behaviors more clearly. In addition, to prove the *separation loss*'s effectiveness, a comparison experiment was performed on the difference according to the loss function in the I3D model, one of the existing crowd behavior classification models, and prove the superiority of the proposed loss function.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00203, Development of 5G-based Predictive Visual Security Technology for Preemptive Threat Response). This work was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A1A01041815).

References

1. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision, pp. 4489-4497. IEEE (2015)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308. IEEE (2017)
3. Dupont, C., Tobias, L., Luvison, B.: Crowd-11: A dataset for fine grained crowd behaviour analysis. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 9-16. IEEE (2017)
4. Bendali-Braham, M., Weber, J., Forestier, G., Idoumghar, L., Muller, P. A.: Transfer learning for the classification of video-recorded crowd movements. In: International Symposium on Image and Signal Processing and Analysis, pp. 271-276. IEEE (2019).
5. Bendali-Braham, M., Weber, J., Forestier, G., Idoumghar, L., Muller, P. A.: Ensemble classification of video-recorded crowd movements. In: 12th International Symposium on Image and Signal Processing and Analysis, pp. 152-158. IEEE (2021)

6. Sreenu, G., Durai, M. S.: Intelligent video surveillance: a review through deep learning techniques for crowd analysis. Journal of Big Data, 6(1), 1-27 (2019)
7. Wang, Q., Ma, Y., Zhao, K., & Tian, Y.: A comprehensive survey of loss functions in machine learning. Annals of Data Science, 9(2), 187-212 (2022)
8. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P.: Focal loss for dense object detection. In: IEEE international conference on computer vision, pp. 2980-2988. IEEE (2017)

RFPN: End-to-end and efficient scene text recognition using Feature Pyramid Network

Ruturaj Mahadshetti, Guee-Sang Lee*, Hyung-Jeong Yang, Soo-Hyung Kim

Department of Artificial Intelligence Convergence,
Chonnam National University, Gwangju, South Korea,
ruturajm9770@gmail.com, gslee@jnu.ac.kr, hjyang@jnu.ac.kr, shkim@jnu.ac.kr

Abstract. Scene text recognition (STR) plays a vital role in various computer vision applications. STR has been a hot research topic in the computer community, and deep learning-based STR methods have achieved remarkable results over the past few years. However, Previous state-of-the-art scene text recognition methods may still produce a notable amount of false outcomes when applied to images captured in real-world environments because of less robust features and semantic information about text. In order to resolve this issue, we propose an approach called RFPN, which integrates ResNet and Feature Pyramid Network (FPN) to grab multi-level relations and enrich the functionality and generalization of the feature extractor. The proposed framework deal with different scale features to improve the robustness of features and semantic information. Extensive experiments on various datasets demonstrate that our method can acquire significant performance accuracy.

1 Introduction

As a branch of computer vision, Scene Text Recognition (STR) has become a popular and interesting research topic in the computer vision community and industrial fields. Recent works have endeavored to enhance the accuracy of scene text recognition using deep networks, such as upgrading the attention approach [1, 3, 4], elevating the backbone networks [1, 5, 7], and applying rectification modules [6, 8, 2]. However, STR is a very strenuous task due to the bad picture quality in natural images, such as low resolution, complicated backgrounds, and different fonts. Many recent studies use deep learning methods to handle these challenges e.g, sequence models [10,17,21], visual feature extraction methods [15,11,12], and rectification modules [13,14,16].

Many modern scene text recognition methods [20,18,19] perform well, but their performance significantly falls with low-resolution and partially occluded images. Low-resolution text images exist in many cases, e.g., a photo taken with fewer focal cameras or an image squeezed to reduce disk usage. When addressing low-resolution images, proposed recognition methods usually use interpolation methods (bicubic, bilinear). However, upsampled images are still blurred. Some failure cases from the [20] framework and RFPN outputs are shown in Fig. 1.

The recent growth of deep learning-based scene text detection methods [30,27,29,28] has shown the effectiveness of a feature pyramid network. They trained

a framework to extract features rather than creating feature extractors manually, which significantly enhanced the precision of text detection.

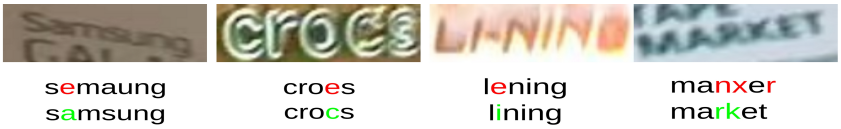


Fig. 1: The comparison of our and existing work such as [21].

In this paper, we focus on addressing the above problem and increasing the ability of extracted features, which improves the robustness of visual features. Fig 2 shows the detailed architecture of the proposed pipeline. We propose a novel framework (RFPN) that integrates ResNet-45 and Feature Pyramid Network. The feature pyramid network (FPN) [24] has features from top to bottom and combines them. It gradually merges them with semantic features to obtain multi-scale features of the input image and finally binds features of different scales. Our method enhances insufficient extracted features in the recognition process and improves accuracy. We use a feature pyramid for upsampling visual features and combine these features to capture the text content. RFPN increases the feature extractor's ability to generate robust visible features about text content that rapidly boosts the recognition performance and overcomes false positive outcomes. The main contributions of the proposed framework can be listed as follows: 1) We introduce a novel RFPN framework that achieves promising recognition accuracy. 2) The RFPN generates robust semantic features of text content from low-resolution images.

Proposed method

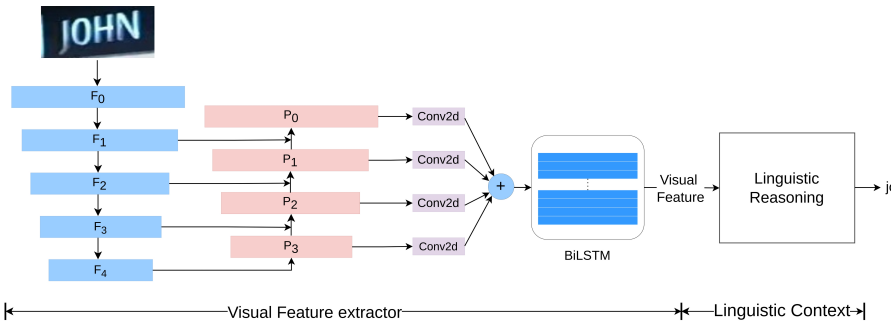


Fig. 2: The proposed pipeline of RFPN, the framework classifies into two parts 1) Visual feature extractor and 2) linguistic reasoning.

The RFPN framework consists of two parts: Feature extractor, Linguistic Reasoning. The Feature extractor is used to extract 2D features. Then, the Linguistic Reasoning model predicts the characters.

A) Feature extractor:

We integrate ResNet-45 and FPN to extract features. The text image sends as input to ResNet, then FPN upsamples features from the stage[5], stage[4], stage[3], and

stage[2]. After upsampling, all stages are converted into similar channel sizes (512) using the (3x3) conv and aggregate all stages. Thus, the feature map (V) size of ResNet-45+FPN is 1/4 of the input image, and the channel number is 512. After extracting features, Bidirectional LSTM(BiLSTM) captures the long-dependencies. BiLSTM produces a better sequence H = Seq. (V).

B) Linguistic Reasoning:

It consists of semantic reasoning and a parallel layer. Multi-head attention mechanism attention helps an STR process to learn a character-level language model representing output class dependencies. Position encoding uses to perceive the pixel location information. Then, the Parallel Prediction layer proposes to predict the characters in parallel.

Experimental Results

A. Dataset

We use SynthText and SynthText90K datasets for training, The performance evaluation on IIIT5K, ICDAR13, ICDAR15, SVT, SVTP, and CUTE80 datasets.

B. Implementation

We use the ResNet-45 as a backbone. Initially, we set the default value of the stride for the first stage, then increment it by 1 for the remaining, and initialize weights by default value. For data augmentation, we use perspective distortion, random rotation, color jittering, and experiment on 2 NVIDIA GTX 2080ti GPUs with batch size 96. To train our model, we use an Adam optimizer with a learning rate of 1e-4. The recognition process includes a-z alphabets and 0-9 digits. We use cross-entropy loss for estimate loss.

Table 1: Comparisons of STR performance across previous methods.

Model	IIIT5K	IC13	IC15	SVT	SVTP	CUTE
SE-ASTER [21]	92.8	93.8	80.0	80.0	81.4	83.6
Cheng et al. [1]	87.4	93.3	70.6	85.9	-	-
SAM [25]	93.9	95.3	77.3	90.6	82.2	87.8
Jiang et al. [33]	94.0	93.5	85.5	93.0	86.7	87.5
Sajid et al.[32]	95.5	96.3	83.8	90.8	86.0	86.9
ABINet[20]	96.2	97.3	86.0	93.5	89.3	89.2
SRN [22]	94.8	95.5	82.7	91.5	85.1	87.8
Zhang et al.[31]	96.5	97.7	85.4	93.0	89.3	91.3
SynthTIGER[26]	89.8	87.9	69.5	84.5	74.6	74.0
Ours*	**95.1**	**94.6**	**86.8**	**92.0**	**87.0**	**88.5**

Conclusion

In this paper, we proposed an efficient scene text recognition method that is end-to-end trainable. We claim the importance of robust and accurate visual features for scene text recognizer. RFPN framework improves the ability of the feature extractor, which generates robust visual features and semantic information. RFPN archives promising results on a different dataset. In the future, we will explore its potential.

Acknowledgment

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF)& funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A3B05049058 & NRF-2020R1A4A1019191).

References

1. Cheng, Fan, Yunlu, Gang, Shiliang , and Shuigeng. Focusing attention: Towards accurate text recognition in natural images. In ICCV, pages 5076–5084, 2017.
2. Zhou, Shuchang, Yao, Cao, and Yin. Icdar 2015 text reading in the wild competition. 2015.
3. Wojna, A. N Gorban, D. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz. Attention-based extraction of structured information from street view imagery. In ICDAR, volume 1, pages 844–850. IEEE, 2017.
4. Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles. Learning to read irregular text with attention mechanisms. In IJCAI, volume 1, page 3, 2017.
5. M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai. Mask textspotter: An end-toend trainable neural network for spotting text with arbitrary shapes. IEEE transactions on pattern analysis and machine intelligence, 2019.
6. Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In CVPR, pages 4168–4176, 2016.
7. Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence, 2018.
8. M.Yang, Y. Guan, M. Liao, X. He, K. Bian, S. Bai, C. Yao, and X. Bai. Symmetry-constrained rectification network for scene text recognition. In ICCV, 2019.
9. Chen, Jin, Zhu, Luo, Wang.:Text recognition in the wild: A survey. ACM ,2021.
10. Litman, R., Anschel, O., Tsiper, S., Litman, R., Mazor, S., Manmatha, R.: Scatter: selective context attentional scene text recognizer. In: CVPR. pp. 11962–11972 (2020)
11. Cheng, Z., Bai, F., Xu, Y., et al.: Focusing attention: Towards accurate text recognition in natural images. In: ICCV. pp. 5086–5094. IEEE (2017)
12. Li, H., Wang, P., Shen, C., Zhang, G.: Show, attend and read: A simple and strong baseline for irregular text recognition. In: AAAI. vol. 33, pp. 8610–8617 (2019)
13. Shi, Wang, Lyu, et al.: Robust scene text recognition with automatic rectification. In: CVPR. 2016.
14. Shi, B., Yang, M., Wang, X., et al.: Aster: an attentional scene text recognizer with flexible rectification. TPAMI (2018)
15. Wang, Hu,: Gated recurrent convolution neural network for ocr. In: NIPS. (2017)
16. Yang, Guan, Liao, et al.: Symmetry-constrained rectification network for scene text recognition. In: ICCV. (2019)
17. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: CVPR. pp. 12113– 12122 (2020)

18. Wang, Yuxin, et al. "From two to one: A new scene text recognizer with visual language modeling network." Proceedings of the International Conference on Computer Vision. 2021.
19. Fang, Shancheng, et al. "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition." Proceedings of the CVPR, 2021.
20. Baek, et al. "What is wrong with scene text recognition model comparisons? dataset and model analysis." *Proceedings of the international conference on computer vision*. 2019.
21. Zhi, Yu , Yang, Zhou, and Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. CVPR, 2020.
22. Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. CVPR, 2020.
23. Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In BMVC,2012.
24. Lin, Dollár, Girshick, He, Belongie. Feature pyramid networks for object detection; Proceedings of the 30th IEEE Conference on CVPR ,2017.
25. Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-toend trainable neural network for spotting text with arbitrary shapes. IEEE, 2019.
26. Yim, Moonbin, et al. "Synthtiger: Synthetic text image generator towards better text recognition models." *ICDAR*. Springer, Cham, 2021.
27. Kang, Jianjun, Mayire Ibrayim, and Askar Hamdulla. "MR-FPN: Multi-Level Residual Feature Pyramid Text Detection Network Based on Self-Attention Environment." *Sensors* 22.9 (2022): 3337.
28. Liu, Xi, et al. "Scene text detection with feature pyramid network and linking segments." *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019.
29. Liu, Fagui, et al. "FTPN: Scene text detection with feature pyramid based text proposal network." *IEEE Access* 7 (2019): 44219-44228.
30. Liang, Min, et al. "Scene text detection via decoupled feature pyramid networks." *International Journal on Document Analysis and Recognition (IJDAR)* (2022).
31. Zhang, Xinyun, et al. "Context-based Contrastive Learning for Scene Text Recognition." AAAI, 2022.
32. Sajid, Usman, et al. "Parallel scale-wise attention network for effective scene text recognition." *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.
33. Jiang, Hui, et al. "Reciprocal Feature Learning via Explicit and Implicit Tasks in Scene Text Recognition." *International Conference on Document Analysis and Recognition*. Springer, Cham, 2021.

An Implementation of the Odor Density Prediction Model Based on the Odor Density Level Data in Cheongju City

Seong-ju Joe[1], Woo-seok Choi[1], Sang-hyun Choi[2]

[1] Dept. Bigdata, Chungbuk National University, Cheongju, South Korea
[2] Dept. Management Information System, Chungbuk National University, Cheongju, South Korea
{fulans2, cdt3017}@naver.com, {chois}@cbnu.ac.kr,

Abstract. Due to the importance of controlling pollution levels through environmental monitoring, it has become a key area for governments and private organizations not only around the world but also in Korea. However, in emerging countries, poor implementation due to insufficient technology, and there are no government-managed areas, especially in Korea, Chungcheongbuk-do. In order to overcome these limitations, studies to predict the odor density are being actively conducted. In this study, a odor level prediction model was designed after 10 minutes, 30 minutes, and 60 minutes by using a machine learning algorithm. As a result of the analysis, the Random Forest Model was the best due to the characteristics of the odor.

Keywords: Odor, Environment Monitoring, Time Series, Machine Learning.

1 Introduction

In addition to the rapid expansion of industrial facilities, the increasing problem of environmental pollution due to environmental offices and various livestock farms is becoming an important social problem.[1] and, It is the subject of the biggest civil complaint caused by the indiscriminate emission of odors. Since odor is a problem directly related to the health of local residents, such as causing discomfort and disgust such as headache and vomiting, supportive services are needed to effectively, efficiently and accurately deliver odor-related information to local residents.[2].

The global environmental monitoring market size reached $18.8 billion in 2021. The International Mining and Resources Conference, a leading consultant on management strategies and market research around the world, estimates that it will grow at a 7.5 percent compound annual growth rate between 2022 and 2027 to reach $21 billion by 2025.[3].

In addition, the amount of odors generated in various environments is monitored in Korea. First, atmospheric TMS linked to the 'Korea Environment Corporation', including a study predicting the spread area. Second, temperature/humidity monitoring, which monitors the environment of a space where temperature and humidity are important during manufacture, such as food and medicine. Third, it manages resources by monitoring the amount of fuel and the amount of water/waste. As such, we are trying to provide various monitoring solutions.[4].

It organizes and implements environmental monitoring and friendly regulations and initiatives to promote and support environmentally friendly industries.[5]. However, despite favorable developments in regulations and standards, emerging countries still have poor implementation of inconsistent environmental regulations and pollution control reforms in their policy structures and implementations. In addition, high costs such as product purchase, installation, and maintenance can also affect the growth of the environmental monitoring market.[5].

As an example of this improvement, Sri Lanka's Central Environment Agency, the only organization that currently follows procedures for data collection across Sri Lanka, has installed mobile devices in major local villages and used in-device devices to collect environmental data throughout the week. However, since real-time air quality monitoring is not carried out, it is difficult for ordinary citizens to check air pollution status information.[6].

Based on the odor density level information collected through various sensors installed in the Chungbuk region, this study intends to implement a odor concentration prediction model by reflecting the area around the sensor as a reference. In particular, Chungbuk was excluded from 50 odor management areas nationwide designated by the Korea Environment Corporation, but the area was selected because of the need for systematic management through the designation of odor management areas[7]

2 Data preprocessing

In this study, the odor density data in Cheongju, Chungcheongbuk-do, provided by the Cheongju Public Data Portal, and the weather observation data provided by the Korea Meteorological Data Open Portal were used. Data from January 1, 2019 to October 27, 2022 were collected to design a model for predicting odor levels. In addition, six zones were designated to check whether the odor density in the adjacent area affects the odor density in the main area. Finally, a dataset of each zone was constructed. Each data is organized every 10 minutes from 0:00 to 23:50 every day.

This researcher made three derivative variables based on the date and time. First, seasonal variables were created by setting winter from November to February, exchange-interseason from March to May and October, and summer from June to September. Second, the time zone was set in density of the average concentration of TMA, and a time variable was created by setting the standard from 22:00 to 01:00 at night, from 2:00 to 5:00 in the morning, from 10:00 to 13:00 in the morning, from 14:00 to 17:00 in the afternoon, and from 18:00 to 21:00. Finally, a weekend variable was created, and Monday, Saturday, and Sunday, which have relatively low odor, were divided into weekends and the rest into weekdays. Temperature, wind direction, wind speed, and humidity information were collected as weather observation variables.

The dependent variable TMA is composed of categorical variables with a total of three classes, Level 0, Level 1, and Level 2, depending on the odor level. In this study, three dependent variables were selected to confirm the performance of the model that changes according to the future point of view to be predicted: the TMA measurement value after 10 minutes, the TMA measurement value after 30 minutes, and the TMA measurement value after 60 minutes.

The variables used in the odor level prediction model are as shown in [Table 1].

Table 1. The variables used in the odor level prediction model

Category	Variable name	Explanation	Type
Date Variables	Month_C	Seasonal variables (Summer, Winter, change_season)	Category
	Week_C	Week variables (Weekday, Weekend)	Category
	Hour_C	Time variabled (02-05, 06-09, 10-13, 14-17, 18-21, 22-01)	Category
Odor density Variables	CH3SH	CH3SH (Methyl Mercaptan) Measured value	numeric
	OU	OU (Odor Unit) Measured value	numeric
Weather observation Variables	Temp	a temperature value	numeric
	Wind_Speed	wind speed value	numeric
	Wind_Direction	Wind direction value	numeric
	Humidity	Humidity value	numeric
dependent Variables	TMA_10m	10 Minutes after TMA Measured value	Category
	TMA_30m	30 Minutes after TMA Measured value	Category
	TMA_60m	60 Minutes after TMA Measured value	Category

There were several sections that were not measured in the Odor data. So, we filled the date with empty values using interpolation. Data were interpolated by restoring the missing interval, which was filled up to 30 minutes up and down, but still exists a null value. Through interpolation and deletion of missing values, 145,451 data are finally left.

3 Analysis

The TMA to be predicted consists of three classes: Level 0, Level 1, and Level 2. However, the ratio of each class was 68%, 30%, and 2%, which had a very serious class imbalance. In particular, since it is important to correctly predict Level 2 in the odor level prediction model, the over-sampling technique SMOTE (Symtheic Minority Over-Sampling Technique) was applied to increase the data corresponding to Level 2 by 20 times.

Table 2. SMOTE-based oversampling result

As-Is			To-Be	
Class	Count(ratio)	→	Class	Count(ratio)
Level 0	79,714(68%)		Level 0	79,714(53%)
Level 1	34,795(30%)		Level 1	34,795(23%)
Level 2	1,851(2%)		Level 2	37,020(24%)

In this study, four algorithms were used: Logistic Regression, Random Forest, XGBoost (eXtreme Gradient Boosting), and Light GBM (Light Gradient Boosting Machine) to design a odor level prediction model that predicts TMA_10m, TMA_30m, and TMA_60m, respectively. and Performance evaluation used K-Fold Cross Validation, a technique for making learning data and evaluation data by crossing datasets several times.

Table 3 below shows the performance evaluation results of the odor level prediction model.

Table 3. Odor level prediction model performance evaluation table

Target	Model Accuracy (K-Fold=5)			
	Logistic Regression	Random Forest	XGBoost	Light GBM
TMA_10m	72.48%	91.12%	92.71%	**92.78%**
TMA_30m	71.81%	89.84%	90.29%	**90.45%**
TMA_60m	71.20%	87.76%	**89.62%**	89.48%

Table 3 shows the results of the odor level prediction model. Light GBM had the highest TMA level prediction accuracy of 92.78% and 90,45% after 10 minutes and 30 minutes, respectively. However, it can be seen that XGBoost is the highest at 89.62% for TMA level prediction after 60 minutes. For the prediction of TMA level after 60 minutes, XGBoost was the highest at 89.62%. For all algorithms, the prediction performance decreased by about 3% as the prediction time increased. However, considering that the prediction time point is six times farther away, it can be seen that the performance of the model is very high.

In general, it is important to correctly predict the bad odor level (Level 2) rather than the no odor level (Level 0). Table 3 shows that the performance of Light GBM is higher than that of other algorithms. On the other hand, Random Forest has a somewhat low overall accuracy, but Level 2 has excellent predictive accuracy. Therefore, this researcher decided that the use of Random Forest Model is the most effective.

Table 3. Light GBM Confusion Matrix

Light GBM		Actual Value		
		Level 0	Level 1	Level 2
Predict Value	Level 0	7,826	817	17
	Level 1	873	18,810	94
	Level 2	0	302	352
Class-specific accuracy		89.9%	94.4%	**76.0%**

Table 4. Random Forest Confusion Matrix

Random Forest		Actual Value		
		Level 0	Level 1	Level 2
Predict Value	Level 0	7,436	917	4
	Level 1	1,263	18,676	71
	Level 2	0	336	388
Class-specific accuracy		85.5%	93.7%	**83.8%**

4. Conclusion

In this paper, a machine learning algorithm was applied to predict the odor level after 10 minutes, 30 minutes, and 60 minutes, and the prediction accuracy was 92.78%, 90.45%, and 89.62%, respectively, showing high performance. In addition, it was confirmed that the odor level prediction model based on the random forest is somewhat inferior in accuracy, but the level 2 prediction performance is very high. Therefore, if the odor level prediction model designed in this study is used, it is expected that excellent odor monitoring will be possible.

References

1. Dae-seung Kim, Odor Removal Technology and Recent Trends in Industrial Sites
2. Ji-Hyuk Kang, A Study on the Electrophysiological Response of the Cerebral Cortex by Olfactory Stimulation: Alpha Activity,51(4):462-467(2019)
3. Inkwood Research, Global Environmental Monitoring Market Forecast 2021-2028(2021)3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. LG U+ Safe and clean management with environmental monitoring Wireless real-time monitoring of various environmental resources and harmful substances, https://www.lguplus.com/
5. INNOPOLIS Foundation, Environmental Monitoring Market(2020)
6. Monitoring of the atmospheric environment through crowdsourcing appears!, https://www.industrynews.co.kr/
7. Status of designation of odor management area(2020.02), https://www.keco.or.kr/

Analyzing Context and Speaker Memory using Pretrained Language Model for Emotion Recognition in Korean Conversation task

Dang-Khanh Nguyen[1,2], Hoai-Duy Le[1,2], Seok-Bong Yoo[1],
Guee-Sang Lee[1], Soo-Hyung Kim[1], and Hyung-Jeong Yang[1 (*)],

[1] Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea
{sbyoo, gslee, shkim, hjyang}@jnu.ac.kr
[2] {khanhnd185, hoaiduy1396}@gmail.com

Abstract. Emotion Recognition in Conversation is an essential phase in multiple applications, such as customer feedback understanding, social media threads, etc. Chonnam National University organized The 4th Korean Emotion Recognition International Challenge (KERC22), which aims to analyze the social behavior of each sentence of the speakers in a Korean conversation. This paper explains our solution to handle the task, our analysis, and our experimental results on the provided dataset. Our method utilizes the power of pretrained language model to obtain the conversation context and speaker memory information. The model achieves the F1 score of 76.50 (%) on the public test set of the KERC22. This result outperforms the baseline and performs top 7th in the competition.

Keywords: Emotion Recognition in Conversation, Korea Pretrained Language Model, BERT.

1 Introduction

Emotion Recognition in Conversation (ERC) is a common and complicated task in affective computing. There are many papers [1, 2, 3, 4] using commonsense knowledge to improve the performance of model in ERC task. However, most of the knowledge base is in English and infeasible to be applied in Korean conversation. Other papers [5, 6, 7] try to build conversation graphs to learn the inter and intra speaker relation in a conversation. Although this method performs well, the complexity of the graph increases together with the number of speakers in conversation.

In this paper, we introduce a straightforward solution for the Emotion Recognition in Korean Conversation. We exploit the enhancement of the pretrained Korean language models [16, 20] trained on many large Korean corpora using unsupervised learning. In our approach, these language models are fine-tuned to extract the speaker memory and contextual information. By fusing the conversation context and the history feature of the speaker, we can boost the accuracy of the model.

For the convenience, in this paper, we use the term “target utterance” to describe the utterance that we want to identify the sentiment or the emotion, and “target speaker” to call the person who speaks the target utterance. Additionally, we define the context including of scene context and conversation context. The scene context is considered as the scene description provided in the dataset. The conversation context includes all utterances from the beginning of the conversation to the target utterance.

2 Related works

The baseline of The 4th Korean Emotion Recognition International Challenge (KERC22) Dataset [8] is an attention-based model. The speaker context is generated from the context of target speaker and other speaker by using attention fusion. Similarly, the scene context is the combination of the scene description and the scene sentiment. Then, the late fusion layer merges the speaker context, scene context and the target sentence feature. The model also exploits the pretrained multilingual BERT [9] as a feature extractor.

Joosung Lee and Wooin Lee proposed CoMPM [10], a model that leverages the speaker memory and the contextual feature embedding to identify the speaker’s behavior. With a straightforward approach, the model still accomplishes good results with English conversation datasets compared to graph-based or commonsense knowledge approaches. The authors also run the model with Korean conversation. However, the detailed analysis and experiments are not discussed.

Xiaohui et. Al. proposed Emotionflow [11], a model that can capture the sequential information of emotions. It comprises a transformer-based encoder and a conditional random field (CRF) module. The pretrained RoBERTa model [12] is utilized to analyze the semantic context. The CRF layer explores the knowledge of the emotion transition in the sequence of utterances. The author conducted the experiments on the MELD, which is an English conversation dataset.

3 Proposed method

The proposed model comprises two branches: the context branch and the speaker-memory branch. The context branch computes the contextual feature of the conversation and the speaker memory branch generate the target speaker’s history information. A fusion layer is used to combine the outputs of two branches. The fusion layer is followed by a fully connected head to generate the final logits. The architecture of the model is described in Figure 1.

In context module, we tokenize the scene description and all utterances in conversation context into sequences of tokens, then, prepend each sequence a special token to mark the role of that sequence in the conversation. The sequence of the scene description is prepended with token <c0>, the sequence of the utterance is prepended with token <si> where the utterance is spoken by the ith speaker. Afterward, these sequences are stacked in series and prepended with a classification token <cls>. The final series of tokens are fed into a pretrained Korea language model to generate. The

embedding of the <cls> token is chosen as the context embedding of the whole conversation.

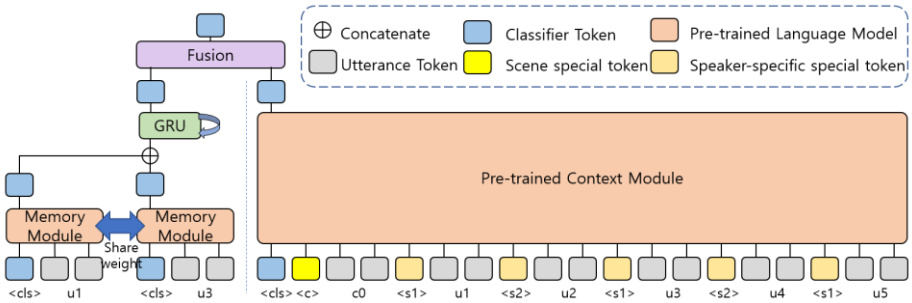


Fig. 1. The architecture of the proposed method. The example conversation is between 2 speakers with 5 utterances. The scene description c0 associates with special token <c>. Speaker 1 corresponds to special token <s1> and speaks utterance u1, u3, u5. Speaker 2 corresponds to special token <s2> and speaks utterance u2, u4. The figure describes the operation flow of the model to generate the prediction for target utterance u5. Speaker s1 would be the target speaker and u1, u3 would be the history utterance of target speaker.

In speaker memory module, all the history utterances of the target speaker are tokenized and extracted by another pretrained language model to obtain the latent feature embeddings. In each utterance, we take embedding of <cls> token as the speaker-specified utterance embedding. The sequence of these embeddings is fed into a Gated recurrent unit (GRU) neural network [18]. Finally, we take the embedding of the last utterance as the speaker memory embedding.

The outputs of two branches are then fused by a fusion layer to get the target utterance feature. A fully-connected layer head is used to produce the logit prediction from this feature. For fusion method, we suggest 3 options including the dot-product operand fusion, attention fusion [17] and summation fusion. Both embeddings come from the same pretrained language model, so it is feasible to simply sum up to get the fused output. Regarding the attention approach, we use context embedding as the output while paying attention to the behavior of the speaker memory embedding.

4 Dataset

We conducted the experiment on the KERC22 Dataset. The dataset consists of 12289 sentences from 1513 scenes of a Korean TV show named ‘Three Brothers’. In this dataset, we consider one scene as a conversation and one sentence as an utterance in the ERC problem. The database is split into train set, public and private test set with the approximated ratio of 3:1:1, respectively. Each sample includes a sentence ID, speaker’s name, the scene description, ID of the scene it belongs to, and the content of the sentence (or utterance). Each entry is labeled with a socio-behavioral emotional state, it could be euphoria, dysphoria, or neutral.

There are 907 conversations in the training set where 134 conversations have only one sentence, the remains have 2 to 64 sentences. A conversation may have zero, one or more scene description, adjacent sentences in a conversation may share the same scene description. Some sentences and some scenes have no description. The dataset suffers from the imbalance of label. Particularly, the number dysphoria, euphoria and neural samples in the training set are 4526, 1967, and 846, respectively.

5 Experiments

5.1 Experiment setting

We choose Pytorch as our development framework. The model is implemented and trained on the machine with NVIDIA RTX 2080 Ti GPU. The model is optimized by Adam Optimizer [19] with the learning rate of 1e-5 and the weight decay of 1e-2. The model only handles one sample in one iteration, which means the batch size equals 1. The performance metric for this ERC task is the micro F1 score and it is monitored in every epoch.

5.2 Experiment results

Compared to other ERC datasets, KERC22 has an additional field, which is the scene context. We would like to examine whether this new information is helpful for the task or not. In order to explore it, we conducted 3 experiments. In the first one, we used conventional dataset without the scene description and trained our model. In the next experiment, we considered the scene description together with the conversation context as the context and fed it to the context module. This configuration is the original idea of our model described in Section 3. In the final setting, we separated the scene context and conversation context, we used a brand-new pretrained language model extracting the scene context to obtain the scene context embedding. In this way, the fusion layer would have 3 inputs instead of 2 inputs as the original model. As shown in Table 1, merging scene and conversation context is the best among 3 options.

Table 1. Different ways to exploit the scene context and the scores on KERC22 public test set.

Model	Micro F1 score (%)
No scene context	74.47
Merging scene context and conversation context	**75.76**
Separating scene context and conversation context	74.12

Next, we would like to understand the role of loss function in the training process. We optimized the model parameter with 3 options of loss function including Poly loss [13], unweighted and weighted cross entropy (CE). The distribution of three labels in the training set is unbalanced. The ratio of dysphoria, euphoria and neural is 5.3 : 2:3 : 1, respectively. Therefore, the weighted CE is expected to balance the contribution of

each sample to the loss. As a result, the model trained with weighted CE loss achieved the best performance compared to unweighted CE and poly loss. The detailed result is described in Table 2.

Table 2. Performance of the models using different loss functions.

Loss function	Micro F1 score (%)
Unweighted CE loss	75.76
Poly loss	74.86
Weighted CE loss	**76.50**

Regarding the fusion method, we came up with 3 fusion techniques: attention fusion, scaled dot-product operand, and summation operand. In the attentive approach, we used the speaker context embedding as the query when the context embedding was considered as the key and value. The scaled dot-product operand is a simple version of the attention mechanism but using no learnable linear layer. Finally, the summation fusion simply sums up the context and speaker-memory embedding because they are both generated from a unify transformed-based architecture. Table 3 shows that the summation fusion gives the best performance among 3 fusion techniques.

Table 3. The performances of the model with different fusion techniques. The model is optimized with weighted CE loss function.

Fusion technique	Micro F1 score (%)
Scale dot-product fusion	75.57
Attention fusion	74.98
Summation fusion	**76.50**

After executing the above experiments, we finalized the optimized configuration for our model, which is merging the scene and conversation context, applying summation fusion and optimizing the model with weighted CE loss. We would like to draw a comparison between our method and other models. Table 4 shows the performance of the baseline, the Emotionflow and our method on the public test set of KERC22. Emotionflow's result is approximately higher than the baseline while our models outperform the remains. Additionally, our model with the electra [15] pretrained language model [14] accomplishes better score than the one using kobert [14].

Table 4. The performances of the models on KERC22 public test set.

Model	Micro F1 score (%)
Baseline	65.67
Emotionflow	66.88
Proposed method (kobert)	75.33
Proposed method (electra)	**76.50**

6 Conclusion

In this paper, we describe our transformer-based approach for the emotion recognition in Korean conversation. Our method is independent from the external knowledge base which is limited in Korean language. With the increasing development of pretrained language models, our model can be modified to operate with tasks on other languages. On the other hand, we also compared the performance of our model with different configurations of loss function and fusion technique. As a result, the proposed method achieves a significant improvement compared to the baseline of KERC22.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (NRF-2020R1A4A1019191).

References

1. Zhong, P., Wang, D., & Miao, C.: Knowledge-enriched transformer for emotion detection in textual conversations. arXiv preprint arXiv:1909.10681 (2019)
2. Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., & Poria, S.: Cosmic: Commonsense knowledge for emotion identification in conversations. arXiv preprint arXiv:2010.02795 (2020)
3. Li, J., Lin, Z., Fu, P., & Wang, W.: Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 1204-1214) (2021, November)
4. Zhu, L., Pergola, G., Gui, L., Zhou, D., & He, Y.: Topic-driven and knowledge-aware transformer for dialogue emotion detection. arXiv preprint arXiv:2106.01071 (2021)
5. Ishiwatari, T., Yasuda, Y., Miyazaki, T., & Goto, J.: Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7360-7370) (2020, November)
6. Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A.: Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. arXiv preprint arXiv:1908.11540 (2019)
7. Sun, Y., Yu, N., & Fu, G.: A discourse-aware graph neural network for emotion recognition in multi-party conversation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2949-2958) (2021, November)
8. The 4th Korean Emotion Recognition International Challenge, Chonnam National University, https://sites.google.com/view/kerc2022
9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, (2018)
10. Lee, J., & Lee, W.: CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. arXiv preprint arXiv:2108.11626 (2021)
11. Song, X., Zang, L., Zhang, R., Hu, S., & Huang, L. Emotionflow: Capture the dialogue level emotion transitions. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8542-8546). IEEE (2022, May)

12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
13. Leng, Z., Tan, M., Liu, C., Cubuk, E. D., Shi, X., Cheng, S., & Anguelov, D.: PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions. arXiv preprint arXiv:2204.12511 (2022)
14. Kiyoung Kim (2020): Pretrained Language Models For Korean, released on Github, https://github.com/kiyoungkim1/Lmkor
15. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
16. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.: Attention is all you need. Advances in neural information processing systems, 30 (2017)
18. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
19. Kingma, D. P., & Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Yang, K.: Transformer-based Korean Pretrained Language Models: A Survey on Three Years of Progress. arXiv preprint arXiv:2112.03014 (2021)

Classification Models of Online News Relevant to Animal Disease for Early Detection of Livestock Disease Outbreak

Saravit Soeng[1], Vungsovanreach Kong[1], HyungChul Rah[2]

[1] Department of Big Data, Chungbuk National University, Cheongju, South Korea
[2] Research Institute of Veterinary Medicine, Chungbuk National University, Cheongju, South Korea
{soengsaravit, kvsovanreach, hrah}@cbnu.ac.kr

Abstract. In recent years, the disease surveillance system has played a crucial role in preventing the spread of diseases. The surveillance system incorporates the process of data collection, analysis, and interpretation of data as well as detection, confirmation, and reports. The traditional surveillance system is mostly operated in a manual way which causes the delay in data availability and a lack of timeliness in detection. With the vast amount of data available online, it is very helpful to the surveillance system to get the data on time and work efficiently. In this research, machine learning-based approaches are proposed for the early detection of disease outbreaks based on online news. The machine learning algorithms like Naïve Bayes, Support Vector Machine, and Random Forest are applied for the classification task in this research. The results show that the models are sufficiently capable of classifying online news relevant to animal disease outbreaks.

Keywords: Early Detection, Livestock Disease, Disease Outbreak, Online News, Machine Learning

1 Introduction

Livestock diseases are a potentially terrible type of agricultural risk. Livestock diseases can seriously harm animal health as well as human health, and also have negative economic impacts by negatively affecting producer incomes, markets, commerce, and consumers [1], [2]. The livestock disease outbreak has made a major concern for the authorities and the public in taking action to prevent the spread of viruses. To lower risks and avoid significant economic losses, disease surveillance is very crucial. These efforts heavily rely on the ability to quickly and precisely detect an outbreak [2]. The surveillance systems participate in the processes of data gathering, analysis, and interpretation of the collected data as well as the detection, confirmation, and reporting of diseases, and also taking into account the public health response [3]. Traditional disease surveillance is mostly a manual procedure that causes a delay of one to two weeks in data availability and lacks timeliness in the detection [4], [5], [6]. The availability of web-based data sources has recently developed as an addition to

conventional surveillance systems, and by delivering real-time statistics and lowering the cost of public health, it has significantly helped with infectious disease surveillance [3], [7]. In this proposed research, we introduce machine learning-based approaches to detect livestock disease outbreaks from online news across various sources.

The proposed research focuses on three kinds of livestock diseases such as African swine fever (ASF), avian influenza (AI), and foot-and-mouth disease (FMD). The news data of these diseases are crawled from various internet-based news sources and then categorized into two groups, that is, relevant and irrelevant, where relevant news represents the disease outbreak and irrelevant news represents the other information of disease rather than the disease outbreak. The proposed method strongly relies on the use of machine learning techniques like Naïve Bayes, Support Vector Machine, and Random Forest for outbreak detection with the help of vector counts and TF-IDF approaches in the feature engineering step for transforming the text data into vector.

2 Proposed Methodology

The proposed methodology combines various components including data collection, data preprocessing, feature extraction, and machine learning model building. Fig. 1 represents the proposed method adopted for disease outbreak detection.

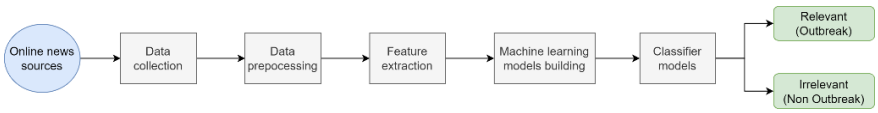


Fig. 1. Proposed method for outbreak detection

2.1 Data Collection

In this research, data is collected from various online news sources via Google News RSS feed and the official websites of the USDA Foreign Agriculture Service, USDA Animal and Plant Health Inspection Service, Food and Agriculture Organization of United Nations, and European Food Safety Authority by using crawling method. The collected data is from 2019 to 2022 with three types of diseases, such as ASF, AI, and FMD. The news data was then manually labeled as relevant and irrelevant, with which relevant news representing the disease outbreak and irrelevant news representing other information about the disease rather than the disease outbreak, by domain experts and computer scientists. After labeling, 360 news articles were determined to be relevant, while the other 360 were determined to be irrelevant.

2.2 Data Preprocessing

Applying preprocessing can enhance a dataset's quality in general and text classification performance in particular. Preprocessing can remove noise from the dataset. In some circumstances, the dataset's quality for text classification tasks can be enhanced by

applying preprocessing techniques such as stopword removal, punctuation mark removal, word stemming and lemmatization [8], [9]. In this study, the text data of news article contents are processed with various techniques to remove digits and words containing digits, non-ASCII characters, stopwords, and non-sense words, as those will not provide meaningful information for the model training. The preprocessing steps help to improve the processing step and also enhance the model efficiency.

2.3 Feature Extraction

To transform text data into numbers and to select the features, the vector counts and TF-IDF techniques are applied in the feature extraction steps. CountVectorizer creates count vectors as features, as the count vector is a matrix notation of the dataset in which every row represents a document from the corpus, each column corresponds to a term from the corpus, and each cell represents the frequency count of a specific term in a certain document [10]. And, TF-IDF score represents the relative importance of a term in the document and the entire corpus [10], [6].

2.4 Machine Learning Model Building

The final step is to train classifiers using the features created in the previous step. There are many choices of machine learning algorithms that can be used to build a model. In this research, we utilized some machine learning algorithms like Naïve Bayes, Support Vector Machine, and Random Forest for implementing the classification task. To build the classifiers, we applied the scikit-learn library to train each model with the labeled data that has been split into 80% and 20% for training and testing, respectively.

3 Results and Discussion

After training and evaluation, the selected models have produced similar results based on the evaluation metrics like precision, recall, F1-score, AUC score, and accuracy. Table 1 summarizes the performances of models based on particular evaluation metrics. Based on the results from Table 1, the Support Vector Machine model slightly outperforms other models in terms of precision, AUC score, and accuracy. The proposed models a bit outperformed the previous research [11] on the classification of news documents based on ASF in terms of accuracy. In addition, the proposed method is able to classify relevant news related to three different kinds of diseases (ASF, AI, and FMD) with the utilization of a single model.

Table 1. Performances of classifier models

Model	Precision	Recall	F1-Score	AUC-Score	Accuracy
Naïve Bayes	0.81	0.88	0.84	0.91	0.83
Support Vector Machine	0.88	0.82	0.85	0.92	0.85
Random Forest	0.82	0.88	0.85	0.91	0.84

4 Conclusion

This paper proposed machine learning approaches for the early detection of livestock disease outbreaks from online news. The machine learning algorithms like Naïve Bayes, Support Vector Machine, and Random Forest have been applied for the task of classification with the help of vector counts and TF-IDF techniques in this research. The models did produce good results in identifying animal disease outbreaks from the online news based on the performance evaluation. The proposed research is expected to help the authorities and health sectors timely detect the livestock disease outbreak and take action to prevent the spread of viruses. In the future, the research may also be applicable to extract the epidemiological information for an outbreak described in the news, such as disease name, outbreak location, date, host, and number of cases.

Acknowledgments. This work was supported by "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ0153412022)" Rural Development Administration, Republic of Korea, and partly by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant number: NRF-2020R1I1A1A01071884).

References

1. Inamura, M., Rushton, J., Antón, J.: Risk Management of Outbreaks of Livestock Diseases. (2015)
2. Bajardi, P., Barrat, A., Savini, L., Colizza, V.: Optimizing surveillance for livestock disease spreading through animal movements. Journal of the Royal Society Interface 9, 2814-2825 (2012)
3. Gupta, A., Katarya, R.: Social media based surveillance systems for healthcare using machine learning: A systematic review. Journal of Biomedical Informatics 108, 103500 (2020)
4. Lee, K., Agrawal, A., Choudhary, A.: Real-time disease surveillance using Twitter data: demonstration on flu and cancer. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1474–1477. Association for Computing Machinery, Chicago, Illinois, USA (2013)
5. Arsevska, E., Valentin, S., Rabatel, J., De Goër De Hervé, J., Falala, S., Lancelot, R., Roche, M.: Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. PLOS ONE 13, e0199960 (2018)
6. Amin, S., Uddin, M.I., alSaeed, D.H., Khan, A., Adnan, M.: Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches. Complexity 2021, 5520366 (2021)
7. Yan, S.J., Chughtai, A.A., Macintyre, C.R.: Utility and potential of rapid epidemic intelligence from internet-based sources. International Journal of Infectious Diseases 63, 77-87 (2017)
8. Hacohen-Kerner, Y., Miller, D., Yigal, Y.: The influence of preprocessing on text classification using a bag-of-words representation. PLOS ONE 15, e0232525 (2020)
9. Analytics Vidhya, https://www.analyticsvidhya.com/blog/2021/06/must-known-techniques-for-text-preprocessing-in-nlp/
10. Analytics Vidhya, https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/
11. Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B.: Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. Computers and Electronics in Agriculture 123, 104-115 (2016)

A novel product recommendation system for global market

Ji Won Jung[1], Soo Hyung Kim[1], Berdibayev Yergali[1], Kyu Ik Kim[1], Jinseok Lee[1], Jin Suk Kim[1]

[1] #704, 582, Daedeok-daero, Yuseong-gu, Daejeon, 34121, Rep. of KOREA

Abstract. Due to the growing market interest in reverse direct purchases, it is necessary to analyze global information about shopping malls in order to recommend reverse direct purchase products for small and medium-sized merchants who lack information about the international market. Therefore, this study is focused on the global market, in particular, consumption and current demand for products. The novelty of the study is to use the recent Learning to Rank algorithm from XGBoost Ranker method to display a ranking of the most requested products and/or categories for each country and each season, which can help sellers/producers understand which country to export to and the best time to do so. Furthermore, the AI model can recommend a suitable market based on classification approaches and a good selling price based on regression models.

Keywords: Recommendation system, Classification, Regression, Learning to Rank, XGBoost Ranker

1 Introduction

By 2030, about 80% of the world economy is expected to move to the global market, and exports of reverse direct purchase (RDP) products by overseas consumers who buy goods in local online stores are growing at an average of 55% per year.

As of 2021, RDP sales in South Korea exceeded 40 million and increased to 50.6% and 13.6 million.

Rising demand in RDB markets is driven not only by increased consumer preference for in-person purchases in the era of Covid-19, but also by the resurgence of online consumption patterns in the era of the global economy.

In particular, as the Korean wave of 2010 drew attention to Korean products such as cosmetics, fashion, and accessories, RDP markets in South Korea began to enter the global market. With online consumption trends accelerating, global major retailers are expanding their territory by rebuilding RDP platforms and expanding their sales networks, and since small merchants are also participating, 108,724 new Korean online merchants, including Amazon, increased their number to 38% from 79,033 sellers in the previous year [1].

This research aims to support the RDP business of small and medium merchants who are unable to follow the global market trends by providing them with global

shopping trends. Thus, we did an analysis of the global online market data and then built an AI recommendation model for RDP products with finding a suitable market and the best price to sell.

2 Data & Methodology

According to Statista[1], in the US as of June 2022, Amazon is the leader in e-commerce market share with 37.8% of the total market. A similar online platform, Taobao leads the local market among e-commerce companies in China, and Shopee in Vietnam.

Thus, we tried to collect relevant and latest data from leading local e-commerce companies for our study.

2.1 Data Preprocessing

Our data includes product code, product sales, rating, sales volume, number of reviews, and purchase dates. We have processed refinement to commonly utilize the data, and the process is as follows.

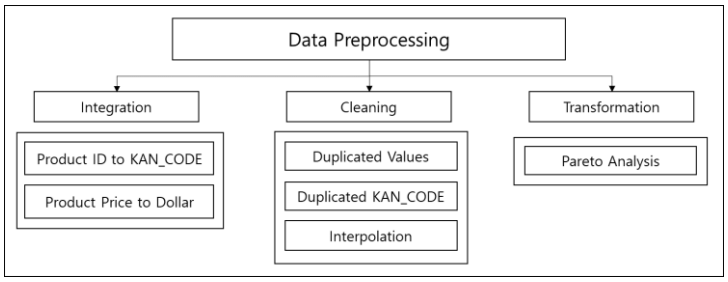


Fig. 1. Data Preprocessing framework

Since the product code is assigned independently of each country, we have standardized them based on Korean article numbers (KAN). Also, due to local currency differences, we exchanged all price values to USD as the standard currency in our study. Next, we remove duplicate transactions and also interpolate among the missing periods for each KAN code. Lastly, a Pareto distribution was performed to determine the relevancy score to be used as the target of the product recommendation.

[1] Statista: https://www.statista.com/statistics/274255/market-share-of-the-leading-retailers-in-us-e-commerce

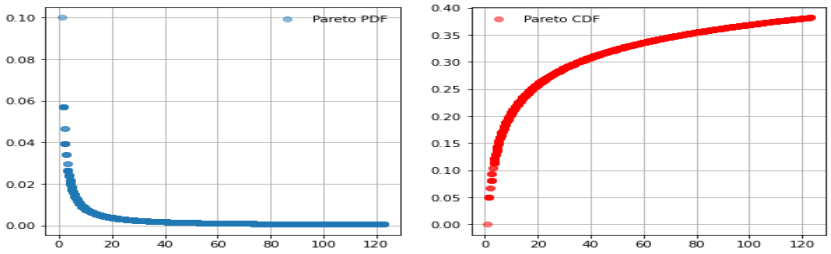


Fig. 2. Pareto distribution

2.2 Feature Extraction

Common information such as KAN code, product sales, and ratings were extracted as basic features to recommend sales products, and statistical indicators such as monthly average, maximum, minimum, and median were calculated in various ways to extract more meaningful features.

Furthermore, in order to measure the trend, we calculated the moving average value by three methods such as simple (SMA), cumulative (CMA) and exponential moving average (EMA) estimation for measuring an increase rate compared to the previous month. As a result, we have the importance of these features as shown below.

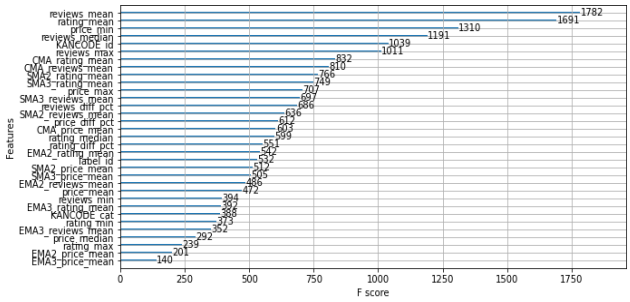


Fig. 3. Feature Importance analysis

3 Model

We approached the Learning to Rank algorithm as the XGBoost Ranker model and a classification method to find a suitable market for sellers and a regression model for a good selling price.

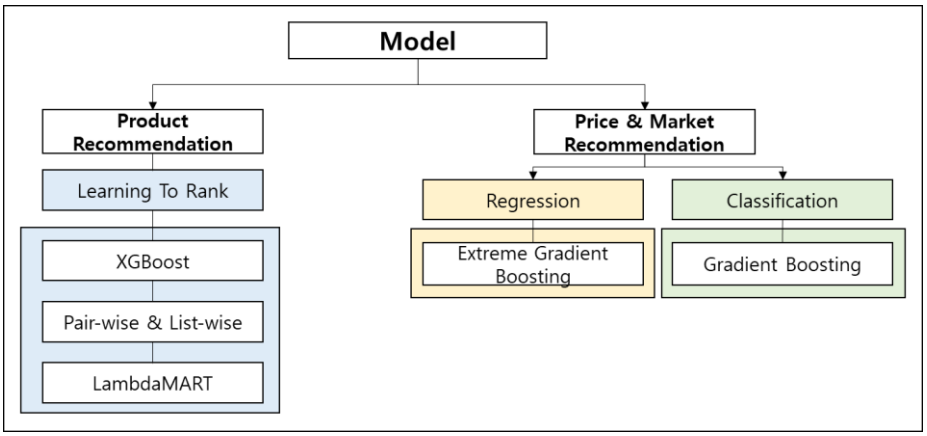


Fig. 4. Recommendation system model

3.1 Product recommendation

For product recommendation, the optimal order was selected based on the relevance between objects using Learning to Rank (LTR), a machine learning supervised learning technique known to be excellent in the recommended area.

Nowadays there are three types of approaches in LTR, such as Point-wise, Pair-wise, and List-wise [2, 3].

- *Point-wise* is a method of calculating scores and sorting lists by considering only one item at a time, and has the disadvantage of not fully utilizing the simplest method or the entire information of the list.
- *Pair-wise* is a method of deriving the optimal order through comparison between the two in consideration of a pair of items at a time, and in this case, it has an advantage that its performance is better than Point-wise.
- *List-wise* is a method of deriving the optimal order by utilizing the entire information in the list, which is more complicated than other methods, but good performance can be expected.

We used Extreme Gradient Boosting (XGBoost) algorithm, in particular, new model as XGBRanker from open source XGBoost python package by using parameter as LambdaMART, because it has been proven to be highly successful algorithms in solving real-world ranking problems [4].

To evaluate a model, we used Normalized Discounted Cumulative Gain (NDCG) metrics, to minimize pair-wise losses, shown in *Equation 1*.[5].

$$obj = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (1)$$

As result, we have high performance in both approaches, with 98% in NDCG metrics (which can be considered as accuracy in Learning to rank), shown in Table 1.

Table 1. Model results

	MSE	*RMSE*	*NDCG*
Pair-wise model	*10.39*	*3.21*	*0.98*
List-wise model	*13.51*	*3.67*	*0.98*

3.2 Price & Market Recommendation

This part of the model consists of two parts: a price recommendation and a market recommendation. Because we are targeting two areas related to price. Thus, the approaches also differ as a regression task to find a good price and a classification task to find a suitable market.

As a target for the regression problem, we used the average selling price of each product, and for the classification problem, we used price groups which categorized by the KMeans clustering method. As a result, we had four price levels, which we labeled as Economy, Regular, Premium, and Luxury.

KMEAN_labels	items	price	rating	reviews	RATIO, %
Luxury	1	580.00	0.00	0.00	0.88
Premium	10	75.84	2.31	83.70	8.77
Regular	37	34.39	2.89	2331.70	32.46
Economy	66	13.82	3.67	944.85	57.89

Finished KMEAN labeling for "32007140000"

Fig. 5. Market levels based on KMeans clustering in one product

Regression and classification were done in the open source PyCaret Python packages. Our final results in both approaches shown in Table 2 & 3.

Table 2. Regression model results

Model	*MAE*	*MSE*	*RMSE*	*R2*
Extreme Gradient Boosting	*12.5863*	*119978.38*	*121.9403*	*0.921*
Gradient Boosting Regressor	*12.8837*	*121629.98*	*121.8854*	*0.9206*
Decision Tree Regressor	*14.1816*	*120291.37*	*129.6951*	*0.9118*
Random Forest Regressor	*14.0503*	*126779.31*	*139.9091*	*0.8906*
AdaBoost Regressor	*37.0856*	*124430.79*	*150.5369*	*0.8441*

Table 3. Classification model results

Model	*Accuracy*	*AUC*	*Recall*	*Prec.*	*F1*	*Kappa*	*MCC*
Gradient Boosting Classifier	*0.789*	*0.9371*	*0.7901*	*0.7953*	*0.7905*	*0.7185*	*0.7196*
Extreme Gradient Boosting	*0.7802*	*0.9379*	*0.7816*	*0.784*	*0.7809*	*0.7069*	*0.7078*
Logistic Regression	*0.7781*	*0.9314*	*0.7797*	*0.7736*	*0.7744*	*0.7042*	*0.7053*
Random Forest Classifier	*0.7723*	*0.9274*	*0.774*	*0.7707*	*0.7709*	*0.6965*	*0.6969*
Extra Trees Classifier	*0.764*	*0.9213*	*0.7656*	*0.7608*	*0.7611*	*0.6854*	*0.6863*

In addition, we have compiled a confusion matrix to see the result of the clas sification model for recommending suitable price groups, as shown in Fig. 6.

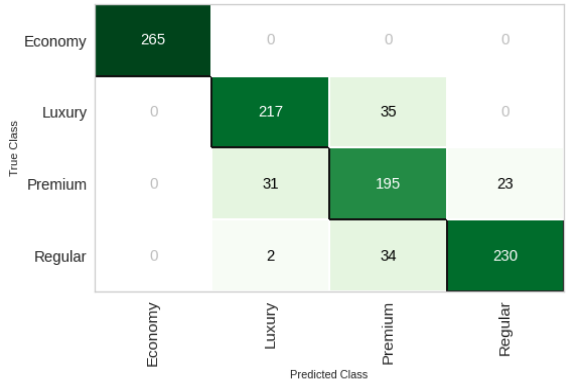


Fig. 6. Confusion Matrix

4 Conclusion

In conclusion, the results of this study can be used as an index to determine the possibility of exporting a specific product by providing recommending sales prices and products for each country to small and medium-sized sellers, who look for RDP business.

On the other hand, since this study used only shopping mall data, especially data on the cosmetics sector, among the various data collected, it is necessary to further expand the scope to the clothing and electronics sectors. Furthermore, it is considered necessary to establish a recommendation system that provides more informative analysis by utilizing other collected data. Therefore, this study helps to understand the implementation of the Learning to Rank algorithm from the XGBoost Ranker approach.

References

1. Korea International Trade Association news (2022-10-21). https://www.kita.net/cmmrcInfo/cmmrcNews/cmmrcNews/cmmrcNewsDetail.do?pageIndex=1&nIndex=71143&sSiteid=1
2. Liu, T. Y. (2009). Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval, 3(3), 225-331.
3. Dive into Deep Learning https://d2l.ai/chapter_recommender-systems/ranking.html
4. Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. Learning, 11(23-581), 81.
5. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

A Fault Diagnostic Model for Electric Rotating Machines

Soo Hyung Kim[1], Yergali Berdibayev[1], Jongho Park[1], Jiwon Jung[1], Seung Hyeok Lee[1], Kyuik Kim[1], Hyeongki Jo[1], Chieteuk Ahn[1], and Jinsuk Kim[1]

[1] Neoforce Co. Ltd., #704, 582, Daedeok-daero, Yuseong-gu, Daejeon, 34121, Rep. of KOREA

Email: mrjommer@naver.com, yergali.nf@gmail.com, breakprejudice@naver.com, isl452145@gmail.com, leeshyk@hanmail.net, han_bando@naver.com, gudrl0517@gmail.com, ahncteuk@gmail.com, 0210kimjs@daum.net

Abstract. This paper aims to develop AI fault diagnosis and prediction model for a real-time monitoring system that collects and analyzes signal data from electric rotating machines in power. The model focuses on the most common defects of a rotating machines: bearing, stator, and rotor. Two common signal analysis techniques — temporal and frequency analysis — and a novel way of using Convolutional Neural Network algorithm were applied to the acquired data for better performance.

Keywords: Prognostics and Health Management, Electric Rotating Machines, Classification Model, Artificial Intelligence, Convolutional Neural Network

1 Introduction

As power generation facilities face operating conditions such as high speed, high load, and high temperature, they are likely to cause major accidents in case of damage or damage in small areas. In particular, the rotor part of the power generation facility is a part that has potential elements of danger and large-scale accidents; therefore, it can be said to be a representative case of early detection.

In order to ensure the reliability of power generation facilities, time-based maintenance (TBM) has been performed each certain period, but this has caused excessive maintenance costs and has shown limitations in preventing sudden failures and accidents.

Unlike TBM, status-based maintenance with AI fault diagnosis and prediction technology can enable stable and reliable system management at low maintenance costs by detecting abnormalities in power generation facilities and predicting future failures to take proactive measures.

Also, according to a global market trend report from Innopolis in South Korea, the market for prognostics and health management using artificial intelligence in the global energy and utility sector is expected to increase from $778 million in 2020 to 27.1% annually, reaching $2.582 billion in 2025.

The purpose of this study is to develop AI fault diagnosis and prediction model for a real-time monitoring system that collects and analyzes signal data from electric rotating machines in power. Specifically, this research deals with creating an experimental device that simulates power plant facilities, acquiring vibration and current sensor signals, preprocessing data, extracting features, and developing an AI fault diagnostic model.

2 Experimental Setup

A plant facility consists of numerous rotational equipment and components such as turbines, generators, pumps, fans, and etc. In this study, we constructed an experimental device simulating a generator to establish a data acquisition environment for an electric rotary facility. As shown in Figure 1, it was built with a motor, a shaft, bearings, and load parts including a stator and a rotor. Also, we attached vibration and current sensors to collect normal and fault status data.

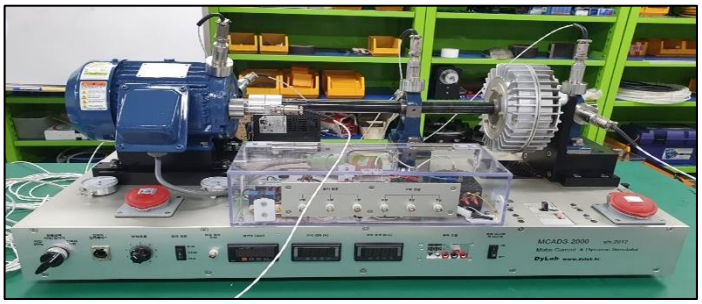


Fig. 1. Experimental Device (MCADS-2000)

The experimental device is composed of a motor, which has a power frequency of 60 Hz, four poles, 2HP, 220/380 V, 6.4/3.7 A 36 stator slots, and 28 rotor bars. The fault types that we have targeted to diagnose by the AI model are described in Table 1.

Table 1. Fault Types of Electric Rotating Machine

Component	Fault Type
Bearing	Outer Race, Inner Race, Rolling Element, Cage
Stator	Shorted-turn Winding, Single Phasing
Rotor	Broken Rotor Bar, Eccentricity

3 Data Acquisition and Preprocessing

3.1 Data Acquisition

Defective electrical rotating machines create abnormal noise and vibration due to friction or distortion of the defective parts. Also, defects in electrical devices affect the flow of motor current. Therefore, we collected vibration and current signal data by running the experimental device for 200 seconds at a speed of 1800 RPM to determine a normal status and eight fault status.

To collect vibration data, we used ADRE SXP 408 of Bentley with a velocimeter vibration sensor of General Motors. As shown in Figure 1, two sensors were attached perpendicularly on the frames of each bearing, and three sensors were attached on the motor in the XYZ axis. For current signal data, we connected cables on current output BNC connectors. The specifications for acquired data are listed in Table 2.

Table 2. Dataset Specification

Category	Sampling Rate	Time(s)	Dataset Name	# Datasets
Vibration	12800Hz	200	Normal Motor, Faulty Bearing, Faulty Stator, Faulty Rotor	9
Current	2000Hz	200		9

3.2 Data Acquisition

In the data processing process, we used a discrete signal separation method to remove Gaussian noise caused by shaft rotation and demodulation band selection to change shapes of the signals to help the data analysis process. The general process of data preprocessing is illustrated in Figure 2.

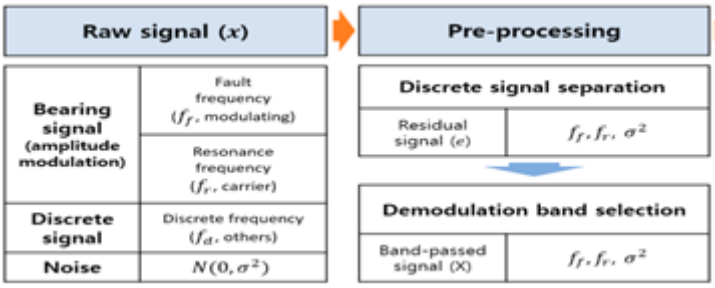


Fig. 2. Signal Preprocessing Process

3.2.1 Preprocessing for Vibration Signals

When measuring vibration signals of the motor, the sensors also catch unwanted signals, such as discrete signals, from other components connected to the motor. Therefore, it is necessary to remove the noises such as discrete signals to correctly analyze the

acquired data. Autoregressive (AR) Model, is often used as a tool to remove the signals (Sim et al., 2020). This model determines the result by linearly relying on its own previous value and on a stochastic term, which means that it can reproduce a discrete signal, which is more consistent and predictable than defect causing signals. The mathematical equation for the AR Model is shown in *Equation 1*. Then, we can isolate defect signals, or fault signals, by subtracting discrete signals from the original signal.

$$y_t = c + \phi_1 y_{t-1} + \phi_1 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t. \quad \textbf{(1)}$$

Furthermore, to clearly extract the defect signal from the remaining residual signal after the discrete signal is removed, we find a suitable band for analyzing the defect signal using Spectral Kurtosis as a demodulation band selection technique. Spectral Kurtosis (SK) is a statistical indicator that represents the relationship between signal impulses and frequency and can be used as the technique when analyzing defect frequency in rotating equipment such as bearings (Kim et al., 2020). The SK of the resonant frequency around the defect frequency is higher, so we modified the waveform according to the value of SK, so that the preprocessed data is useful for analyzing fault signals.

3.2.2 Preprocessing for Current Signals

Unlike vibration data, current data showed a trend of more constant waveform and modest changes in peaks; therefore, it was difficult to predict discrete signals using AR Model. Instead, we used a discrete wavelet transform (DWT) filter to remove its discrete signals. Wavelet transforms are signal decomposition techniques that find correlation and importance of a signal by changing size and position of its waveform. Among them, a DWT can be used for removing discrete signals by making a low-pass filter (Dolabdjian et. al 2002).

We used DWT as a low-pass filter while using Daubechies 4 as a base function for removing the unwanted signals and then applied SK demodulation. The comparison between preprocessed data for vibration and current is illustrated in Figure 3.

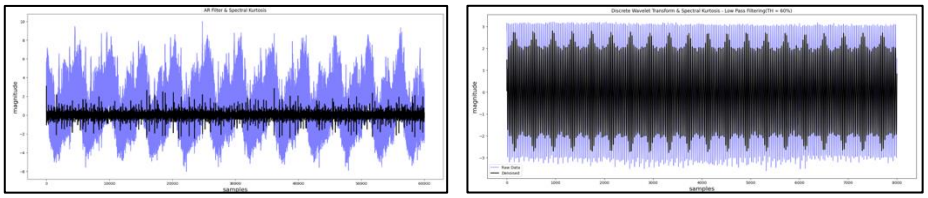


Fig. 3. Vibration Preprocessing vs Current Preprocessing

4 Feature Extraction and Diagnostic Model

In this research, we extracted features for the diagnostic model by applying time and frequency analysis techniques and also Convolutional Neural Network (CNN) method, and then we selected ten features, which showed higher feature importance.

4.1 Temporal Analysis

Time-domain signals can be analyzed and extracted as features by calculating statistical indexes of the data (Sim et al., 2020). Therefore, we used mean, standard deviation, skewness, root mean square, crest factor, etc. as temporal analysis features for the model. The list of the indexes is shown in Table 3. We created eleven indexes from both vibration and current signals to analyze their trends. After we created features, we normalized the values by using the z-score method to form a train dataset.

Table 3. Descriptive Statistics

Abbreviation	Full Name	Brief Explanation	Formula
MEAN	Mean	Average	$\frac{\sum X}{N}$
RMS	Root mean square	Value that generally tends to get bigger as the degree of fault in the bearing increases	$\sqrt{\frac{\sum X^2}{N}}$
STD	Standard deviation	Value representing the dispersion of a signal	$\sqrt{\frac{\sum (X-\overline{X})^2}{N-1}}$
PEAK	Peak	Maximum value of signal absolute value	$\max(\lvert X\rvert)$
SK	Skewness	The asymmetry of the probability density function of the vibration signal	$\frac{\frac{1}{N}\sum (X-\overline{X})^3}{STD^3}$
KUR	Kurtosis	The sharpness of the probability distribution of the vibration signal, and if this value is close to 3, it is closer to normal distribution	$\frac{\frac{1}{N}\sum (X-\overline{X})^4}{STD^4}$
CF	Cres factor	The ratio of peak values to the RMS of a signal	$\frac{PEAK}{RMS}$
CL	Clearance factor	Peak value divided by the square of root mean	$\frac{\max(\lvert X\rvert)}{\left(\frac{\sum \sqrt{\lvert X\rvert}}{N}\right)^2}$
SF	Shape factor	RMS divided by mean	$\frac{RMS}{MEAN}$
IF	Impulse factor	The ratio of peak values to the mean of a signal	$\frac{PEAK}{MEAN}$
P2P	Peak to peak	The difference between maximum and minimum values of the signal	$\max(X) - \min(X)$

4.2 Frequency Analysis

Time-domain data can be transformed to frequency domain by using the signal decomposition method called Fast Fourier Transform (FFT). We applied envelope analysis to remove any modulated amplitude by the high resonance signal around the defect causing frequency before we applied FFT. That is, we transformed the preprocessed signal by using the Hilbert equation to smooth the signal by enveloping the waveform (Sim et al., 2020). The following is the equation of the envelope method, where $x(t)$ is the original signal and $\hat{x}(t)$ is the enveloped signal.

$$x_{analytic}(t) = x(t) + j\hat{x}(t) \quad (2)$$

Table 4. Fault Frequency Equations

Component	Fault Type	Fault Frequency Equation
Bearing	Inner Race	$BPFI = \frac{Nb}{2} S\left(1 + \frac{BD}{PD} cos\beta\right)$
	Outer Race	$BPFO = \frac{Nb}{2} S\left(1 - \frac{BD}{PD} cos\beta\right)$
	Ball	$BSF = \frac{Pb}{2Bd} S\left\{1 - \left(\frac{Bd}{Pd}\right)^2 (cos\beta)^2\right\}$
	Cage	$FTF = \frac{S}{2}\left(1 - \frac{BD}{PD} cos\beta\right)$
Stator	Short-turned Winding	$f_x = f_i\left\{\frac{n}{p}(1-s) \pm k\right\}$
	Single Phasing	
Rotor	Broken Rotor Bar	$f_x = f_i(1 \pm 2ks)$
	Eccentricity	$f_x = f_i\left\{1 \pm \frac{k(1-s)}{p}\right\}$

In the frequency domain, fault signal frequency of each fault type in our models can be calculated as shown in Table 4. After we specified the fault signal frequencies, we used them as frequency analysis features as well as root mean square from low frequency range, which is between 0 ~ 500Hz, and high frequency range, which is between 0 to its maximum Hz. Totally, we created eleven features from the frequency-domain data.

4.3 Convolutional Neural Network Analysis

While time and frequency domain analysis methods are well-known methodology to diagnose failures of rotating machines, we also considered that a significant analysis could be possible by comparing the shapes of waveform of normal-state and fault-state data. Therefore, we designed a Convolutional Neural Network to extract features since CNN is specialized in detecting and understanding patterns (Krizhevsky et al., 2017). To effectively use CNN, it is necessary to transform one-dimensional data into a two-dimensional matrix (Li et al., 2017). Therefore, we created sixteen features from CNN decomposition, as illustrated in Table 5.

Table 5. CNN Model Architecture

Layer(type)	Output Shape	# Params
Conv2d(Conv2D)	(None,256,500,512)	25600
Max_pooling2d(MaxPooling2D)	(None,128,250,512)	0
Batch_normalization(BatchNormalization)	(None,128,250,512)	2048
Conv2d(Conv2D)	(None,128,250,256)	3277056
Batch_normalization_6(BatchNormalization)	(None,128,250,256)	1024
Conv2d(Conv2D)	(None,128,250,128)	295040
Max_pooling2d(MaxPooling2D)	(None,64,125,128)	0
Batch_normalization(BatchNormalization)	(None,64,125,64)	512
Conv2d(Conv2D)	(None,64,125,64)	32832
Batch_normalization (BatchNormalization)	(None,64,125,64)	256
Global_average_pooling2d(GlobalAveragePooloing2D	(None, 64)	0
Dense	(None, 16)	1040

4.5 Diagnostic Model

To diagnose failures of electric rotating machines, we designed the classification models for bearing, rotator, and stator by constructing training dataset for each model, as illustrated in Figure 4.

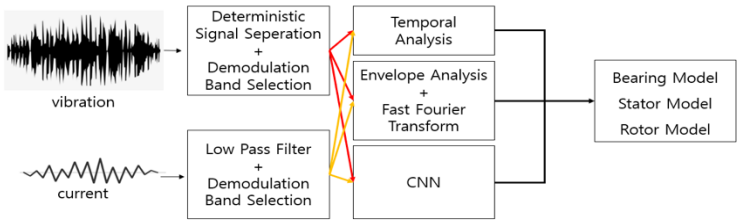


Figure 4. Diagnostic Model Pipeline

We had chosen classification models that are noted as efficient in supervised classification problems: Extra Trees, Extreme Gradient Boost, Random Forest, CatBoost (El Menzhi & Chiementin, 2022). Then, to reduce deviation and biased results, we normalized the dataset by using z-score method, and to validate our result, we cross-validated five times and averaged the result.

4.6 Model Result

Finally, to evaluate the performance of the diagnostic models, we compared accuracy, precision, and recall that are performed by the aforementioned, five classification models.

As shown in the Figure 5, 6, and 7, the CatBoost classification model showed the highest accuracy, precision, and recall for the bearing model, while the Random Forest classification model and Extra Trees classification model performed well in the stator model and rotor model. Overall, all of the three diagnostic models performed well to diagnose abnormality in the electric rotating machines.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.9605	0.9984	0.9605	0.9620	0.9608	0.9506	0.9509	0.8120
rf	Random Forest Classifier	0.9598	0.9980	0.9598	0.9617	0.9601	0.9498	0.9501	0.2480
et	Extra Trees Classifier	0.9592	0.9979	0.9592	0.9605	0.9595	0.9489	0.9491	0.5740
xgboost	Extreme Gradient Boosting	0.9585	0.9982	0.9585	0.9594	0.9586	0.9481	0.9483	0.6440
svm	SVM - Linear Kernel	0.8175	0.0000	0.8175	0.8281	0.8121	0.7719	0.7773	0.2060

Figure 5. Bearing Model

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9490	0.9969	0.9490	0.9501	0.9491	0.9236	0.9240	0.0820
catboost	CatBoost Classifier	0.9471	0.9970	0.9471	0.9484	0.9471	0.9207	0.9213	0.5840
xgboost	Extreme Gradient Boosting	0.9433	0.9968	0.9433	0.9447	0.9434	0.9149	0.9155	0.7960
et	Extra Trees Classifier	0.9423	0.9970	0.9423	0.9439	0.9425	0.9135	0.9141	0.4800
svm	SVM - Linear Kernel	0.9413	0.0000	0.9413	0.9426	0.9414	0.9120	0.9126	0.0260

Figure 6. Stator Model

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.9462	0.9963	0.9462	0.9468	0.9461	0.9192	0.9196	0.0720
xgboost	Extreme Gradient Boosting	0.9442	0.9947	0.9442	0.9447	0.9442	0.9163	0.9166	0.7760
rf	Random Forest Classifier	0.9413	0.9955	0.9413	0.9418	0.9413	0.9120	0.9123	0.0780
catboost	CatBoost Classifier	0.9394	0.9950	0.9394	0.9409	0.9395	0.9091	0.9098	0.5920
svm	SVM - Linear Kernel	0.7837	0.0000	0.7837	0.8089	0.7785	0.6755	0.6888	0.0220

Figure 7. Rotor Model

5 Conclusion

In the future, we will advance the method proposed in this paper to develop a system that diagnoses possible failures not only in electric rotating machines but also in mechanical ones and monitors the condition of power plants and predicts future failures.

References

1. Innopolis Report (2021). Prognositics and Health Management Market, *Global Market Trend Report*, 8
2. Sim, J., Kim, S., Park, H. J., & Choi, J. H. (2020). A tutorial for feature engineering in the prognostics and health management of gears and bearings. *Applied Sciences*, *10*(16), 5639.
3. Kim, S., An, D., & Choi, J. H. (2020). Diagnostics 101: A tutorial for fault diagnostics of rolling element bearing using envelope analysis in MATLAB. *Applied Sciences*, *10*(20), 7302.
4. Dolabdjian, C., Fadili, J., & Leyva, E. H. (2002). Classical low-pass filter and real-time wavelet-based denoising technique implemented on a DSP: a comparison study. *The European Physical Journal-Applied Physics*, *20*(2), 135-140.
5. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.
6. Li, Z., Qin, Z., Huang, K., Yang, X., & Ye, S. (2017, November). Intrusion detection using convolutional neural networks for representation learning. In *International conference on neural information processing* (pp. 858-866). Springer, Cham.
7. El Menzhi, L., & Chiementin, X. (2022). Supervised Classification Of induction Motors faults. In *E3S Web of Conferences* (Vol. 336, p. 00051). EDP Sciences.

Detection of bubble defects in contact lenses using YOLOv5

Sung-hoon Kim, In Joo, Gi-nam Kim, Kwna-Hee Yoo,
Dept. of Computer Science, Chungbuk National University, South Korea
{sidsid84, jooin95,kgn4192,khyoo}@chungbuk.ac.kr

Abstract. A digital transformation is in progress in various manufacturing fields, and the introduction of big data and artificial intelligence technologies is also in progress in contact manufacturing companies.Consequently, artificial-intelligence-based methods for the accurate detection of defects occurring in contact lens manufacturing processes have been developed. In this study, we detected the location and frequency of bubble defects in contact lenses using the YOLOv5 model. We trained the model using 10,485 images to obtain consistent results for various lens types. As a result of a test on 300 images, an accuracy of 96% was obtained at the 0.257 position, which is the confidence score in the actual test.

Keywords: deep-learning, contact-lens, object detection, smart factory

1 Introduction

With the advent of the fourth industrial revolution, government and private companies have presented increased interests in the introduction and development of cutting-edge technologies such as big data and artificial intelligence (AI). Under these circumstances, such technologies have created new value-added services and enabled profit realizations. Along with such technological developments, smart factories that apply advanced technologies to various manufacturing fields have also been introduced [1].

The application target of this study was a company specializing in contact lens manufacturing; notably, this company is currently in the process of digital transformation. In particular, we evaluated the application of AI to the contact lens inspection process, which is the last stage of the company's product production.

The image data collected during this process are shown in Fig 1. The object classes included two types of bubbles and an airBubble, and the bubble object was classified as a defect. The airBubble was a temporary impurity that appeared during the imaging of the inspection equipment and was classified to be normal.

In addition, as depicted in Fig 2, a pair of original black and white image data with different lighting for one lens was collected.

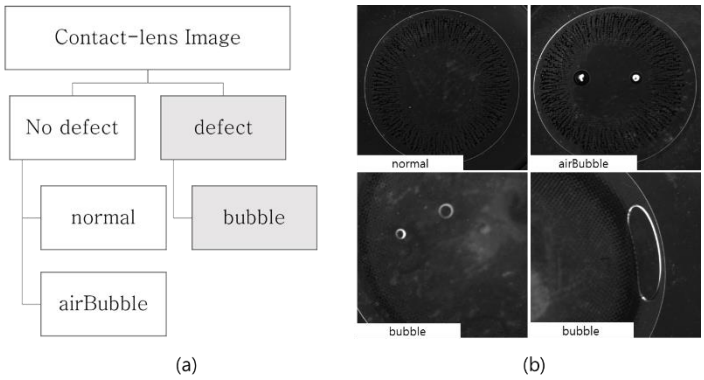


Fig. 1. (a) Classification of contact lens defects. (b) Examples of classes

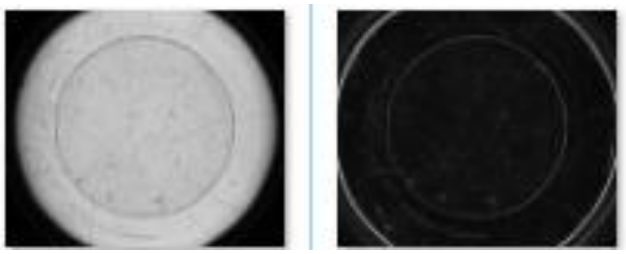


Fig. 2. Pair of original data with different lighting

Currently, the aforementioned inspection process relies on manual sorting; therefore, the entire production cannot be inspected, and the accuracy remains at approximately 85%. We aimed to achieve a processing speed of less than 1.4 s per product and an accuracy of more than 90% with the incorporation of AI into the process.

Notably, several attempts have been made in the field of contact lenses with the foregoing objective. Sophia et al. conducted a study using deep learning to classify non-lens, contact lens, and cosmetic lens wear [2]. Raghavendra et al. [3] proposed a convolutional neural network-based ContlensNet for the same purpose and recorded a high accuracy. Zin et al. [4] classified the boundary line of a lens, and based on this, they identified whether the lens was used based on a Support Vector Machine(SVM).

Generally, the common objective of previous studies was to identify whether contact and cosmetic lenses were used, which are factors interfering with the iris recognition technology. Therefore, recognizing the boundary of a lens was the focus of these studies.

However, because our study aimed to detect defects in contact lenses, a new approach was required.

2 Contact lens defect detection using YOLOv5

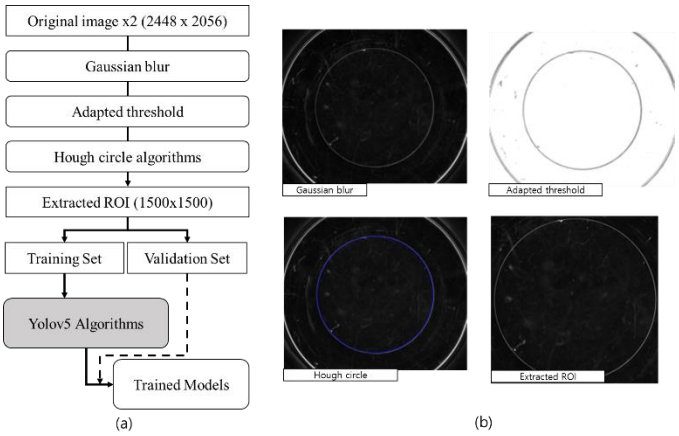


Fig. 3. (a) Flow chart of defect detection. (b) Description of changes in the image according to each step

As depicted in Fig 3, the image data consisted of a pair of grayscale images, and the original image size was 2448 × 2056 pixels. The original image was subjected to a noise removal process using a Gaussian blur filter, and following this, the feature was maximized for the circular boundary through an adaptive threshold process. For the modified image, a circular region of interest (ROI) area was identified using the Hough circle detection algorithm. Subsequently, the ROI was extracted with a margin of 40 pixels. After executing the same process for white and black images, a pair of images was concatenated to one image with two channels.

We used the YOLOv5 object detection algorithm to analyze the prepared ROI image. In particular, YOLOv5 slices an image and adds a cross-stage partial network (CSPNet) to the backbone network. Consequently, YOLOv5 significantly reduces the skeleton of the network system. With its lightweight model size, the object recognition speed can be as high as 140 fps when running on a server.[5][6].

We divided the 10,485 prepared data into 8,388 images for training and 2097 images for validation. The training parameter was set to 1000 epochs; however, the best model was derived at epoch 243, the number of anchors was eight, and stochastic gradient descent was used as the optimizer.

3 Experimental results

Fig 4 presents the results obtained for the validation set during training. The F1 score was 0.872, with the highest value at a confidence score of 0.257. For the precision, 0.977 AP@0.5 was observed in the bubble class, and 0.953 AP@0.5 was observed in the airBubble class. The mAP@0.5 was measured to be 0.965.

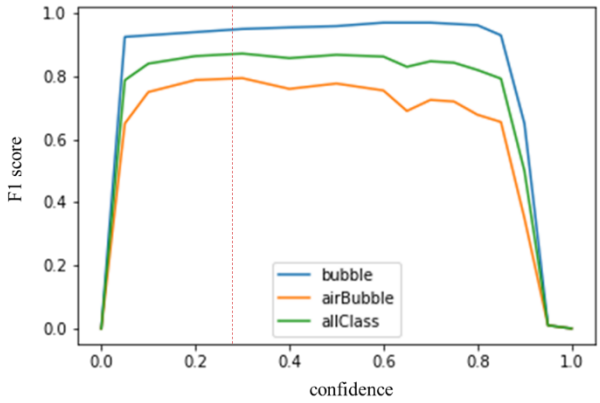


Fig. 4. F1 score with the confidence score for the validation set

The actual data were verified based on the previously obtained confidence score. A test was conducted on 300 lens images, and "Defect" and "No defect" products were classified. At this time, Actual was based on the judgment of the process operator.

Table. 1 Confusion matrix in the test data

Predict \ Actual	Defect	No defect
Defect	179	12
No defect	0	109

The test results are listed in Table. 1; an accuracy of 96% was obtained according to the following formula. The total test time for 300 lenses was 245 s, with an inspection speed of 0.82 s/piece.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4 Conclusions and future research

We managed to achieve our initial goals; however, our analysis could be improved in several aspects. In particular, numerous cases with erroneous classifications of normal products as defective were noted.
Notably, this could be addressed through post-processing. However, considering the inspection speed, altering the model structure appears desirable.

Acknowledgments. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2022-2020-0-01462) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation) and by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2019-0-01183) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

1. Yoon, Y. H., Lee, J., Lee, E., Moon, B. M., Seo, J. H., Lee, J., ... & Sung, S. Policy suggestions on the smart factory based on the survey results from smart factory suppliers. Journal of the Korean Society for Quality Management, 48(1), 1–11. (2020)
2. Gino Sophia, S. G., Ceronmani Sharmila, V. Recognition, classification for normal, contact and cosmetic iris images using deep learning. International Journal of Recent Technology and Engineering, 8(3). (2019)
3. Raghavendra, Ramachandra, Kiran B., Raja, Christoph Busch. Contlensnet: Robust iris contact lens detection using deep convolutional neural networks. 2017 IEEE Winter Conference on Applications of Computer Vision. (2017)
4. Zin, N. A. M., Asmuni, H., Hamed, H. N. A., Othman, R. M., Kasim, S., Hassan, R., ... & Roslan, R. Contact lens classification by using segmented lens boundary features. Indonesian Journal of Electrical Engineering and Computer Science, 11(3), 1129–1135. (2018)
5. Zheng, J. C., Shi-Dan, S. and Shi-Jia, Z. Fast ship detection based on lightweight YOLOv5 network. IET Image Processing. 16(6), 1585–1593 (2022)
6. Glenn, J., Alex, S., Jirka, B., NanoCode012, Ayush, C., TaoXie, Liu, C., Abhiram, V., Laughing, T., yxNONG, Adam, H., lorenzomammana, AlexWang1900, Jan, H., Laurentiu, D., Marc, Yonghye, K., oleg, wanghaoyang0106, Yann, D., Aditya, L., ml5ah, Ben, M., Benjamin, F., Daniel, K., Ding, Y., Doug, D., Francisco, I. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, Apr (2021)

Robust lane detection using saturation and lightness

Sumin Kim[1] and Youngbae Hwang[1]

[1] Dept. of Electronics Engineering, Chungbuk National University, Cheongju-si 28644, Rep. of Korea, ksm4465ksm@chungbuk.ac.kr, ybhwang@cbnu.ac.kr

Abstract. Extracting effective binary images is important for lane detection. To generate binary images, only using the specific channel of color space may be problematic due to various factors in the driving environment. In this paper, we present a method to extract binary images for robust lane detection by fusing a saturation channel of HLS color space and a lightness channel of CIELab color space. After applying pre- and post-processings to the binary images, lane is detected accuracely by line fitting.

Keywords: Lane Detection, Saturation, Lightness, Line Fitting

1 Introduction

One of important procedures for lane detection is extraction of binary images in various color spaces [1][2]. The complex methods of post-processing are performed to deal with this step. If regions of lane are not properly obtained, it is difficult to fit the lane correctly. To handle this problem, binary images for lane regions were extracted using the complementary fusion of a saturation channel in HLS color space and a lightness channel in CIELab color space. After pre-processing and post-processing are applied to binary images, we show that lane is well fitted to test images.

2 Method

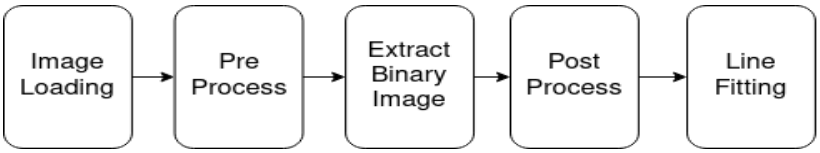


Fig. 1. Procedure for lane detection

The sequence of algorithms for detecting lanes is shown in Fig.1. After resizing an image to reduce computation, only the lane region was left after removing the sky region. Next, the saturation channel in HLS color space and the lightness channel were fused to generate binary images for lane detection. Lightness has characteristic that can detect bright lane regions. However, if the illunimation is strong, the reflection causes erroneous cases. In the case of saturation, the lane color can be detected clearly in spite

of blurness. Additionally, it does not generate artifacts from the illumination. Based on the assumption that both of two channels can detect lanes well, after applying XOR operation to binary images of two channels, the result of XOR is subtracted from each channel value. Then, the final binary image is obtained through AND operation of these lightness and saturation channels. The lanes are detected through a line fitting algorithm with RANSAC as dividing left and right sides of ego-lane.

3 Experiment

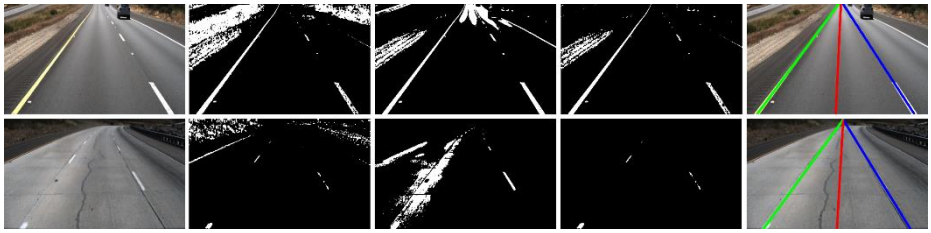


Fig. 2 From the first to the fifth columns are original images, binary images from the satuation, binary images from the lightness, fusion results of binary images of the saturation and the lightness channels, and lane detection results, respectively

In the case of the third column in Fig. 2, lane regions are well found, but false regions occur due to light reflection on the road. In the case of the second column, although lane regions are blurred, there is no problem from the reflection. Therefore, by combining two channels, binary images can be obtained as in the fourth column. Then, other regions outside the lane are removed using vertical line constraint of Bird Eye View (BEV). By applying line fitting with RANSAC, lanes are detected accurately.

Acknowledgement

This work was partially supported by the Grand Information Technology Research support program (IITP-2022-2020-0-01462), by Institute of Information & communications Technology Planning & Evaluation (IITP) (No.2022-0-00970).

References

1. Kim, Kwang Baek, & Song, Doo Heon.: Real Time Road Lane Detection with RANSAC and HSV Color Transformation. *Journal of Information and Communication Convergence Engineering*, *15*(3), 187–192. (2017)
2. C. Ma and M. Xie.: A Method for Lane Detection Based on Color Clustering, *2010 Third International Conference on Knowledge Discovery and Data Mining*, pp. 200-203, doi: 10.1109/WKDD.2010.118. (2010)

Weight Prediction of Korean Cattle with Weather Information Using Time Series Data Analysis Method

Sora Kang[1], Wanhyun Cho[2], Myung-Hwan Na[2],

[1] Department of Mathematics and Statistics, Chonnam National University, 61186, Republic of Korea
[2] Department of Statistics, Chonnam National University, 61186, Republic of Korea
sc12love@gmail.com
{whcho, nmh}@jnu.ac.kr

Abstract. Korean cattle is mainly bred for food, and growth may be inhibited due to stress in the process of measuring the weight of livestock or in a rapidly changing environment. Therefore, it is necessary to automatically predict the weight using the breeding environment and characteristics of Korean cattle. In this study, we try to predict the weight during the breeding period by considering the influence of age and environment of Korean cattle through deep learning techniques such as LSTM and Attention-LSTM models. As a result, the attention-LSTM technique showed excellent performance by showing the lowest RMSE, MAE, and MAPE. Finally, growth and environmental factors that affect the weight of Korean cattle are reviewed, and a breeding strategy that can control the quality and production of Korean beef through the FDA technique is proposed.

Keywords: Weight of Korean cattle, LSTM, Attention, Functional data analysis, Time series data analysis.

1 Introduction

Korean Cattle is one of the most popular meats, and is mainly raised for food rather than agriculture. In particular, there is a lot of interest in the breeding period and weight of Korean cattle, as can be seen from the article titled 'Super Hanwoo' whenever the maximum weight is updated. Although the breeding scale and form of Korean cattle are diversifying, the occurrence of high temperature stress due to extreme weather such as heat waves is directly linked to the death of many cattle.

Therefore, research that can measure automatically is needed to reduce the stress of livestock, and it is time to estimate the weight of Korean cattle in consideration of the weather environment. Various studies are being conducted in this regard. Kang et al. (2021) studied a method for predicting the weight of Korean cattle using image processing technology, and Kang et al. (2021) attempted to build a model using a Box-Cox transformation linear regression model.

This study examines the relationship between the weight of Korean cattle, age, and weather information, and attempts to predict their weights through long short term memory (LSTM) and Attention-LSTM models, which are known to show excellent performance in time series data. As a result, we intend to present a breeding strategy for each breeding stage according to the breeding management method through the functional data analysis (FDA) technique so that we can use this to prepare for the weather environment.

2 Material and Analysis Method

2.1 Analysis Data

The data used in this study are weather environment data and growth data collected by week from two farms raising Korean cattle.

The environmental data is the synoptic meteorological observation (ASOS) data provided by the Meteorological Data Open Portal (https://data.kma.go.kr/cmmn/main.do) of the Korea Meteorological Administration, and it is measured in units of one hour. New variables such as mean, max, min, and cum were created based on each observation value. Regions without environmental data were replaced with environmental data with the same latitude.

Growth data are weekly data from August 2021 to June 2022 (26 weeks to 52 weeks), and the survey period is slightly different for each individual Korean cattle. The target farms are raising Korean cattle for breeding in Jangheung-gun, Jeollanam-do and Gokseong-gun, Jeollanam-do.

The response variable is the weight of Korean cattle, and the explanatory variables are biological characteristics such as month (meaning how many months old you were born) and environmental factors such as temperature (℃), precipitation (mm), wind speed (m/s), humidity (%), sunlight (hr). Growth was considered to be affected by the environment the day before, and environmental data were made weekly according to the date of growth.

2.2 Data preprocessing

Prior to analysis, pre-processing of the data was performed to increase the data level.

First, when there was no survey period for growth data, the missing weekly data was interpolated using linear interpolation, and the missing environmental data was replaced with the weekly average.

Second, out of a total of 28 Korean cattle, 22 were used as training data (the number of data is 764), and 6 were used as test data (the number of data is 189). The test data was divided in consideration of the farm household.

Next, the explanatory variables used in the analysis model were scaled to equalize the distribution and range of the data. The model having only month as explanatory variables was transformed using the Min-Max scaling technique, and the model having month and environment as explanatory variables was transformed using the standard scaling technique. The min-max scaling technique can reduce the range by reducing the value to a value between 0 and 1, and the standardized scaling technique makes the mean 0 and standard deviation 1, and is suitable considering that the units of the environment are different.

Finally, since it is a time series data, the data structure was modified according to the analysis model. By converting 4 weeks' characteristic values into data, the short-term model predicts the next week, and the long-term model predicts one month later.

2.3 Analysis Method

LSTM is a type of recurrent neural network and is a model with long and short-term memory. There are input gate, forget gate, and output gate, and the sigmoid function in these three gates controls the gate. In particular, as the length of each cell increases, the information storage capacity decreases, leaving close information and discarding unnecessary information through the forget gate. However, there is a limit to putting all information in a fixed vector.

Attention-LSTM appeared to compensate for this limitation, and it explores the entire data at every time point and reflects it with a higher weight on the more relevant parts rather than the same weight. The attention distribution can be obtained through the soft-max function by obtaining the attention score from the encoder part, and the sum of the attention weights is 1. Next, the attention value may be obtained by weighting the attention weight and the hidden state of each encoder. The result obtained by delivering the attention mechanism to the LSTM of the decoder part increases the connectivity between data.

FDA is a statistical analysis method that provides all information that changes with time in time series data. Through a regression model using an estimation method that minimizes Integrated Square Error (ISE), growth and environmental factors affecting the weight of Korean beef can be identified, and the level of factors that can maximize the weight of Korean cattle can be found. By visually showing changes over time, it is possible to intuitively understand the appropriate breeding environment level during the breeding period.

$$ISE = 1/n \ \sum(g`(x)-g(x))2 \ . \quad \textbf{(1)}$$

3 Results

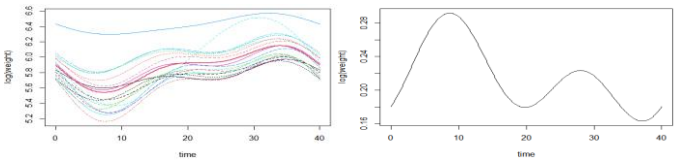
3.1 Correlation Analysis

In this study, correlation analysis was performed to find out the relationship between the weight of Korean cattle, month after birth, and environments.

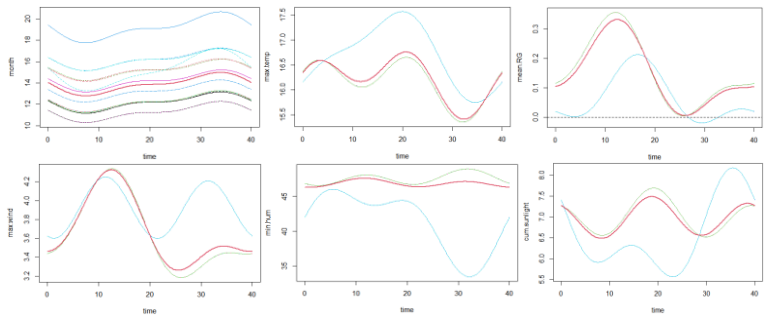
As a result of examining the relationship between weight, month after birth and environmental factors, the correlation was high in the order of month, average precipitation, minimum humidity, cumulative sunlight, maximum temperature, maximum wind speed. Month, maximum temperature, maximum wind speed and cumulative sunlight have a positive correlation, while mean precipitation and minimum humidity have a negative correlation.

3.2 FDA Analysis

Fig. 1. is a line graph explaining the trend of measured values (color lines) and averages (red line) for each individual for the log value of Korean cattle weight, month, and five environmental factors for 40 weeks reported by functional data. First, although there was no significant difference in the weight of Korean cattle at most time points, it can be seen that there is a difference in the weight level of each individual in the 10th stage (end of November). Second, the month was different for each individual Korean cattle, and the two farms showed different breeding environments during the breeding stage.



(a) Trend and variance of weight over time by week



(b) Trend of month and environment factors over time by week

Fig. 1. Trend of weight of Korean cattle and explanatory variables over time by week

Fig. 2. shows the beta value estimated by the multivariate functional regression analysis. Due to collinearity with other explanatory variables, it was possible to determine when the log value of the weight of Korean cattle had a significant effect using month, maximum temperature and minimum humidity.

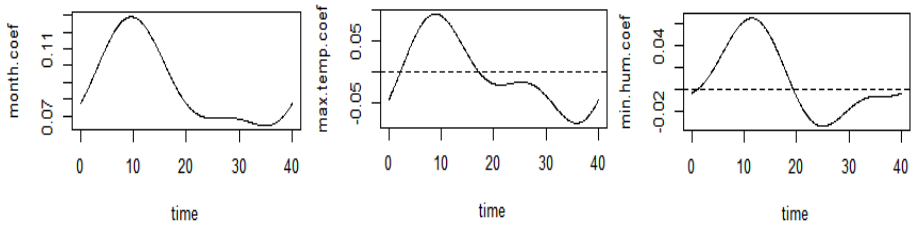


Fig. 2. Estimated parameter values for month, max.temp and min.hum

Through **Fig. 1.** and **Fig. 2**, we suggest month after birth and breeding environment that can maximize the weight of Korean cattle. In order to increase the weight of Korean cattle, it is important to manage them that are 13 to 14 months until the 10th stage. In the 10th stage of the breeding period, maintaining the maximum temperature at 16.0-16.5℃ and the minimum humidity at 45-50% has a positive effect on weight, and to increase body weight in the later stage, the minimum humidity should be maintained at 45-50%.

3.3 Evaluation of Models

To compare the performance of each model, root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were used as evaluation indicators, and the results are shown in **Table 1.** The best model was the Attention-LSTM model considering both month and environments, and it was found to be RMSE=5.030, MAE=4.067, and MAPE=1.087. It was found that the model considering the breeding environment together showed lower values than the model considering only the month, showing better predictive power. In particular, it was confirmed that the Attention-LSTM model significantly reduced the error than the LSTM model.

Table 1. Performance evaluation results on short-term prediction.

Model	Explanatory variable	RMSE	MAE	MAPE
LSTM	Month	13.097	10.125	2.937
	Month + Env	8.460	6.194	1.874
Attention-LSTM	Month	11.298	9.593	2.536
	Month + Env	5.030	4.067	1.087

Fig. 3. shows the results of predicting the weight of Korean cattle with the Attention-LSTM model considering all variables, which is the best learning model for the test data. It can be seen that the predicted weight of Korean cattle well predicted the actual weight of Korean cattle during the breeding period.

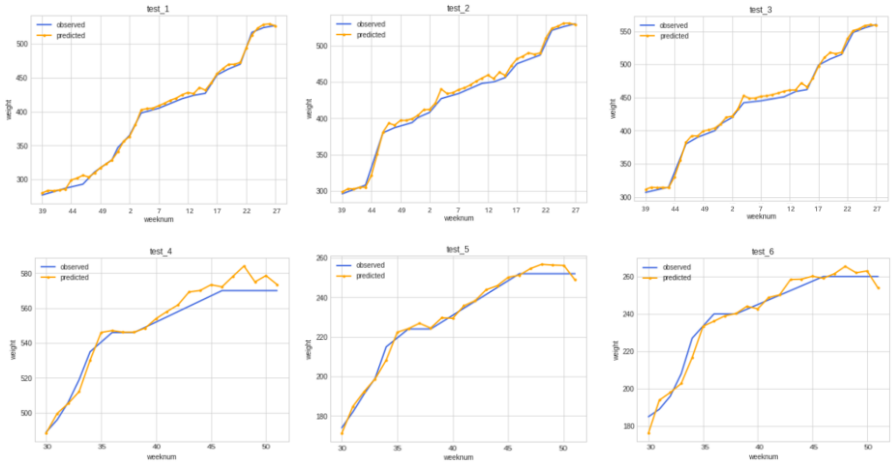


Fig. 3. Predicted result using the best Attention-LSTM model.

The predictive performance of Attention-LSTM model with the same conditions, called considering both month and environment, was evaluated. **Table 2.** shows the performance results of the learning model on the test data according to the prediction period, such as after 1, 2, 3, and 4 weeks.

The average 1-month prediction performance was RMSE=3.549, MAE=3.403, and MAPE=1.260, but the performance index values for the results of predicting the weight of Korean cattle after 1 week, 2 weeks, 3 weeks, and 4 weeks is increasing. It means that the error increases and the long-term prediction performance deteriorates.

Table 2. Performance evaluation results on long-term prediction.

Model	Forecast period	RMSE	MAE	MAPE
Attention-LSTM	After 1 week	8.504	6.712	1.819
	After 2 weeks	9.427	6.817	1.851
	After 3 weeks	11.496	8.239	2.194
	After 4 weeks	11.678	8.132	2.223
	total 1 month	3.549	3.403	1.260

As a result, it is possible to estimate the weight of Korean cattle in the short term, but it is insufficient in the long term. Nevertheless, it is meaningful to suggest an efficient breeding strategy that makes it possible to give low-weight Korean cattle a

large amount of feed and proper weight Korean cattle a small amount of feed for a short period of time.

4 Conclusions

In this study, growth and environmental factors affecting the weight of Korean cattle were investigated using weekly Korean cattle data from August 2021 to June 2022. In addition, a model for predicting the weight of Korean cattle was built, and the optimal age and weather conditions were proposed to maximize the weight of Korean cattle during the breeding stage. A Korean cattle weight prediction model was constructed using the deep learning technique LSTM and Attention-LSTM model, and the breeding strategy was confirmed using the FDA technique, a statistical technique. Finally, the performance of the models was compared using the metrics of RMSE, MAE, and MAPE. This study is meaningful in helping to find ways to prepare for bad breeding conditions, although climate phenomena cannot be controlled. In the future, if various growth information other than uncontrollable weather conditions can be collected and utilized, it is expected to contribute to increasing the weight of Korean cattle.

Acknowledgments. This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry(IPET) through Smart Plant Farming Industry Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(421017-04). This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF)

References

1. Kang Y, Na M, Cho W, Kim S.: Prediction of Korean cattle weight using image processing technology. In Proceeding of KIIT Conference 2021. 72-74. (2021)
2. Kang S, Na M, Cho W, Kim S, Go H, Cho K.: Prediction of weight of Korean cattle and pigs using statistical data analysis method. Proceedings of the Korean Society for Data Analysis Conference, 2020: 3-6. (2021)
3. Wu, P., Huang, Z., Pian, Y., Xu, L., Li, J., Chen, K.: A combined deep learning method with attention-based LSTM model for short-term traffic speed forecasting. Journal of Advanced Transportation, 2020. (2020)
4. Wang J, Chiou J, Muller H.: Functional data analysis. Annual Review of Statistics and its application. 3: 257-295. (2016)

Detection Model for Wearing Hardhat in workplace Based on Deep Learning

Ju-yeon Lee[1], Woo-seok Choi[1], Joong-hun Cho[1],
Sang-hyun Choi[2*]

[1]Dept. Bigdata, Chungbuk National University, Cheongju, South Korea
[2]Dept. Management Information System, Chungbuk National University, Cheongju, South Korea
{yeony_yy, jojh87,chois}@cbnu.ac.kr, {cdt3017}@naver.com

Abstract. Most of the safety accidents that occur in the workplace are caused by not wearing protective equipment such as hardhats. Therefore, the construction industry is making efforts by emphasizing the safety rules of workers or monitoring the wearing of protective equipment. However, deploying a safety manager for this purpose is expensive, and it is practically difficult to endlessly check whether workers are wearing protective equipment in the field. Therefore, this study proposed a plan that can be automatically monitored by CCTV in the workplace by designing a deep learning-based hardhat identification model.

Keywords: Construction Safety, HardHat, Faster RCNN, Deep Learning, Computer Vision

1 Introduction

In the workplace, large and small safety accidents occur frequently and the number of safety accidents increases every year[1]. In addition, the incidence of safety accidents caused by not wearing protective equipment such as hardhat was 88.9%, which was about 8 times higher than the incidence of safety accidents when wearing personal protective equipment, 11.1%[2]. In summary, it is currently possible to prevent it just by using a hardhat in many workplaces, but safety accidents continue to occur due to workers' insensitivity to safety. However, it is practically impossible for safety managers to continue monitoring workers for wearing hardhats in the actual field. To solve this problem, this researcher proposes a method that automatically identifies whether a worker wears a hardhat and gives an alarm by applying a deep learning algorithm to a camera installed in the workplace.

* Corresponding author

2 Analysis

Faster R-CNN[4] introduced Region Proposal Network (RPN), a network that performs candidate region extraction to solve the problem that the Selective search algorithm used for candidate region extraction in Fast R-CNN[3] operates on the CPU and causes bottlenecks. The RPN is located between the feature map and RoI Pooling in the structure of the Fast R-CNN and uses Anchor boxes of various sizes and aspect ratios to generate region proposals. Applying an anchor box created using sliding window techniques to Feature maps creates many region proposals that present object locations of various sizes and performs regression and classification for each location.

In this study, a total of 7,763 data were used to design a model for detecting whether to wear a hardhat in the workplace with 7,057 train data and 706 test data. 7,057 images have been split into 80% for training, 20% for validation.

The performance evaluation of the model for detecting whether to wear a hardhat in the workplace was evaluated using AP@loU = 0.5 and precision, recall, and f1-score. The performance evaluation results are shown in Table 1, and the AP of each class showed 73% of Head and 79% of Helmet.

Table 1. Evaluation of detection by using Faster RCNN models

Model	mAP	Precision	Recall	F1-Score
Faster $RCNN_{resnet50fpn}$	76.4 %	83.1%	78.3%	80.5 %

3 Conclusion

In this paper, the Faster RCNN was used to automatically check whether workers in the workplace wore hardhat. As a result, the mAP was finally recorded at 76.42%. Using the model proposed in this study, it is expected that safety managers can replace the work of monitoring whether workers wear hardhats and contribute to the prevention of safety accidents in the workplace.

Reference

1. You, H. J., You, Y. T. and Kang, K. S., “A study on the efficiency improvement of the safety management personnel system in apartment construction site.” Korea Safety Management & Science Korea Safety Management & Science, Vol. 19, No. 1, pp. 87-94 (2017)
2. Occupational Safety and Health Research Institute. “Cause of Industrial Accident in 2014”, OSHRI Research Report (2016)
3. Ross, Girshick. Fast R-CNN. arXiv preprint arXiv:1504.08083v2. (2015)
4. Shaoqing, Ren., Kaiming, He., Ross, Girshick., Jian, Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv preprint arXiv:1506.01497v3. (2016)

Comparison of Machine Learning Algorithms for Predicting Rice Yield Using Multispectral Images

Dahyun Kim[1], Wanhyun Cho[2], Sangkyoon Kim[2], Myung Hwan Na[2],

[1] Department of Mathematics and Statistics, Chonnam National University, 61186, Republic of Korea
[2] Department of Statistics, Chonnam National University, 61186, Republic of Korea
{qhqk7132, whcho, narciss76}@jnu.ac.kr, nmh@chonnam.ac.kr

Abstract. To predict the yield (kg/10a) of rice, vegetation indices were calculated from multispectral images of unmanned aircraft taken several times. Although many models have been developed to predict the yield of rice, the actual prediction performance is not excellent due to weather changes and pests. In this study, the yield of rice was predicted only by the vegetation index without using weather factors and growth survey data. Three machine learning models were trained with each vegetation index, and errors of the model were evaluated. These models included support vector machines (SVM), k-nearest neighbors (k-NN), and random forest (RF). The best performance model was the k-NN model of the NDRE vegetation index.

Keywords: vegetation index, multispectral image, support vector machines, k-nearest neighbors, random forest.

1 Introduction

Recently, smart farm technology is developing to collect environmental and growth information to analyze, model, and automate data. The most important focus in analysis and modeling is the prediction of crop growth and yield. The total yield of rice can be estimated from the yield per unit area (kg/10a) and the cultivated area. Estimation of yield per unit area (kg/10a) is affected by weather factors and the growth characteristics of rice.

In this study, without using weather factors, the yield was predicted through the vegetation index that can represent the growth characteristics. Various vegetation indices were calculated through time series multispectral images taken with an unmanned aerial vehicle. The vegetation index is made based on the high reflectance of Near-Infrared (NIR) in healthy vegetation. The vegetation index is possible to determine growth abnormalities. Among vegetation index, Normalized Difference Vegetation Index (NDVI) is the most used. Normalized Difference Red Edge (NDRE) is more suitable for observing the mid to late period of crop growth. Green

Normalized Difference Vegetation Index (GNDVI) is useful for measuring photosynthetic rate and monitoring plant stress. Finally, Leaf Chlorophyll Index (LCI) is an index that evaluates the chlorophyll content of areas that are completely covered with leaves.

Three machine learning algorithms support vector machines (SVM), k-nearest neighbors (k-NN), and random forest (RF) were used to predict the yield (kg/10a) of rice through the change of the vegetation index over time. These algorithms are non-linear, and the determination of hyperparameters affects the accuracy of the model. Therefore, grid searches were used to find the appropriate hyperparameters. The accuracy of the model was evaluated by calculating the root mean square error (RMSE), mean absolute percentage error (MAPE), and mean percentage error (MPE) as error indicators of the model.

Kang et al. (2021) calculated various vegetation index using multispectral images of rice. A model for predicting rice yield and protein content was developed using partial least squares method, ridge regression method, and artificial neural network (ANN) model.

2 Material and Method

2.1 Data Collection

This study used datasets from 'major crop growth images'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)' These are multispectral images and yield (kg/10a) data of rice in Jinju, Gyeongsangnam-do from 2018 to 2020. The sowing date of rice is May 4, the planting date is June 6, and the harvest date is October 17-18. There are 5 bands of the multispectral image: BLUE, GREEN, RED, RED EDGE, and NIR. The multispectral imaging period is about two months from early August to early October, the growth period of rice. The number of shots is 8 to 10.

2.2 Calculation of vegetation index

The vegetation index was calculated by combining four bands except for BLUE in multispectral images taken with an unmanned aerial vehicle.

NDVI is the most common and is used to measure the density of vegetation. NDVI value is higher when the vegetation is healthy and the reflectance of infrared light is higher. It is a vegetation index that is often used for drought monitoring or agricultural production forecasting. It has a value from -1 to 1.

$$NDVI = (NIR - RED) / (NIR + RED) . \quad (1)$$

NDRE is an index sensitive to leaf chlorophyll content on soil background effects. NDRE is used instead of NDVI when NDVI is almost 1 due to increased chlorophyll content in the late stages of cultivation of crops such as rice. It highlights problematic plants, from nutrient deficiencies to pest and disease damage.

$$NDRE = (NIR – RED\ EDGE) / (NIR + RED\ EDGE) . \quad (2)$$

GNDVI uses the green band instead of the RED band in the NDVI formula. It is often used to evaluate the moisture content and nitrogen concentration of plant leaves. It is more sensitive to chlorophyll concentration compared to the NDVI index. It is useful for measuring photosynthetic rate and monitoring plant stress.

$$GNDVI = (NIR - GREEN) / (NIR + GREEN) . \quad (3)$$

LCI is an index that evaluates the chlorophyll content of a completely covered area. It is formulated using three bands: RED, RED EDGE, and NIR.

$$LCI = (NIR – RED\ EDGE) / (NIR + RED) . \quad (4)$$

2.3 Proposed Method

Since the shooting date and interval are different for each year, the vegetation index value was obtained on a daily basis using the interpolation method. The spline interpolation method is a method of finding a smooth function with a low order polynomial. In the section where the y value changes rapidly, it is important to select the interpolation point well. A machine learning model was developed to predict the number of yields of rice through the change of the single unit vegetation index obtained using the interpolation method.

SVM takes a relatively long training time, but it is a popular model because of its good performance. To separate data based on the boundary line, learn to maximize the margin, which is the distance between the boundary line and the closest data. In a regression problem where the dependent variable is continuous, try to include as many observations as possible within the margin. A kernel is used to process non-linear data, and there are representative polynomial kernels and Gaussian time kernels.

The k-NN is an algorithm that predicts values through the nearest K observations. Therefore, the prediction results obtained from k-NN tend to be similar to those of neighbors. Choosing an appropriate k is a very important factor in determining performance. If k is too small, an overfitting problem that considers even the noise component of the data occurs. Conversely, if k is too large, the k-NN's ability to predict local information in the data is lost.

RF repeats the process of creating one decision tree by randomly selecting some features among all features. Outputs one prediction value per decision tree. Among

the prediction values made by several decision trees, the value that comes out the most is determined as the final prediction value.

The quality of a regression model is how well its predictions match up against actual values. RMSE, MAPE, and MPE were calculated to evaluate the error of the regression model.

3 Results and Discussion

3.1 Spline Interpolation

The shooting dates and intervals of multispectral images in 2018, 2019, and 2020 are different. In order to input to one model as an input, the spacing must be set equally. Using spline interpolation, the value of the vegetation index of unobserved points between observation points was calculated in units of days. As a result, the dates with data in common for 2018, 2019, and 2020 are about two months from August 9 to October 5. Fig. 1 shows the result of spline interpolation. A model was created to predict the yield of rice using changes in NDVI, NDRE, GNDVI, and LCI over time as variables.

3.2 Evaluation of Prediction Models

In regression analysis, the difference between the actual value and the model estimate is called residual. Residuals for all points in the dataset can be calculated and play an important role in determining the usefulness of the model. There are several regression error metrics that can be calculated through residuals. In this study, the regression model was evaluated using root mean square error (RMSE), mean absolute percentage error (MAPE), and mean percentage error (MPE).

RMSE can facilitate interpretation by converting error indicators into units like real values. Overall, the RMSE of SVM was higher than that of other models, and the RMSE of k-NN model was lower. The model with the lowest RMSE value is the k-NN model using the NDVI vegetation index, and the RMSE is 23.923 kg/10a. MAPE can interpret the prediction of the model as a percentage, how far away it is from the actual value. Like RMSE, MAPE also showed high errors in the order of SVM, RF, and k-NN. The model with the lowest MAPE value is the k-NN model using the NDRE vegetation index, and the MAPE is 3.373%. MPE is the elimination of the absolute value operation in MAPE. If there are many negative or positive errors, whether there is bias in the regression model can be evaluated. As a result of calculating the MPE, the RF model was the most unbiased. The MPE of the NDVI and NDRE vegetation index of the k-NN model, which performed well in the RMSE and MAPE error indicators, was compared. The model bias of NDVI vegetation index was more than double that of NDRE. Therefore, the best performing model was the k-NN model of the NDRE vegetation index.

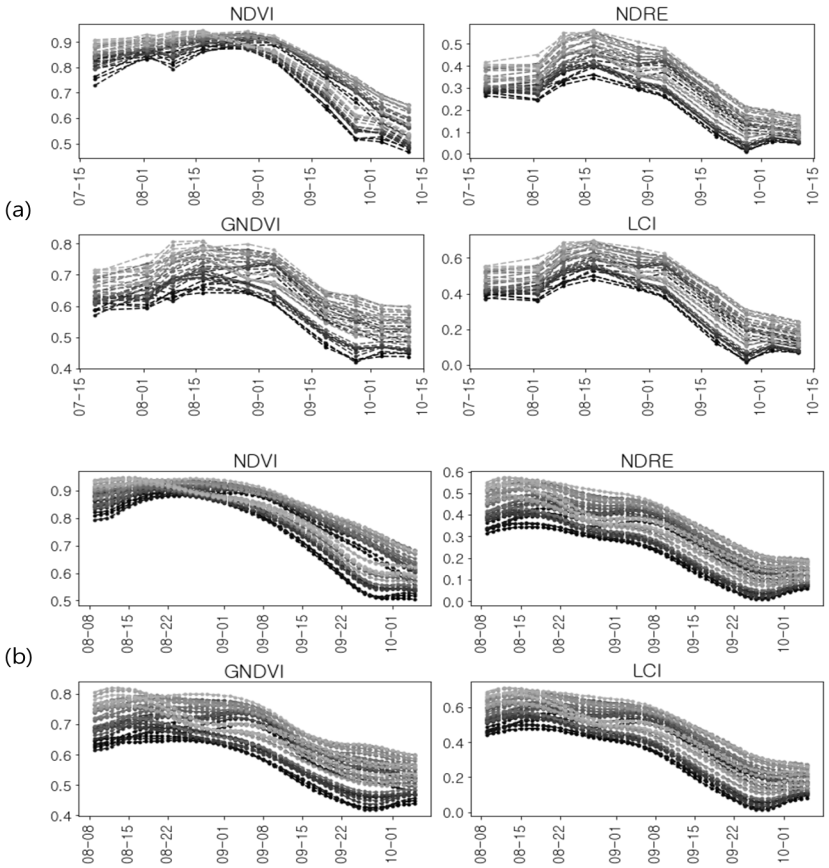


Fig. 1. Changes in vegetation index over time (a) before and (b) after of spline interpolation.

Table 1. Error rate of machine learning model.

Model	Vegetation Index	RMSE	MAPE (%)	MPE (%)
SVM	NDVI	28.146	4.707	-0.647
	NDRE	26.622	4.038	1.006
	GNDVI	26.432	4.104	0.305
	LCI	26.091	4.112	1.095
k-NN	NDVI	23.923	3.394	1.451
	NDRE	23.995	3.373	0.688
	GNDVI	25.241	3.793	0.618
	LCI	24.653	3.538	0.835
RF	NDVI	25.727	4.144	0.028
	NDRE	24.874	3.864	0.698
	GNDVI	24.549	4.106	-0.017
	LCI	24.199	3.822	0.769

4 Conclusions

In this study, the yield was predicted using time-series multi-spectral images taken with an unmanned aerial vehicle during the mid to late stage of the rice growth process. The vegetation index was calculated by combining the bands of the multispectral image, and the change of the vegetation index according to the passage of time was investigated. With these changes in the vegetation index, a model was created using three machine learning algorithms, and the performance of the model was evaluated as an error index. The performance of the k-NN model was the best, and among them, the errors of NDVI and NDRE were low. In the MPE index, which evaluates the model's bias, the NDRE has less bias than the NDVI, so it can be said to be the best performing model. Therefore, when the k-NN model using the NDRE vegetation index is used to predict the yield of rice, it will be most accurate.

Acknowledgments. This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through the Open Field Smart Agriculture Technology Short-term Advancement Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(32204003)

References

1. Kang Y., Nam J., Kim Y., Lee S., Seong D., Jang S., Ryu C.: Assessment of Regression Models for Predicting Rice Yield and Protein Content Using Unmanned Aerial Vehicle-Based Multispectral Imagery. Remote Sensing. 13, 8, 1508. (2021)
2. Yoon J., Yoon Y., Kim Y.: Utilization of Vegetation Indice in Agricultural Field. Journal of Agriculture & Life Science. 55, 5, 1-9. (2021)
3. Gregory A. Carter, William G. Cibula, Richard L. Miller.: Narrow-band Reflectance Imagery Compared with Thermal Imagery for Early Detection of Plant Stress. Journal of Plant Physiology. 148, 5, 515-522. (1996)
4. Lei Y., Mahdi A., Mujahid A., Amin J., Belgacem B., Muhammad F. Javed., Nermin M. Salem.: Comparative Analysis of the Optimized KNN, SVM, and Ensemble DT Models Using Bayesian Optimization for Predicting Pedestrian Fatalities: An Advance towards Realizing the Sustainable Safety of Pedestrians. Sustainability. 14, 10467. (2022)
5. X. Zhou, H. B. Zheng, X. Q. Xu, J. Y. He, X. K. Ge, X. Yao, T. Cheng, Y. Zhu, W. X. Cao, Y. C. Tian: Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 130, 246-255. (2017)

Deep learning-based model for rapid prediction of in-hospital clinical deterioration

Trong-Nghia Nguyen[1], Ngoc-Tu Vu[1], Bo-Gun Kho[2], Guee-Sang Lee[1], Hyung-Jeong Yang[1], Soo-Hyung Kim[1,*], Aera Kim[1]

[1] Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea

[2] Pulmonology and Critical Care Medicine, Chonnam National University Hospital, Gwangju, Korea

{trongnghia7171, tuvungocnd, imdrkbg}@gmail.com,
{gslee, hjyang, shkim, arkim}@jnu.ac.kr

Abstract. In this study, we develop a deep learning application system with high interpretability and diversity in input features for the prediction of in-hospital clinical deterioration. The high ability to understand input features helps the system to stick to the actual context. The use of the Transformer structure has made our method superior to comparative models in many respects when testing on a large and challenging data set with a 0.652 F1-score, 0.77 sensitivity, 0.837 AUROC, and 0.839 AUPRC.

Keywords: Clinical Deterioration, Rapid Response System, Deep Learning, Machine Learning.

1 Introduction

Deterioration in hospitals is a serious problem for medical systems. Approximately 209,000 patients are treated for cardiac arrest in hospitals each year [1]. This index has increased since the emergence of the coronavirus pandemic 2019 (COVID-19), which has become a burden on public health [2]. In this context, rapid response systems (RRS) are constantly being researched and developed to prevent treatment delays that are caused by an overload due to large numbers of hospitalized patients. Many studies have focused on developing a "risk score" - an indicator to assess the loss of danger to a patient's clinical condition. In this investigation, we develop a Rapid Response System of applying deep learning and transformed architect techniques to improve the predictive quality of the rapid response system through a probability algorithm.

* Corresponding author.

1.1 Related works

When it comes to "risk score", it must be mentioned classic methods like Modify Early Warning Score [3], and National Early Warning Score [4]. These assessment methods are mainly based on the diagnostic experience of specialists and physicians. Through the input of patient vitals (Heart Rate (HR), Respiration Rate (RR), Body Temperature (BT), etc.) at a time, MEWS and NEWS can estimate the condition of that patient (Normal/Abnormal) at that moment. The limitation of the above methods is that the range of input features is too narrow in the context of the development of electronic health records (EHRs). This makes it impossible for us to take full advantage of the variety of input features. Thus, machine learning/deep learning application approaches were born.

In recent years, a large number of studies aimed at increasing the likelihood of predicting the early sign of clinical deterioration have been published. Typical can be mentioned MEWS ++, a variant of MEWS aimed at improving the ability to predict clinical developments in hospitalized patients through machine learning models [5]. This system trains three classical machine learning methods (random forest (RF), linear support vector machine, and logistic regression) on a large dataset. By comparison, it surpassed traditional MEWS with an increase of 37% in sensitivity, 11% in specificity, and 14% in the area under a receiver operating characteristic (ROC) curve (AUROC). Besides, DeepSigns [6]: a method for creating a computational model that can predict the deterioration of a patient's health in such a way that appropriate treatment can be started as soon as possible, has been developed based on the integration of Deep learning techniques, Recurrent Neural Networks [7], and the Long Short-Term Memory [8], were also introduced. Furthermore, we have also published a study [9] with the application of TabNet [10] - Interpretable Learning for tabular data on the same topic. Experimental results on a private dataset show that the proposed method outperforms machine learning models with 66% AUROC and 29.1% of the area under a precision-recall curve (AUPRC). However, this study still has many limitations such as limitations on input features type, and low sensitivity leading to the system skipping time points that should be alerted.

1.2 Method and Contribution

From the above background and motivation, we aim to develop deep learning based RRS (Deep-RRS) to satisfy the following factors: (1) Diversity in input data types - Instead of using only numeric features (vital signs or laboratory tests), our method applies Transformer architecture [11] to optimize the interpretability and understanding of the model for textual features. (2) Increase the sensitivity of the system - Improve the ability to detect "abnormal time points".

2 Proposed Method

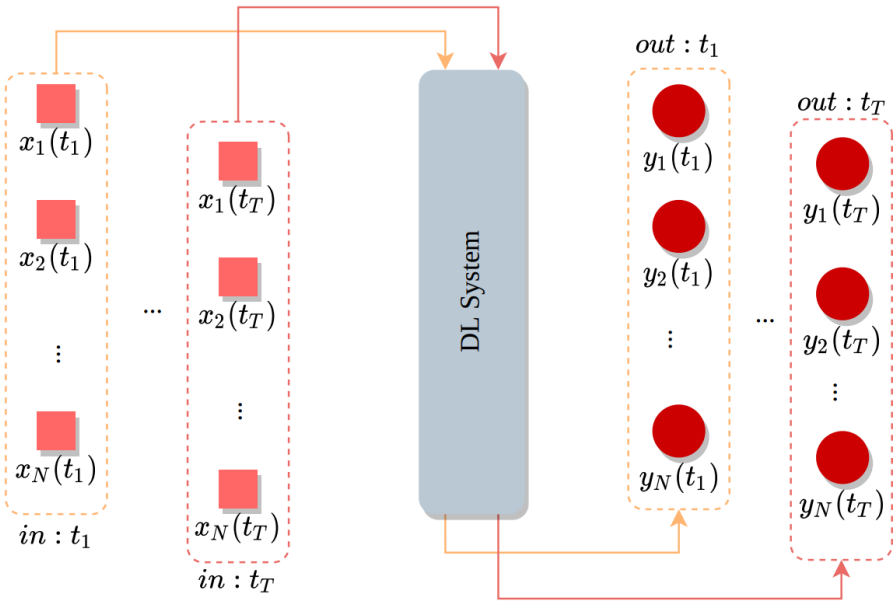


Figure 1: Overall pipeline of our proposed system. Input: patient's clinical variables (x) at timepoint t; Output: abnormal probability (y) corresponding to input measurement values x at time t.

The overall proposed system is illustrated in **Figure 1**. The main goal of the system is to assess whether the clinical condition of the patient at a given time is abnormal or normal through abnormal probability. Our system takes as input the patient's clinical indicators and general features at a certain time point. Here, the Machine Learning (ML)/Deep learning models (DL) trained on the large data set will process the input information and return the patient's abnormal probability at the corresponding time. By estimating a suitable threshold (0.5 in this problem), we can determine the final abnormal status from the above probability index. Thus, the "risk score" in our system is the abnormal probability.

2.1 Feature Selection

We mentioned the diversity of input data types as an advantage of this study over our previous method. However, more than simply using more features to increase the system's performance, this study performs physician-assisted feature selection so that these features really make sense for the medical side and are suitable for the emergency system.

From 37 features of the 2021 RRT CNUH dataset - the dataset used for this study, we conduct pre-processing, screening, and selecting the most 11 suitable and optimal

features for the system. Used features include patient clinical variables and general information, which are divided into 2 types:

Numerical features (n = 6): Age (general) and vital signs (clinical): Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Hear Rare (HR), Respiration Rate (RR), Body Temperature (BT).

Categorical features (n = 5): All are the patient's general information: Gender, Diagnosis, Hospitalization Department, Hospitalization Route, and Inpatient Ward.

2.2 Deep Learning Model Architecture

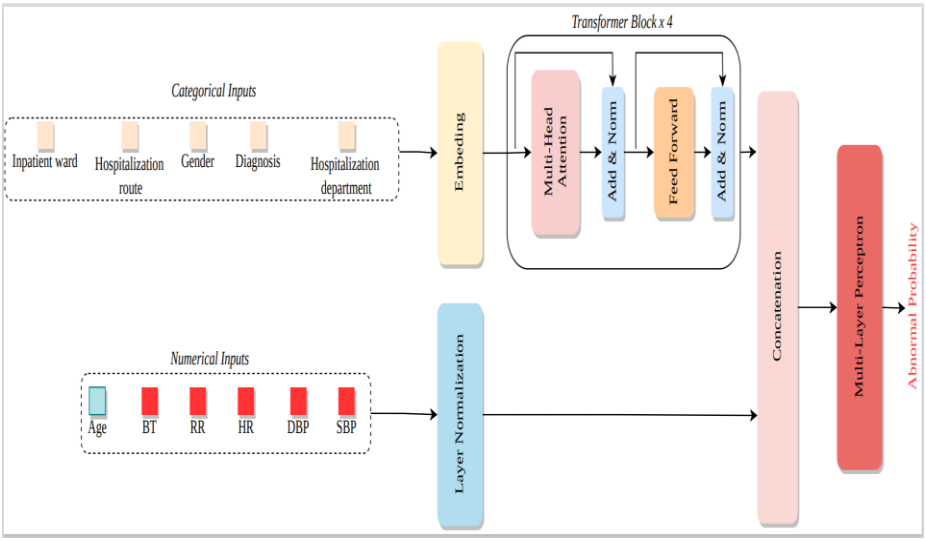


Figure 2: Proposed Transformer based model for in-hospital abnormal status prediction in Deep-RRS.

Our proposed DL model (**Figure 2**), is based on TabTransformer's structure [12], consisting of 3 main elements: an embedding layer, 4 Transformer layers, and a multi-layer perceptron. Which, each Transformer layer consists of a multi-head attention layer, followed by a positional transition layer. Among the input features, we have Diagnosis - a field containing the doctor's assessment and comment about the patient's condition through a sentence. Therefore, we use Transformer constructs because of their powerful natural language processing capabilities.

Assume that we have a set of input features $x = \{x_{cat}, x_{num}\}$, where $x_{num} \in \mathbb{R}^n$ denote all n numerical features and $x_{cat} = \{x_1, x_2, \ldots, x_m\}$ illustrates all m categorical features. model's pipeline can be summarized into the following steps:

1. x_{num} is normalized by Layer Normalization while x_{cat} is embedded to get the embedded features $E_\varphi(x_{cat})$.
2. Embedded features $E_\varphi(x_{cat})$ are fed into 4 Transformer blocks for obtaining contextual embedding features. The sequence Transformer layers could be

formulated as S_ϕ which processes $E_\varphi(x_{cat})$ to get the contextual embedding $\mathbb{C} = \{c_1, \ldots, c_m\}$ corresponding for each x_i, $i \in \{1, \ldots, m\}$.

3. Concatenating contextual embedding features $\mathbb{C}$ and normalized numerical features together.
4. Concatenation is used as input of MLP (denoted as $\mathcal{M}_\vartheta$) to get the prediction probability y.

Let $\mathcal{K}$ be the cross-entropy for the whole classification task, our model minimizes the loss function $\mathcal{L}(x, y)$ for enhancing all model's parameters by the first-order gradient approaches. Which, the model defines φ as the column embedding, ϕ for the Transformer layer, and ϑ for the top of MLP.

$$\mathcal{L}(x, y) = \mathcal{K}(\mathcal{M}_\vartheta(S_\phi(E_\varphi(x_{cat})), x_{num}), y). \quad (1)$$

3 Experiment Results

3.1 Dataset & Experiment configuration

For the experimental process, we performed the study on the 2021 Rapid Response Team Chonnam National Hospital dataset (RRT-CNUH). This dataset was collected and screened by the Rapid Response Team of Hakdong Chonnam National University Hospital during the period from February 1, 2021, to November 30, 2021. **Table 1** aggregates demographics on training/test sets and some used characteristics of the dataset.

Table 1 Cohort demographics of the RRT-CNUH dataset.

		Total N (%)	***Training (%)***	***Test (%)***
Number patients		25,329	18,470 (80)	6,859 (20)
Normal (%)			15,775 (81.4)	5,873 (85.6)
Abnormal (%)			3,602 (18.6)	1,335 (14.4)
General Information	Age (mean±std)		64.2 ± 17.6	64.0 ± 18.1
	Gender (male/female)		9,752/8,718	3,605/3,254
Vital signs (mean±std)	SBP		110.5 ± 38.7	110.7 ± 38.9
	DBP		65.4 ± 27.3	65.5 ± 27.4
	HR		69.6 ± 27.3	69.4 ± 27.2
	RR		17.6 ± 6.5	17.6 ± 6.4
	SBT		34.6 ± 8.4	34.6 ± 8.0

We split this dataset by 80% for training and 20% for testing. The mean age of the patients in the data set fluctuates at 64 and the male/female ratio is balanced. However, the rate of clinical performance deterioration (abnormal) for the whole samples has a large difference (18.6% for train and 14.4% for test). Therefore, the models need to be sensitive enough to extract the minority samples from the entire data.

For comparison models, we divide them into 3 main groups:

Traditional Method: MEWS - This method will lose its advantage over other models because it only calculates based on 5 vital signs features.

Machine Learning Models: Extreme Gradient Boosting (XGBoost), Random Forest (RF), Multi-Layer Perceptron, Light Gradient Boosting - Effective machine learning models for tabular classification based on tree and boosting principles. These models make full use of input features and are hyper-parameterized by Sklearn's GridsearchCV library.

Deep Learning Models: TabNet [10], DANets [13] - Two novel models have been proven to outperform machine learning methods in recent years. TabNet is the method we applied in our previous research.

With the characteristic of the binary classification task and the unbalanced nature of the data, besides F1-score, we use AUROC and AUPRC as comparison metrics. Besides, to evaluate the sensitivity of the model, the Sensitivity score is also used.

3.2 Results

Table 2 Models performance metrics.

Method	*F1*	*Sensitivity*	*AUROC*	*AUPRC*
MEWS	0.471	0.01	0.526	0.465
XGBoost	0.599	0.65	0.705	0.693
RF	0.612	0.69	0.736	0.724
MLP	0.553	0.44	0.635	0.621
LGB	0.602	0.68	0.707	0.695
TabNet [10]	0.612	0.73	0.715	0.705
DANets [13]	0.532	0.37	0.597	0.573
Deep-RRS (proposed)	0.652	0.77	0.837	0.839

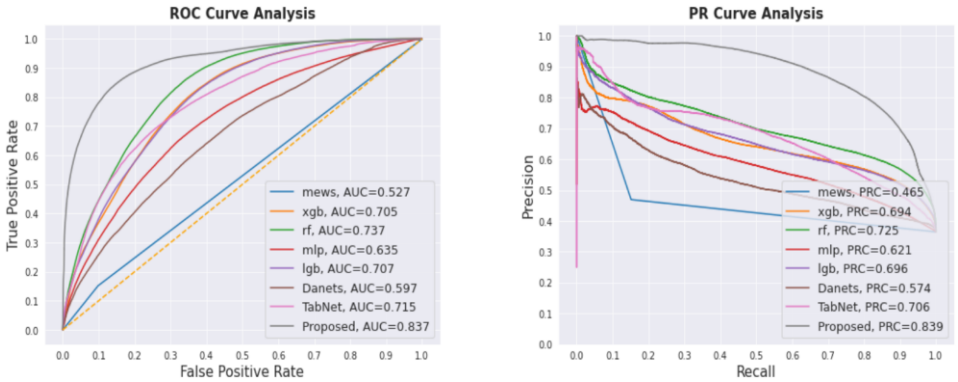


Figure 3 ROC & PRC curve analysis.

Table 2 shows the overall performance of comparison approaches and **Figure 3** illustrates the ROC curve and Precision-recall curve analysis, respectively. Our

proposed model outperformance comparison approaches in all evaluation metrics with 0.652 of F1-score, 0.77 of sensitivity, 0.837 of AUROC, and 0.839 of AUPRC. Deep-RRS showed significant improvement in the case of AUROC and AUPRC with an increase of up to 0.101 of AUROC and 0.115 of AUPRC (compared to RF (0.736 of AUROC and 0.724 of AUPRC) - best of the rest). The high sensitivity has also shown the model's ability to overcome the "missing alarm" problem.

From the above results, it could be seen that the ability to optimize the characteristics of the input data through extracting robust contextual embedding features has increased the efficiency of the prediction process. Exploiting the context of categorical features has enhanced the interpretability of the system. In particular, the use of the Transformed technique - a powerful structure in natural language processing has helped the model make the most of textual features such as Diagnosis instead of having to use limited features for MEWS or must encode them for machine learning models.

4 Conclusion

Deep-RRS has been shown to be superior in predicting clinical deterioration by time step. The advantage of being able to take advantage of a variety of input features helps the system to overcome comparison methods due to its high interpretability. Furthermore, the high sensitivity of the system has been verified through testing and evaluation. However, the limitation of this study is its output as the system can only predict each time step. Therefore, we will improve the versatility of the time series-based system in the future.

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-2020R1A4A1019191). This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF)& funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961). The corresponding author is Soo-Hyung Kim.

References

1. Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R. Mackey, R. H. (2017). Heart Disease and Stroke Statistics- 2017 Update: A Report from the American Heart Association. Circulation, 135(10), e146–e603.
2. R. M. Padilla and A. M. Mayo: Clinical deterioration: A concept analysis, Journal of Clinical Nursing, vol. 27, no. 7-8, pp. 1360–1368, 2018.
3. C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," QJM – Monthly Journal of the Association of Physicians, vol. 94, no. 10, pp. 521–526, 2001.
4. Royal College of Physicians, "National Early Warning Score (NEWS) 2 Standardizing the assessment of acute-illness severity in the NHS: Additional Implementation Guidance," Rcplondon. Ac. Uk, no. March 2020.

5. A. Kia, P. Timsina, H. N. Joshi, E. Klang, R. R. Gupta, R. M. Freeman, D. L. Reich, M. S. Tomlinson, J. T. Dudley, R. Kohli-Seth, M. Mazumdar, and M. A. Levin, “MEWS++: Enhancing the prediction of clinical deterioration in admitted patients through a machine learning model,” Journal of Clinical Medicine, vol. 9, no. 2, 2020.
6. D. B. da Silva, D. Schmidt, C. A. da Costa, R. da Rosa Righi, and B. Eskofier, “DeepSigns: A predictive model based on Deep Learning for the early detection of patient health deterioration,” Expert Systems with Applications, vol. 165, no. March 2020, p. 113905, 2021.
7. Rumelhart, David E; Hinton, Geoffrey E, and Williams, Ronald J (Sept. 1985). Learning internal representations by error propagation. Tech. rep. ICS 8504. San Diego, California: Institute for Cognitive Science, University of California.
8. Hochreiter, S., & Schmidhuber, J"urgen. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
9. T.-n. Nguyen, T.-h. Vo, B.-g. Kho, and G.-s. Lee, “Deep Interpretable Learning for a Rapid Response System,” pp. 8–10.
10. S. O. Arik and T. Pfister, “TabNet: Attentive Interpretable Tabular Learning,” 2019.
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.
12. X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, “TabTransformer: Tabular Data Modeling Using Contextual Embeddings,” no. Rong 2014, 2020.
13. J. Chen, K. Liao, Y. Wan, D. Z. Chen, and J. Wu, “DANets: Deep Abstract Networks for Tabular Data Classification and Regression,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 4, pp. 3930–3938, 2022.

Effective Weighting Scheme for Frequent Subgraphs Extracted from Knowledge Graph

Haemin Jung[1], Kwangyon Lee[2]

[1]Department of Industrial Engineering, Yonsei University,
Seodaemun-gu, Seoul 03722, South Korea
hmjung@yonsei.ac.kr
[2]Industrial AI Research Center, Chungbuk National University,
Cheongju 28116, South Korea
kylee22@cbnu.ac.kr

Abstract. Frequent subgraph mining is used as a method for finding patterns in the knowledge graph. A vector representation or embedding scheme is required to actually use the extracted subgraph as input to an application. Today's deep learning algorithms can provide a way to embed subgraphs during the training process with a specific task objective. The resulting dense vector of a subgraph might be a good representation in vector space, but humans are unable to understand or describe the specific triple information that the original subgraph possesses. In this study, we explored sparse vector representations of frequent subgraphs using various weighting schemes. Each weighting scheme was tested using a basic content-based filtering algorithm and evaluated how effectively it represents the information in the subgraph.

Keywords: Frequent Subgraph Mining, Weighting Scheme, Knowledge Graph

1 Introduction

Knowledge Graphs (KGs) are graph-structured databases which consist of units called triples [1]. Knowledge graphs are used as target data for various analyses, one of which is frequent subgraph mining (FSM). An FSM method tries to find frequently occurring subgraphs in a large graph [2], and if this is applied to a KG, semantic information or related triples can be extracted. To use the semantic information for machine learning algorithms or real-world applications, a way of vector embedding (or representation) is required. Recently, some deep learning algorithms can provide a way of graph embedding, but the resulting dense vectors are hard to understand for human users. Also, they can only be generated through a certain specific task.

In this study, we explored sparse vector representations of frequent subgraphs, extracted from a KG of user-item interactions. We referred to the previous paper regarding the KG and FSM algorithms [3]. For each user's frequent subgraph, we applied three simple weighting schemes and generated three types of vectors accordingly. We then tested it through a content-based filtering problem with a leave-one-out setup. In other words, using the similarity of user's frequent subgraph vector

and item vector, we checked if the item that the user interacted last in the dataset could be recommended in a higher rank.

2 Weighting Schemes for Frequent Subgraphs

Through FSM algorithm, we could extract frequent subgraphs from a KG that has information about user's history of interactions with items [3]. Fig. 1 shows a frequent subgraph of a certain user. The subgraph consists of triples or attributes that the items, which the user has been interacted with, had in common. Note that types of the properties (edges between nodes) are ignored in this illustration. In the example, we could notice that the user-preferred features are A, B, and C, while D is not the one. The frequency of each feature is written next to the arrow.

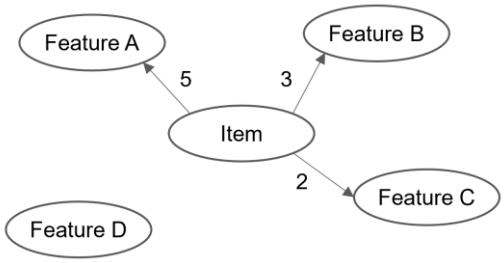


Fig. 1. Example of an extracted subgraph.

To represent this subgraph as a sparse vector, we could apply three simple weighting schemes: (1) binary, (2) frequency, and (3) ratio to maximum.

Binary. A subgraph is represented as a vector with a value of 0 or 1 depending on whether each triple is included in the subgraph or not. It ignores the strength of the user's preference.

Frequency. A subgraph is represented as a vector containing the number of occurrences of each feature. Frequency information is directly reflected as weight.

Ratio to maximum. A subgraph is represented as a vector similar to the frequency vector, but the weights are adjusted by dividing by the maximum frequency. It is a way to emphasize the relativity of preference more.

If the weight of the triple t in the vector is denoted by w(t), each weight is simply calculated as follows.

$$w(\mathrm{A})=1,\ w(\mathrm{B})=1,\ w(\mathrm{C})=1,\ w(\mathrm{D})=0 \quad (1)$$

$$w(\mathrm{A})=5,\ w(\mathrm{B})=3,\ w(\mathrm{C})=2,\ w(\mathrm{D})=0 \quad (2)$$

$$w(\mathrm{A})=\mathbf{5}/\mathbf{5},\ w(\mathrm{B})=2/\mathbf{5},\ w(\mathrm{C})=3/\mathbf{5},\ w(\mathrm{D})=0/\mathbf{5} \quad (3)$$

3 Evaluation Results and Conclusion

The dataset used in the experiment to create the KG is from https://www.kaggle.com/rounakbanik/the-movies-dataset, which is about movies and their metadata. The metadata were collected through the TMDB Open API, and the features includes actors, directors, countries, genres, keywords, etc. We used dataset called 'ratings_small' which contains 66,459 interactions from 634 users to 6674 items.

For evaluation, we left the last movie watched by each user and extracted a frequent subgraph from the graph consisting of the other movies that the user interacted. Then we generated a list of 100 movies including the last movie with 99 movies unseen by the user. After representing subgraphs to vectors according to the three weighting schemes, we calculated cosine similarity between the subgraph vectors and the selected 100 items vectors, which are binary vectors. Figure 2 briefly shows the concept of this similarity calculation. In the figure, Item* has three features: B, C, and D. Since there are no frequencies of triples in the graph of items, the vector can only be binary. When the cosine similarities between 100 items and the subgraph were calculated, we listed the 100 movies based on the similarity value and then check the ranking of the target movie.

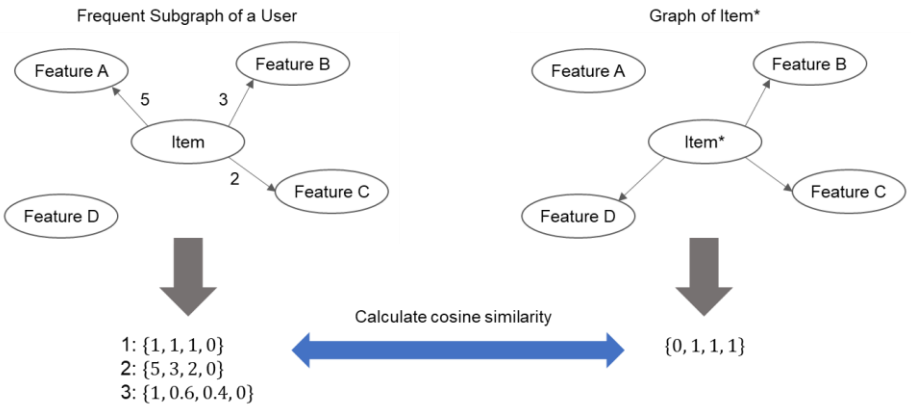


Fig. 2. Example of an extracted subgraph.

As a performance measure, we used Hit Ratio (HR) at rank 10. Namely, if the target movie (the last movie that the user watched) is in the top 10 list, it has a value of 1, otherwise 0. These values were calculated for all users, then averaged to get HR@10. Table 1 shows the performance of three weighting schemes.

Table 1. Similarity-based Content-based Filtering Performance

Weighting Scheme	HR@10
1. binary	0.298
2. frequency	0.311
3. ratio	0.305

The effective weighting scheme for expressing user preference in this test was frequency-based. In other words, the frequency in the subgraph of user serves as an indicator of user preference, and it is more effective to directly reflect strong preference. Although the performance difference according to the weighting scheme was not large, we could figure out that the weighing scheme do have some effect on the performance. However, since the effect of the weighting schemes may vary depending on the characteristics of the dataset or the target downstream task, tests on more diverse tasks are necessary.

Acknowledgments. This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2022-2020-0-01462) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00795, Development of Moire-Pattern Type 3D Camera System for AI Based Analysis).

References

1. L. Ehrlinger and W. Wöß, “Towards a definition of knowledge graphs,” in Jt. Proc. Posters Demos Track 12th Int. Conf. Semant. Syst., Leipzig, Germany, 2016. Available: http://ceur-ws.org/Vol1695/paper4.pdf
2. Jiang, C.; Coenen, F.; Zito, M. A Survey of Frequent Subgraph Mining Algorithms. Knowl. Eng. Rev. 2013, 28, 75–105.
3. K. Lee et al., “Learning knowledge using frequent subgraph mining from ontology graph data,” Appl. Sci., vol. 11, no. 3, pp. 932, Jan. 2021. DOI: http://dx.doi.org/10.3390/app11030932.

Detecting small objects on a PCB using YoloV5

In Joo[1], Sunghoon Kim, Ginam Kim, Kwan-Hee Yoo[1*]

[1] Dept. of Computer Science, Chungbuk National University, South Korea
{jooin95, sidsid84, khyoo}@chungbuk.ac.kr
*Corresponding Author

Abstract. The recent advancements of the 4th industrial revolution have automatized numerous humans tasks in the semiconductor manufacturing industry; in particular, defect detection is being rapidly digitized. During semiconductor production, real-time data analysis and discriminating technology are necessary. The application of artificial intelligence technology for detecting the presence or absence of various defective classes that occur in semiconductor manufacturing is emerging. In this study, the location and frequency of occurrence of various defects in the process control board (PCB) were detected using the YoloV5 model. The types of defects are largely between scratches, cracks, foreign substances, contamination, and colors. As the amount of data was unbalanced, the data was augmented and as a result of a test on 5,000 images of such data, a precision of 94.2% was obtained at the 0.5 position, which is the confidence during actual inspection.

Keywords: process control block, artificial intelligence, data preprocessing, data augmentation, YoloV5

1 Introduction

Recently, smart factories are maximizing process efficiency and management through data collection and analysis. We automated the process control block (PCB) through AI and developed a program to control high-quality products in real time. Since PCB is the first stage of electronic device manufacturing, a simple error can lead to major defects in the final product [1]. Therefore, it is important to detect a faulty PCB to avoid further obstacles and correct errors at the lowest possible cost. Defect inspection is the most critical step in PCB manufacturing [2]. PCB inspection consists of two main processes: defect detection and defect classification. The conventional method detects defects through visual inspection, image pattern analysis, or machine learning algorithms; we proceed with defect detection and classification using a convolutional neural network (CNN) [3]. We classify five bad classes in PCB using YoloV5 among CNNs. In this study, we investigate previous similar methods in PCB defect detection, describe the proposed method, and explain our results. Our PCB data had five types of defects: scratches, cracks, color differences, contamination, and foreign substances.

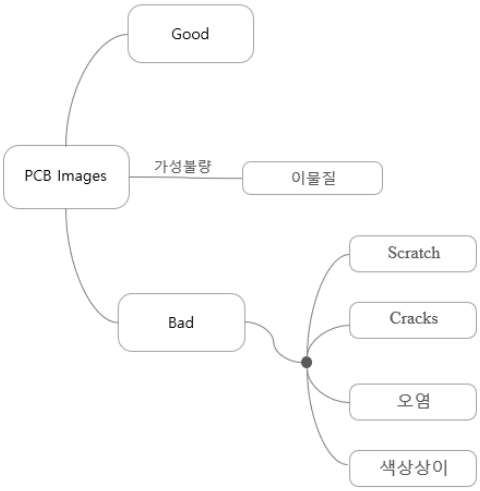


Fig. 1. PCB bad data type

The size of one PCB image is 4096 × 3000 × 24 and considering that the maximum size of one PCB is 240 mm × 90 mm, good resolution is maintained. After amplifying 200 pieces of bad data using augmentation, the image was converted to a size of 1024 × 1024 × 24 through preprocessing and trained. Chapter 2 below describes the preprocessing and methodology.

2 YOLO Model for detecting PCB defects

In this section, we propose a method for classifying defects on a PCB according to the procedure shown in Figure 2. The data size is 4096 × 3000 × 24. Therefore, the sheet is too large for AI learning, and so the image data is processed and normalized through ROI, cutting unnecessary parts of the image, and converting the image size to 1024 × 1024 × 24. Then, to increase the amount of data, data augmentation was performed by rotating and flipping the image by 90˚ and changing the color. The data set was divided into 80% and 20% for training and testing, respectively. We did this by using the YoloV5 model. The model was trained with a batch size of 16 and 1000 epochs. The learning rate was multiplied by 0.1 for every 100 progresses from 0.01. The Yolo model we used finds the bounding box of an object in an image and recognizes its class.

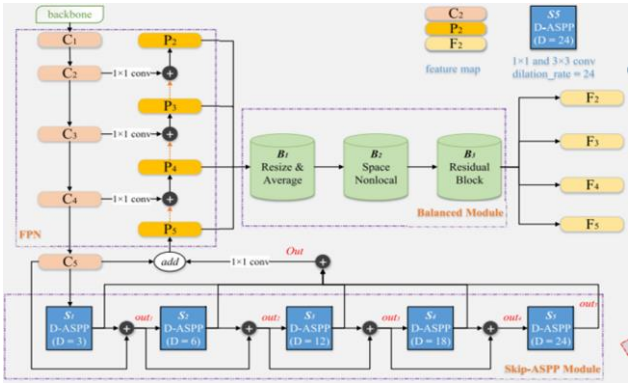


Fig. 2. Overall approach for detecting defects in PCB images [4]

The Yolo model features are specialized for detecting protruding objects and have a pyramidal structure, which greatly facilitates the development of object detection using CNNs. Furthermore, feature maps outputted from thin layers generally have a larger spatial size and maintain sophisticated and detailed low-level patterns. Feature extraction is also performed with the use of the CSP-Darknet backbone and general intersection of union (GIoU)

3 Experimental Results and Future Work

The criterion for good or bad is to measure the green sheet for air bubbles and scratches. Two results can be derived from applying the proposed model. We first classified defective and good products. As shown in Fig. 4, the proposed model had a detection accuracy of up to 99% for good and defective parts.

We performed defect detection as described below. AI detection accuracy was calculated by dividing the number of PCBs with AI judgment by the total number of inspected PCBs × 100, and the criteria for PCBs with AI judgment were compared with two results. The first is a bad AI reading, where the bad class matches, and if the coordinate regression predicted value and the intersection of union (IoU) [5] of ground-truth are 0.5 or greater, the prediction is estimated to be successful.

$$Accuracy = \frac{True\,Positives}{True\,Positives + False\,Positives}$$

Fig. 3. Overall approach for detecting defects in i-ceramic images

Second, we used the number of PCBs to determine that the inspector was judged as normal through conventional defect detection.

The pictures below are the bad features we detected using YoloV5. The right part is the error detected by AI, and the left part is the data used for training. As a result of bad detection, the object range and probability of the bad class are detected.

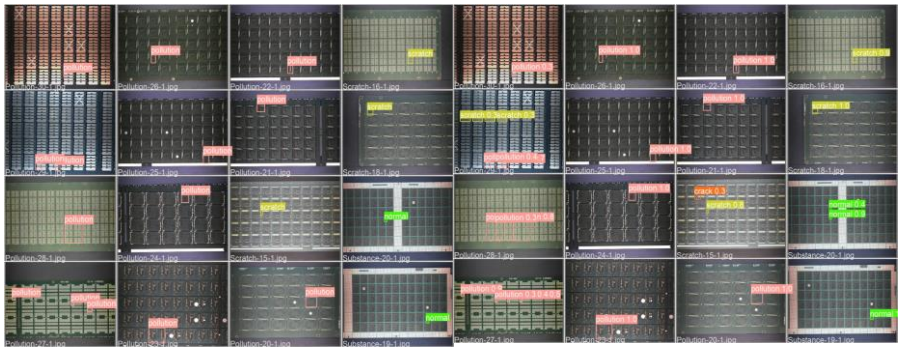


Fig 4. Classification report of predicts for defective and good products

The bounding boxes of every defect are provided in our dataset for the location of each defect to be affirmed; besides, the existing bounding box makes it possible for the images to be utilized as labeled data in object detection tasks.

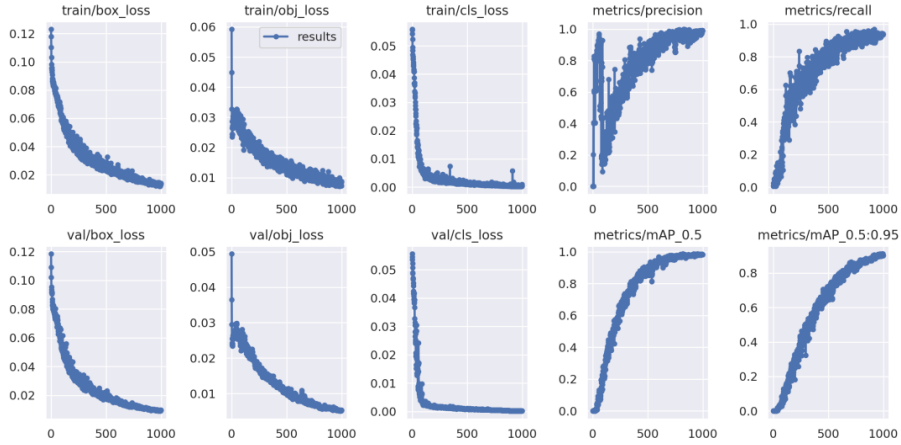


Fig 5. Classification report of predicts for defective and good products

Figure 5 shows the results of the learning of the YoloV5 model. As the amount of data increases, the results may vary; however, for now, the detection rate is approximately 95%. The limitation of this study is that YoloV5 does not detect objects that are too small as defective. In future studies, we will modify YoloV5 to improve small feature detection.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2022-2020-0-01462) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

References

1. Chauhan, Ajay Pal Singh, and Sharat Chandra Bhardwaj. "Detection of bare PCB defects by image subtraction method using machine vision." Proceedings of the world congress on engineering. Vol. 2. 2011.
2. Anitha, D. B. and Mahesh Rao. "A survey on defect detection in bare PCB and assembled PCB using image processing techniques." 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). IEEE, 2017.
3. Sang-Won Suh, et al. "Development of Checker Switch Failure Detection System Using CNN Algorithm." Journal of the Korean Society of Mechanical Engineers 18.12 (2019): 38-44.
4. Zhu, Xingkui, et al. "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021...
5. Rezatofighi, Hamid, et al. "Generalized intersection over union: A metric and a loss for bounding box regression." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Simple Yet Effective Data Augmentation for Imbalanced Solar Panel Soiling Image Dataset using Image Manipulation

Eul Ka[1,*], Seungeun Go[1,*], Ulziitamir Davaadorj[1], Geun-Min Hwang[1], Minjin Kwak[2] and Aziz Nasridinov[1,*]

[1]School of Computer Science, Chungbuk National University, Cheognju 28644, South Korea

[2]Department of Big Data, Chungbuk National University, Cheognju 28644, South Korea

{rkdmmf99, gotmddms05, tamiraa, miindiin99, aziz}@chungbuk.ac.kr, hgma12(at)naver.com

* Both authors contributed equally to this research

*Corresponding Author

Abstract. A large dataset of diverse solar panel images can increase the performance of soiling segmentation models and avoid overfitting problems. However, collecting a large training dataset in the solar panel field is challenging as it takes much time and resources. Existing datasets suffer from the imbalance problem due to a lack of diversity in soiling shapes and solar panels. In this paper, we propose a simple yet effective data augmentation to solve the imbalance issue of the solar panel soiling image (SPSI) dataset using image manipulation. We verified the usefulness of the proposed dataset (called SPSI+) using three state-of-the-art semantic segmentation methods. The experiment results show that the performance of soiling segmentation methods can be improved when the proposed SPSI+ dataset is used as a training dataset.

Keywords: Solar panel, deep learning, image segmentation

1 Introduction

Recently, the global interest in solar power generation has significantly increased. Predicting the amount of solar power generation is essential, as it enables consumers and businesses to plan their consumption. Various causes (e.g., temperature or panel surface soiling) may change the amount of power generation. Panel surface soiling is the main factor that causes reduced power generation (John et al., 2015). Therefore, accurately locating the soiling on the solar panel can help predict the amount of power generation.

A dataset of solar panels with enough information on various soiling can significantly increase the performance of soiling segmentation models. Mehta et al. proposed the Solar Panel Soiling Image (SPSI) benchmark dataset. The SPSI dataset consists of 45,754 panel images with six types of soiling. However, it has two limitations. First, the SPSI dataset consists of only one type of solar panel.

Considering that there are many types of solar panels in the field, the model trained based on the SPSI dataset may misclassify patterns of certain solar panels as soiling. Second, out of 45,754 panel images in the dataset, only 84 images contain the unique shape of soiling. In other words, the soiling region can be wrongly segmented when the model is trained based on this dataset.

In this paper, we will demonstrate that with simple data augmentation on the SPSI dataset, we can achieve even higher accuracy in locating the soiling on the solar panel. For this, we applied various image manipulation techniques to the original SPSI dataset. As a result, we produced a balanced dataset (called SPSI+) containing 46,840 images of unique soiling shapes on various types of solar panels. To validate the usefulness of the proposed SPSI+ dataset, we implemented several state-of-the-art semantic segmentation methods to detect soiling on the solar panel.

2. Our Model

The details of our model is shown in Figure 1. It contains of five parts: data collection, data preprocessing, data augmentation, prediction and comparison with state-of-the-art image segmentation methods.

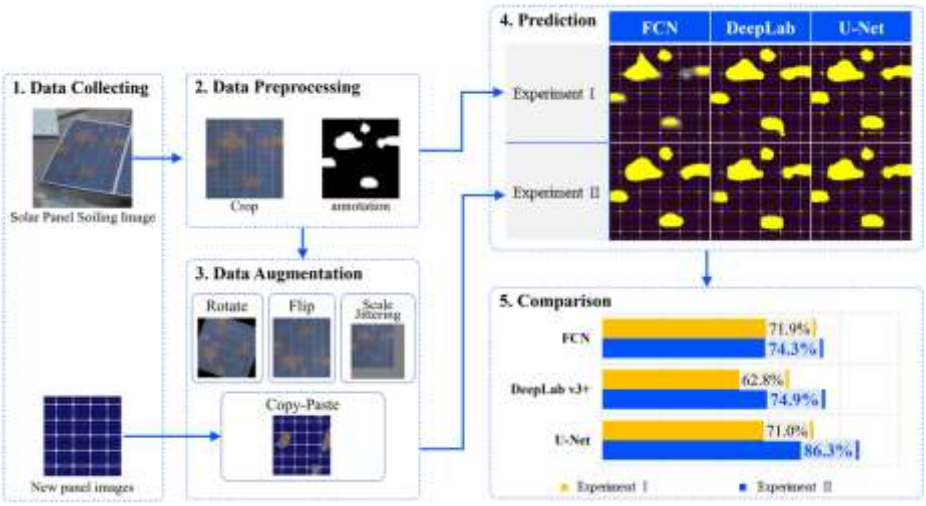


Figure 1. Overview of the proposed methodology.

Data Collection and Preprocessing. We used 29,190 images (excluding clean panels) from the original SPSI dataset. We first crop the solar panel region from the images. We then manually annotate the soiling from each image as the SPSI dataset does not provide the soiling annotation. As for different types of solar panels, we used ten images of panels that contain various patterns.

Data Augmentation. We first selected all 84 unique soiling images from the original SPSI dataset. We then manipulated each image by rotating, vertically and

horizontally flipping, and scale jittering. Here, all manipulations occur randomly, 50 times per image. For scale jittering, we used the resize range from 0.5 to 2.0 of the original image size. Finally, we used the Copy-Paste method (Ghiasi et al., 2021) to copy the soiling part from augmented images of solar panels and pasted them to other types of solar panels. Table 1 shows the details of the SPSI+ dataset after augmentation.

Prediction. Three state-of-the-art semantic segmentation models are used to verify SPSI+ dataset: Fully Convolutional Networks (FCN) (Long et al., 2015), DeepLab v3+ (Chen et al., 2018), and U-Net (Ronneberger et al., 2015).

Table. 1. Details of the proposed SPSI+ Dataset.

Type	Number
Unique soiling image In SPSI Dataset	84
Copy-Paste without manipulation	756
Copy-Paste with manipulation	42,000
Crack Images	4,000
Sum	46,840

3. Performance Evaluation

To verify the usefulness of the proposed SPSI+ dataset, we performed two kinds of experiments as follows:

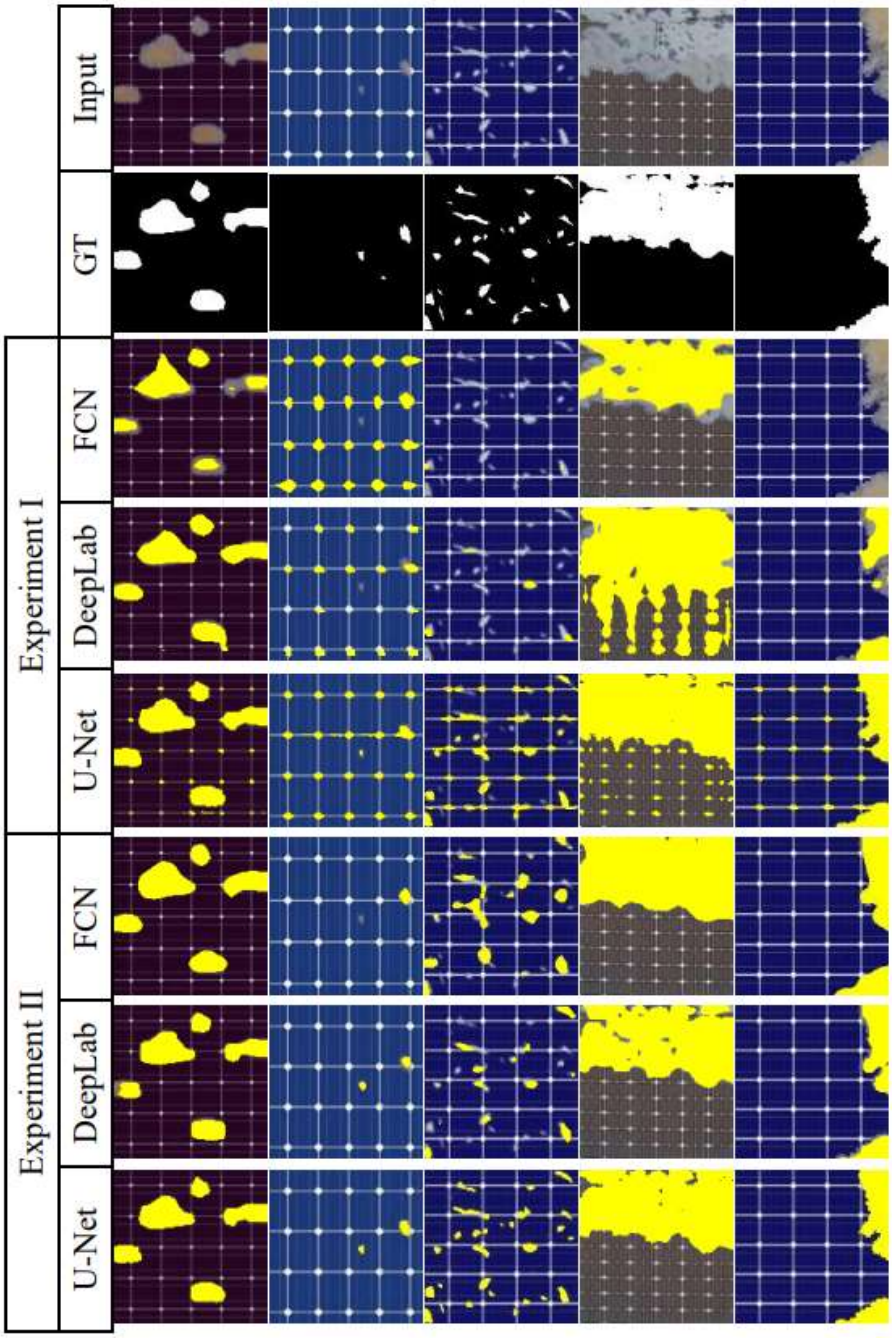
- Experiment I: Performance of soiling segmentation methods when the original SPSI dataset is used as a training dataset.
- Experiment II: Performance of soiling segmentation methods when the proposed SPSI+ dataset is used as a training dataset.

Table 3 shows a visual comparison of each Experiment I and II. From Table 2, we can observe that the performance of soiling segmentation significantly degrades when the original SPSI dataset is used as a training dataset (see Experiment I). That is, soiling segmentation models misclassify patterns of different types of solar panels as soiling (i.e., circle markers on the solar panels). In addition, we can also observe that the soiling region is wrongly segmented when the model is trained based on the original SPSI dataset (see Experiment I). On the other hand, we can achieve more accurate soiling segmentation when soiling segmentation models are trained based on the proposed SPSI+ dataset (see Experiment II). We used the Jaccard Index to quantify the performance. Table 2 shows the Jaccard Index of the models trained in Experiment I and II.

Table. 2. Visual comparison of predicted results

Experiment	FCN	DeepLab	U-Net
I	71.9%	62.8%	71.0%
II	**74.3%**	**74.9%**	**86.3%**

Table. 3. Visual comparison of predicted results.



4. Conclusion

In this paper, we have proposed a new dataset, called SPSI+, using simple yet effective data augmentation techniques. We have also verified the usefulness of the proposed dataset using three state-of-the-art semantic segmentation methods. The experiment results demonstrated that the proposed dataset could solve not only the imbalance issue of the original SPSI dataset but also boost the performance of semantic segmentation models by 2.4% in FCN, 12.1% in DeepLab v3+, and 15.3% in U-Net. To the best of our knowledge, this is the first solar panel soiling image dataset that contains unique soiling shapes on various types of solar panels. To encourage the community to explore more novel research topics, we made the SPSI+ dataset available at: https://spsiplus.github.io/

Acknowledgments. This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW)(2019-0-01183) supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation).

References

1. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV). 801-818. Munich, Germany. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T. Y.; Cubuk, E.
2. D.; Le, Q.V.; and Zoph, B. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Virtual Conference. 2918-2928.
3. John, J. J.; Rajasekar, V.; Boppana, S.; Chattopadhyay, S.; Kottantharayil, A.; and TamizhMani, G. 2015. Quantification and modeling of spectral and angular losses of naturally soiled PV modules. IEEE journal of photovoltaics 5(6): 1727-1734.
4. Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 3431-3440. Boston, MA, USA.
5. Mehta, S.; Azad, A.; Chemmengath, S.; Raykar, V.; and Kalyanaraman, S. 2018. Deepsolareye: Power loss prediction and weakly supervised soiling localization via fully convolutional networks for solar panels. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 333-342. Lake Tahoe, NV, USA.
6. Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (MICCAI). 234-241. Munich, Germany.

Recognition of Various Behaviors of Pigs Using Deep Learning Algorithm

Sooram Kang[1], Sangkyoon Kim[2], Wanhyun Cho[2], Myung Hwan Na[2],

[1] Department of Mathematics and Statistics, Chonnam National University, 61186, Republic of Korea
[2] Department of Statistics, Chonnam National University, 61186, Republic of Korea
slkang2001@gmail.com, {narciss76, whcho, nmh}@jnu.ac.kr

Abstract. For farmers, it is important to diagnose the growth and health of livestock. Proper growth is related to energy consumption and business performance. In addition, the health is related to the presence of disease. The growth and health of livestock can be inferred from the patterns of activity and food intake. However, such information is difficult to be obtained without expert intervention. In order to overcome these, this study proposes a deep learning method that can classify livestock behavior without human intervention. A data set was constructed by collecting images through a camera installed in a farmhouse and defining the behavioral patterns of livestock into four classes. For the usability in edge terminals, the YOLO model, which is light and shows constant performance among object detection algorithms, is used. As a result of the experiment, the proposed model showed high performance with mAP 99% in all classes.

Keywords: Pig behavior monitoring, Behavior classification, YOLO

1 Introduction

In general, it is known that activity, food and water intake are related to the growth and health status of livestock. In particular, it is very important to know the health status of livestock in advance because an extreme situation in which the livestock of the entire farmhouse must be slaughtered can occur if the livestock becomes ill. Therefore, it is important for famer to use this information to grasp the status of livestock and predict business performance. However, it is physically difficult to monitor individual livestock from the perspective of a farm that raises a large number of livestock. Despite a lot of effort, it is simply impossible to visually identify a health of livestock without the help of a professional. Behavioral analysis of livestock is one of the methods to judge the health of livestock and realize animal welfare. To solve this problem, attempts are being made to use computer vision, deep learning algorithms and sensors.

In this study, using a deep learning model, pig activity and behavior such as food and water intake were classified from images. The data used in the study were

collected by installing a camera on the ceiling of a barn at a pig-raising farm. For labeling for behavior classification, only images with large differences were extracted, and then four classes of lying, standing, eating, and drinking were created. For operability in edge terminals, YOLO v4 which is known to be light and have good performance among multi-object detection models, was used.

2 Experimental Material

2.1 Data Collection

The data set used in the study was collected from a pig farm in Hampyeong, Jeollanam-do, South Korea. The cage is 5m wide and 3m long, with a circular feeding container in the middle of the cage and two water dispensers on the wall. The camera was installed on a metal bar 2.8m above the cage floor and filmed 24 hours a day. Images were collected from two cages with 7 and 8 pigs, respectively. Pigs raised on farms are largely classified into three stages as piglet, finishing pig and shipping pig. Especially, data of finishing pigs whose growth have increased rapidly was collected in this study. The frame rate of the camera is 1 frame per second and the resolution of the image is 640x360 pixels. In the experiment, images collected for 10 days were analyzed and pigs between 30 kg and 80 kg were targeted. Fig.1 shows an example of collected image.

Fig. 1. Example of collected pig images by camera installed 2.8m above from the floor.

2.2 Image Labeling

Classes that are useful to identify and classify specific patterns among various behaviors of pigs identified through the collected data was defined. Finishing pig spent a lot of time lying down and sleeping. During waking hours, they could smell or

walk around, but it was difficult to define those activities to specific patterns. In addition, they ate regardless of the time, and especially, when one individual ate its food, it was confirmed that the other individuals also ate their food. Through these data observations, four representative classes were defined which are lying, standing, eating, and drinking. Among the 864,000 images collected for 10 days, 7,889 images with large differences between frames were extracted and labeled. As described above, since pigs spent a lot of time lying down, the class 'lying' was the most with 28,917 out of 44,434 counts. Next, standing, eating, and drinking had the most in the order of 6,575, 6,753, and 2,189, respectively.

3 Experimental Method

3.1 YOLO v4

Modern object detectors mainly consist of two parts which are the backbone and the head. The backbone is the part that transforms the input image into a feature map. VGG16 and ResNet-50 which are pre-trained with ImageNet dataset are representative backbones. The head is the part that performs the location operation of the feature map extracted from the backbone. In the head, predicting classes and bounding boxes are performed. The head is largely divided into dense prediction and sparse prediction, which is directly related to whether the object detector is one-stage detector or two-stage detector. Representative one-stage detectors using dense prediction heads include YOLO and SSD. Unlike the two-stage detector, predicting classes and bounding box regression are integrated.

The neck is a part that connects the backbone and the head. It refines and reconfigures the feature map. Representative examples include Feature Pyramid Network(FPN), Path Aggregation Network(PAN), BiFPN and NAS-FPN. In case of YOLO v4, it uses CSP-Darkent53 as backbone, SPP and PAN as neck, and YOLO v3 as head. In addition, it incorporates a variety of trending techniques that increase accuracy and speed up calculations such as bag of freebies and bag of specials.

3.2 Detection Metrics

IoU(Intersection over Union) is an indicator that evaluates the accuracy of estimating the location of an object. It is evaluated through the size of the area where the ground truth which is the location of the real object and the model's prediction overlap. The large overlapping area of the two boxes means that the model estimated the position of the object well. IoU has a value between 0 and 1. In general, an IoU of 0.5 or higher is considered as correct. Precision is the ratio of data whose prediction and actual value are positive among subjects for which the model has predicted as positive, and recall is the ratio of data whose prediction and actual value are positive among subjects whose actual value is positive.

The confidence score is an indicator of how certain it is that the object detected by the model belongs to a specific class. Therefore, if an appropriate threshold is set for the confidence score, a bounding box with low accuracy of object detection can be regarded as a detection failure. The curve showing the precision value according to the change of the recall value is called PR curve(Precision-Recall Curve). The average of the precision values is called AP(Average Precision), and is generally calculated as AUC(Area Under Curve) value of PR curve. AP is a comprehensive index obtained by considering both precision and recall, and it is easy to quantitatively compare the performance of two different object detection models. AP measures the detection performance of a model for a single object. To measure the overall detection performance of a model for all objects, mAP(mean average precision) is used, which averages all AP values.

4 Experimental Results

Of the total 7,889 images, 4,898 were used for training the YOLO v4 model, and 2,100 were used for validation. And the model was tested with 891 images. There was a total of 7,164 counts of classes in the 891 test images, of which 4,276 are lying, 1,835 are eating, 796 are standing, and 257 are drinking. As a result of the test, the number of false positives for each class was 8 for lying, 6 for eating, 6 for standing and 5 for drinking, and the mAP was 99.47%. Each class showed a very high accuracy of 99%. Fig. 2 shows the number of false positives and true positives and mAP as the experimental results. Fig. 3 shows the P-R curve for each class. All classes showed very high accuracy of over 99%.

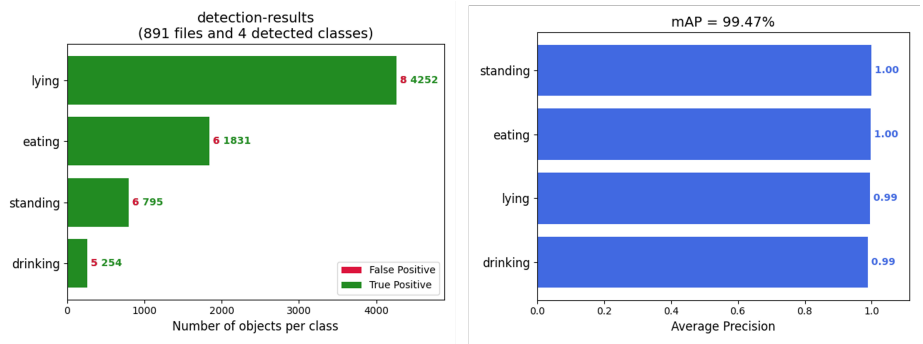


Fig. 2. Detection results on left and mAP on right of YOLO v4 model.

5 Conclusions

Monitoring the growth and health of livestock is a very important but difficult task for the farmer. Assessing the condition of individual livestock requires expert intervention or the use of specific equipment. Since direct measurement of livestock

status is difficult, it is necessary to infer it using indirect indicators such as activity, food intake and water intake. However, while a large number of livestock are raised in a farmhouse, the input of labor force and energy is very limited. For the reason, measurement using such indirect indicators is also difficult. In order to overcome those difficulties, this study proposed a model that identifies and classifies activity, food intake and water intake, which are indicators that can infer the state of livestock without human intervention. To this end, a camera was installed in the farmhouse to collect image data, and by analyzing the collected images, four classes that can specify the behavioral patterns of pigs were defined.

The four classes of lying, standing, eating, and drinking classified with a high degree of accuracy by proposed model, although the value of classes is relatively unbalanced because finishing pigs spend a lot of time lying down. YOLO v4, which is light while has high performance among object detection algorithms, was used under the consideration to transplant the model in edge terminals and to classify pattern pig behavior in real time in a further study. Through this study, it is expected to infer the growth and health status of livestock by classifying and patterning the behavior of livestock, which can be used as a useful tool in farms.

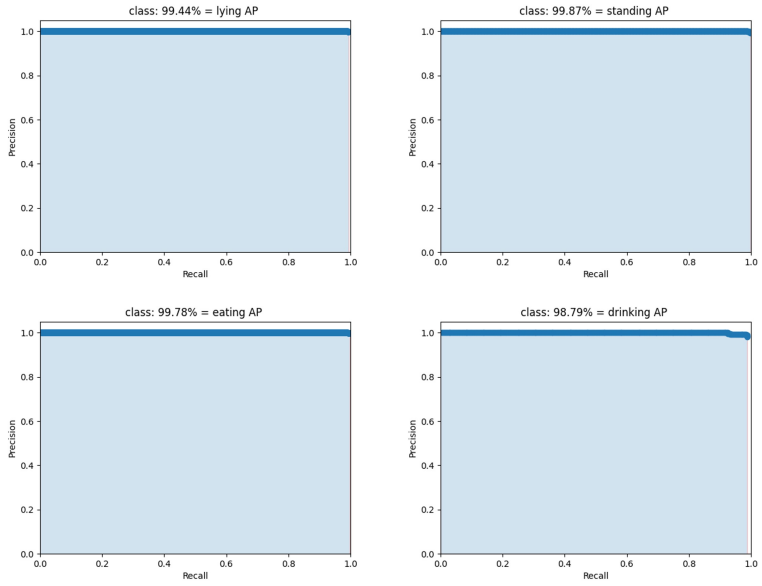


Fig. 3. PR Curves of each class by YOLO v4 model.

Acknowledgments. This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry(IPET) through Smart Plant Farming Industry Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(421017-04)

References

1. Yang, Q., Xiao, D.: A review of video-based pig behavior recognition. Applied Animal Behaviour Science. vol. 233, 105146. (2020)
2. Yang, Q., Xiao, D., Lin, S.: Feeding behavior recognition for group-housed pigs with the Faster R-CNN. Computers and Electronics in Agriculture. vol. 155, pp. 453-460. (2018)
3. Tu, S., Yuan, W., Liang, Y., Wang, F, Wan, H.: Automatic Detection and Segmentation for Group-Housed Pigs Based on PigMS R-CNN. Sensors. vol. 21, 9, 3251. (2021)
4. Nasirahmadi, A., Sturm, B., Edwards, S., Jeppsson, K.-H., Olsson, A.-C., Muller, S., Hensel, O.: Deep Learning and Machine Vision Approaches for Posture Detection of Individual Pigs. Sensors. vol. 19, pp. 2--15. (2019)
5. Bochkovskiy, A., Wang, C., Liao, H.-Y.: YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv:2004.10934. (2020)

The 10th International Conference on Big Data Applications and Services (BIGDAS 2022), November 24-26, 2022 – Jeju Island, South Korea

Topic Modeling-Based Case Analysis for Inductive Social Science Research Methods

Flor Gutierrez De la Cruz[1], *Keunhyung Kim*[2†]

[1]*Master Degree Student, Faculty of Data Science for Sustainable Growth (Management Information Systems Major), & BK21 Social Data Science Research Center, Jeju National University, Jeju, Korea. gutierrezdelaflor@gmail.com*

[2†]*Professor, Faculty of Data Science for Sustainable Growth (Management Information Systems Major), Graduate School, Jeju National University, Korea. khkim@jejunu.ac.kr*

Abstract. In this research topic modeling technique is extended to generate causality information from a set of text documents such as online reviews, and an example of its use is presented. The expansion of topic modeling and case analysis are carried out in three stages. In the first step, an arbitrary number of topics are extracted from online reviews through a vectorization process and a topic modeling process, and a topic name is given to each topic based on the interpretation of keywords included in the extracted topics. In the second step, the topic ratio data for each document is converted into a CSV file, which is integrated processing data, through a post-processing process. In step 3, a causal model is established using topics of integrated processing data as independent variables and external variables as dependent variables, and regression analysis is performed to verify this. The extension method of topic modeling proposed in this research can be used in various ways in areas such as Internet search results and news articles as well as online reviews.

Keywords: Topic Modeling, Social Science Research, Inductive Method, Seongsan Ilchulbong, Regression Analysis.

I Introduction

Social science is an academic field that scientifically identifies various phenomena in our society composed of people and creates new knowledge. Existing social science research methods mainly used structured data collected based on structured questionnaires to find out people's opinions and thoughts. Before collecting survey data, the research hypothesis was derived through a deductive method through a literature review. In terms of scientific verification of the established hypothesis, the data collected by the survey were analyzed for hypothesis verification purposes. Hypothesis derivation by deductive methods may be powerful in terms of theoretical rigor and consistency, but there may be limitations in deriving innovative and diverse hypotheses.

Text data that records people's thoughts and behaviors is an important source for social science. A large number of online reviews posted on websites such as online shopping malls are a type of textual data freely written by many reviewers. An online review written by a specific reviewer may correspond to a survey on the reviewer. Analyzing unstructured text data, such as online reviews, by methods such as text mining, a consistent pattern of text can be found, and based on this, a new hypothesis representing social phenomena can be derived. Since a hypothesis is derived based on data, it can be called an inductive hypothesis. The concept of deriving a new hypothesis from a repetitive pattern inherent in a large amount of text data already generated is the inductive social science research method proposed in this paper.

Although existing topic modeling techniques can generate useful summary information on a set of text documents, they have limitations as a tool for generating social science knowledge. This is because the basic structure of knowledge should be one that can clarify causal relationships. Existing topic modeling cannot derive a causal relationship between topics. In this research, we extend topic modeling to derive causal relationships between topics.

In this research, we intend to present an analysis case of Seongsan Ilchulbong Online Review by extending the topic modeling technique to generate causal information from a set of text documents such as online reviews.

II. Theoretical Background

2.1 Topic Modeling

Topic modeling is a text analysis technique that finds a topic in a large collection of documents [1]. From the point of view of text analysis, topic is defined as a bag of words, it represents a pattern in which specific words appear together repeatedly. The result of topic modeling is generally a list of words that compound a topic, allowing us to infer topics that exist in the document. Text documents are a mixture of topics, consisting of a variety of topics, and each topic is a word mixture composed of words.

Latent Dirichlet Allocation (LDA) is a representative topic modeling algorithm [2]. Topic modeling using LDA can be seen as a process of assigning each word in a document set to each topic. Through the optimal word-topic assignment, a list of words with a high correlation with each topic is found, and the topic composition of each document is identified. LDA uses the Dirichlet distribution, a prior probability distribution for estimating the probability distribution of topics, and the word probability distribution. The reason why the Dirichlet distribution is used in topic modeling is that when the Dirichlet distribution is multiplied by other polynomial distribution functions, it becomes a Dirichlet distribution form again, which is advantageous in creating a posterior distribution using the observed word.

2.2 Prior Research

Online comments are specific text data in which individuals as a part of society express their thoughts and opinions on the internet, these online comments which are voluntary and sincere expressions of experiences and opinions can contribute to social science research. In the analysis of online comments, the text mining technique is widely used since online comments are a collection of text documents [3]

Topic modeling makes it possible to determine the important topics contained in a set of documents, which are the topics cited in online comments [2]

Many studies analyze text data using topic modeling in various domains. <Table 1> shows studies using topic modeling techniques.

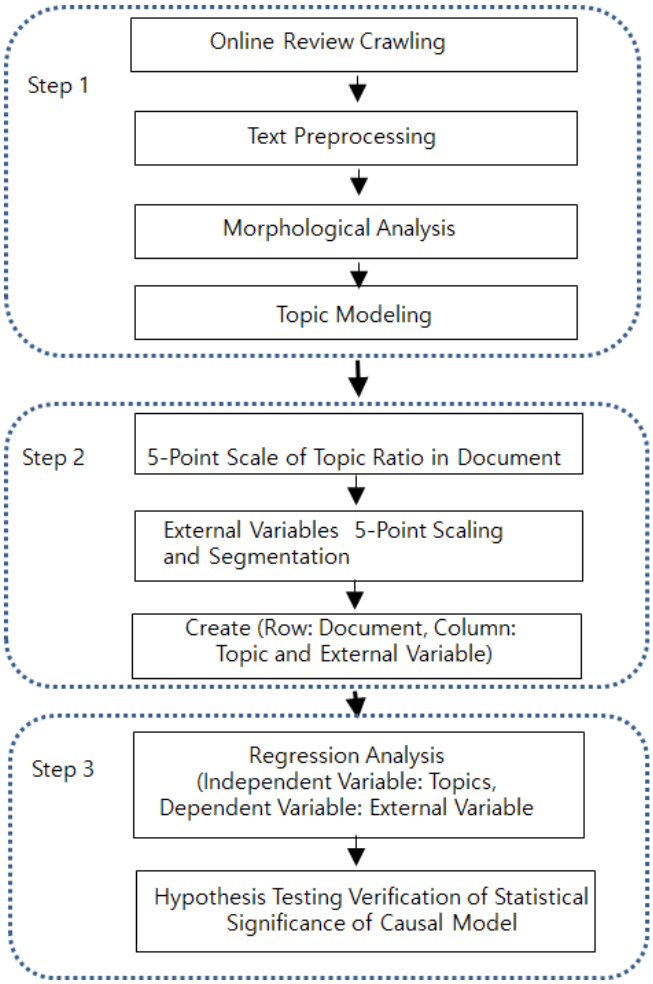
<Table 1> Summary of Research Trends Based on Topic Modeling

Research	Content
Analysis of consumer perception of eating out due to the spread of COVID-19: Application of topic modeling and network analysis [4]	Topic modeling and semantic network analysis were conducted to analyze consumer perceptions of eating out in the COVID-19 era. Frequency analysis of keywords in the text is predominant, and there is no causal model analysis.
Research Trends on Factors Affecting Quality of Life in Cancer Survivors: Text Network Analysis and Topic Modeling [5].	Using semantic network analysis and topic modeling, the keywords of the study on cancer survivors' quality of life influencing factors are identified, and the characteristics of the network by major topics are checked to explore the knowledge structure of the study on cancer survivors' quality of life.
Topic modeling and sentiment analysis of the service quality of complex resorts in Korea [6].	Based on online reviews written by overseas tourists who experienced Jeju Complex Resort, important service quality factors that can affect tourists' satisfaction are derived using topic modeling, and tourists' emotions contained in online reviews are classified into positive, negative, and neutral.
Global Green New Deal Research Trend Analysis by Topic Modeling Analysis [7].	Various academic studies related to Green New Deal in Korea, European Green Dea, and Green New Deal in the United States are compared and analyzed using topic modeling, and major research trends are derived.
Artificial Intelligence and Human Security: Analysis of the Security of Artificial Intelligence in Europe Using Topic Modeling Techniques [8].	Major topics are derived using topic modeling in the security discourse of media articles in Europe on artificial intelligence.
Analysis of research trends in domestic and foreign financial security using topic modeling analysis techniques [9].	To derive key research fields in the financial security field and to present directions, a comparative analysis of major research trends at home and abroad is conducted through topic modeling.

III. Case Analysis

3.1 Analytical design

Analyze actual online reviews by extending topic modeling, online reviews use reviews of tourism products from TripAdvisor (www.tripadvisor.co.kr), a tourism website. <Figure 1> shows the trip advisor's online review analysis process in three stages. In the first stage, line reviews are automatically collected from the TripAdvisor website by on-crawler. Data is collected using a previously developed web crawler in a Python anaconda environment. A text preprocessing process is performed to remove unnecessary words from the collected online reviews. Each sentence of the online review to which text preprocessing is applied is divided into morphemes, which are the smallest words with meaning by applying the tokenization technique. Morphemes are generated using a KoNLPy morpheme analyzer. K topics are extracted from online reviews and converted into morpheme words through a vectorization process and a topic modeling process. The extracted topics are given a topic name based on the interpretation of the keywords. In the second step, the topic ratio data for each document is converted into integrated processing data (csv file) based on a 5-point scale through a post-processing process. In step 3, regression analysis for causal model verification is performed using topics of integrated processing data as independent variables and external variables as dependent variables.



<Figure 1> Analysis Process of Online Review Text

3.2 Data Collection

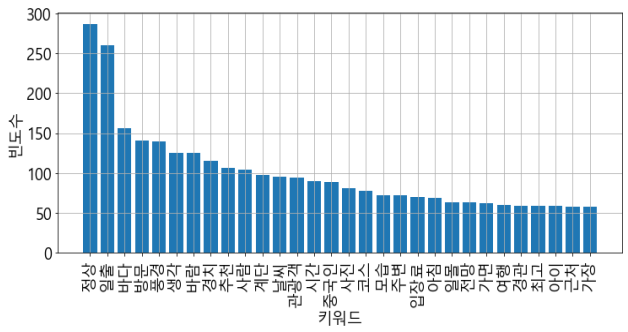
Online reviews of Seongsan Ilchulbong Peak were collected by developing a Python crawler. Python crawler was developed using the Selenium library in the Python Anaconda environment. <Figure 2> shows the structure of the collected online reviews. Online reviews were registered between 2015 and 2020.

	Number of Reviews	Writing Period	Data Structure
Review of Seongsan Ilchulbong Peak	1000	2015 to 2020	Number: Numeric Content: Text (character type) Rating: Numeric Date: Date type

<Figure 2> Online Review Data

3.3 Text Preprocessing and Basic Analysis

Noun words extracted from online reviews were vectorized into a Term-Document Matrix, and the top 30 frequently expressed words were printed on a bar graph and a word cloud. As can be seen in <Figure 3>, words such as 'peak', 'sunrise', and 'sea' frequently appeared.

Frequently Appearance List(Top 30)	
정상, 일출, 바다, 방문, 풍경, 생각, 바람, 경치, 추천, 사람, 계단, 날씨, 관광객, 시간, 중국인, 사진, 코스, 모습, 주변, 입장료, 아침, 일몰, 전망, 가면, 여행, 경관, 최고, 아이, 근처, 가장 (Peak, sunrise, sea, visit, scenery, thinking, wind, landscape, recommendation, person, stairs, weather, tourist, time, Chinese, photo, course, shape, surroundings, admission, morning, sunset, view, mask, travel, scene, best, child, near, most)	
	
Bar Graph	Word Cloud

<Figure 3>Most frequent Appeared Words in Seongsan Ilchulbong Online Review

3.4 Topic Modeling and Causality Analysis

<Table 2> Shows the case of extracting three topics for the Seongsan Ilchulbong Review. The words included in Topic 1 and Topic 2 often overlap, so they were named 'Tourism Activity A' and 'Tourism Activity B'. Topic 3 was named 'family experience' based on words such as 'family' and 'child'.

<Table 2> Main Topics of Seongsan Ilchulbong (3 Topics)

Topic Name	Key Words(Top 20)
Tourism Activity A (Topic 1)	'0.040*"sunrise" + 0.028*"visit" + 0.024*"peak" + 0.018*"tourist" + 0.018*"sea" + 0.015*"scenery" + 0.015*"people" + 0.014*"chinese" + 0.012*"view" + 0.012*"time" + 0.011*"stairs" + 0.011*"wind" + 0.010*"recommended" + 0.009*"china" + 0.008*"feel" + 0.008*"sunset" + 0.007*" top" + 0.007*"landscape" + 0.007*"near" + 0.007*"most"
Tourism Activity B (Topic 2)	'0.034*"peak" + 0.019*"course" + 0.018*"think" + 0.016*"wind" + 0.014*"photo" + 0.014*"around" + 0.012*"sunrise" + 0.010*"sea" + 0.010 *"scene" + 0.009*"admission" + 0.009*"sunset" + 0.008*"parking lot" + 0.008*"landscape" + 0.008*"best" + 0.007*"weather" + 0.007*"best" + 0.007*"scenery" + 0.006*"people" + 0.006*"essential" + 0.006*"mask"
Family Experience (Topic 3)	'0.022*"weather" + 0.020*"peak" + 0.017*"landscape" + 0.015*"scenery" + 0.014*"recommended" + 0.013*"Sea" + 0.012*"think" + 0.011*"Sunrise" + 0.011 *"length" + 0.010*"health" + 0.010*"far" + 0.010*"stairs" + 0.009*"time" + 0.009*"admission" + 0.009*"beautiful scenery" + 0.009*"climb" + 0.009*"middle" + 0.008*"surroundings" + 0.008*"child" + 0.008*"family"

<Table 3> Shows the case of extracting four topics for the Seongsan Ilchulbong Peak review. Topic 1 was named 'Personal Experience A', Topic 2 was named 'Tourism Activity A', Topic 3 was named 'Tourism Activity B', and Topic 4 was named 'Personal Experience B'.

<Table 3> Main topics of Seongsan Ilchulbong (4 topics)

subject name	Keywords (Top 10)
Personal Experience A (Topic 1)	'0.023*"landscape" + 0.020*"sea" + 0.019*"sunrise" + 0.017*"recommended" + 0.015*"peak" + 0.012*"weather" + 0.011*"scene" + 0.011*"scenery" + 0.010 *"chinese" + 0.010*"far" + 0.010*"think" + 0.010*"surroundings" + 0.009*"mask" + 0.009*"top" + 0.009*"nature" + 0.009*"feel" + 0.009*" visit" + 0.008*"most" + 0.008*"family" + 0.008*"wind"
Tourism Activity A (Topic 2)	0.037*"peak" + 0.019*"wind" + 0.019*"stairs" + 0.018*"travel" + 0.017*"sea" + 0.016*"sunrise" + 0.015*"view" + 0.015*"scenery" 0.012* "people" + 0.012*"recommended" + 0.011*"landscape" + 0.011*"visit" + 0.010*"admission" + 0.010*"chinese" + 0.009*"tourist" + 0.009*"parking lot" + 0.009*"middle" + 0.009*"world" + 0.009*"here" + 0.009*"crater"
Tourism Activity B (Topic 3)	0.023*"tourist" + 0.021*"scenery" + 0.018*"sunrise" + 0.017*"peak" + 0.016*"china" + 0.015*"visit" + 0.015*"course" + 0.015*"chinese" + 0.013* "sea" + 0.011*"wind" + 0.011*"view" + 0.011*"previous" + 0.010*"child" + 0.010*"round trip" + 0.009*"photo" + 0.008*"nowadays" + 0.008*"scene" + 0.007*"weather" + 0.007*"up" + 0.007*"walk"
Personal Experience B (Topic 4)	0.035*"sunrise" + 0.032*"peak" + 0.022*"think" + 0.017*"time" + 0.017*"visit" + 0.016*"people" + 0.013*"photos" + 0.012*"stairs" + 0.012* "sunset" + 0.010*"course" + 0.010*"shape" + 0.009*"morning" + 0.009*"wind" + 0.009*"entry" + 0.009*"arrival" + 0.008*"soon" + 0.008*"near" + 0.008*"scenery" + 0.007*"weather" + 0.007*"admission"

<Table 4> Shows the results of regression analysis by setting topics extracted from the Seongsan Ilchulbong Review as independent variables and rating as dependent variables. The topics extracted by setting the number of topics to 2 were 'Tourism Activity' and 'Experience Activity' did not significantly affect 'rating'. The topics extracted by setting the number of topics to 3 such as 'Tourism Activity A', 'Tourism Activity B', and 'Family Experience' had a significant effect on 'rating' at the level of $p<0.01$. As the regression coefficient value is negative, the specific weight of each topic negatively affects the rating. In other words, if three topics are extracted from the Seongsan Ilchulbong Review, it can be seen that reviews with a high proportion of topics related to 'Tourism Activities' or 'Family Experience' among online reviews include negative content such as complaints and dissatisfaction. The topics extracted by setting the number of topics to 4 such as 'Personal Experience A', 'Tourism Activity A', 'Tourism Activity B', and 'Personal Experience B' had a significant effect on 'rating' at the level of$p<0.05$. As the regression coefficient value is positive, it can be seen that the proportion of each topic in the review positively affects the rating.

<Table 4> Regression Analysis Results Using the Dependent Variable as Rating

Number of Topics	Independent Variable (Topic Name)	Regression Coefficient	Standard Error (SE)	t	Significance Probability	Adj R^2	F
3	(Constant Number)	5.1989	0.185	28.092	0.000	0.017	6.619***
	Tourism Activity A	-0.1060	0.026	-4.065	0.000***		
	Tourism Activity B	-0.0775	0.027	-2.849	0.004***		
	Family Experience	-0.0925	0.026	-3.512	0.000***		
4	(Constant Number)	3.9926	0.229	17.426	0.000	0.005	2.251*
	Personal Experience A	0.0639	0.029	2.208	0.027**		
	Tourism Activity A	0.0738	0.030	2.500	0.013**		
	Tourism Activity B	0.0473	0.031	1.547	0.122		
	Personal Experience B	0.0710	0.028	2.546	0.011**		

*:p<0.1, **: p<0.05, ***:p<0.01

Depending on the number of extracted topics, the possibility of positive or negative changes in the direction of topics such as tourism activities affecting the rating can be interpreted as a result of low differentiation between topics and poor accuracy of the given topic name. In addition to the fact that the differentiation for each topic is not high, it is necessary to interpret it in consideration of the fact that the R2 and F values of the regression model are low overall. As a result, if the semantic clarity of the topics derived from topic modeling is not high, the accuracy of the regression analysis results may not be high. Nevertheless, it is meaningful in that it confirmed the possibility of deriving a causal model and setting an inductive hypothesis through an extended analysis of topic modeling on the premise that the accuracy of topic modeling increases due to technological development.

IV. Conclusion

Existing social science research is a deductive method, and it was common to derive research hypotheses deductively through literature review and perform hypothesis verification through surveys. Social science research by deductive methods may be robust in terms of theoretical rigor and consistency of established research hypotheses, but there may be limitations in deriving innovative and diverse hypotheses. In this research, the topic modeling technique was expanded to enable inductive hypothesis establishment, and an analysis case of Seongsan Ilchulbong's online review based on the proposed technique was presented.

In the case analysis of this paper, the differentiation of the derived topics was not high, so the model suitability of the derived regression model was also not high. Nevertheless, it is meaningful in that it confirmed the possibility of deriving a causal model and setting an inductive hypothesis through an extended analysis of topic modeling on the premise that the accuracy of topic modeling increases due to continuous technological development.

The expansion method of topic modeling proposed in this research can be used in various ways. It can be used not only for online reviews but also for Internet search results and analysis of news articles. Depending on the text to be analyzed, various external variables other than the text can exist, and various causal models between the topic list and the surrounding variables can be constructed. Based on the causal model, a hypothesis to explain new social phenomena can be established and verified. The adopted hypothesis can be created with new social science

knowledge. The new text analysis method developed in this study will serve as an opportunity to increase the value of online reviews and text data generated in various fields.

References

1. Hofmann T., “Probabilistic Latent Semantic Analysis”, Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1999, pp.289-296.
2. Blei, D., A. Ng., and M. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol.3, 2003, pp.993-1022
3. Kim Geun-hyung, Expansion of Opinion Mining by Entity Association Model, Journal of the Korea Information Processing Society, Vol. 18-D, No. 4, 2011, pp.237-244.
4. Shin Seo-young, Lee Beom-joon, "Consumer Perception Analysis of Eating Out due to the Spread of COVID-19: Using Topic Modeling and Network Analysis," Hotel Management Research, Vol. 30, No. 8, 2021, pp.790.
5. Park Jin-hee, Jeon Mi-sun, Bae Sun-hyung, and Kim Hee-joon, "Research Trends on Factors Influencing the Quality of Life of Cancer Survivors: Text Network Analysis and Topic Modeling," Central Nursing Research, Vol. 21, No. 4, 2021, pp.231-240.
6. Shin Joo-ha, Lim Hee-jin, and Lee Byeong-joo, "Topic Modeling and Emotional Analysis of Service Quality of Domestic Complex Resort," Tourism Research Journal, Vol. 35, No. 11, 2021, pp.191-206.
7. Yoo Jae-ho, Jo Yeon-hee, Jeon Je-chan, "Analyzing Global Green New Deal Research Trends by Topic Modeling Analysis," Journal of Climate Change Chemistry, Vol. 12, No. 4, 2021, pp.289-298.
8. Hwang Won-joon, "Artificial Intelligence and Human Security: An Analysis of the Security of Artificial Intelligence in Europe Using Topic Modeling Techniques," Journal of the 21st Century Political Society, No. 31, No. 2, 2021, pp.55-82.
9. Chae Ho-geun, Lee Ki-hyun, and Lee Ju-yeon, "Analyzing domestic and foreign financial security research trends using topic modeling analysis techniques," Journal of the Korean Industrial Information Society, Vol. 26, No. 1, 2021, pp.83-95.

Analysis on Trends of Fall Accidents at Small-Scale Construction Sites in South Korea

Seung-Hyeon Shin[1], Hyeon-Ji Jung[2], Minjun Kim[2] and Jeong-Hun Won[3]

[1] Department of Big Data, Chungbuk National University, Cheongju, Republic of Korea
[2] Department of Disaster management, Chungbuk National University, Cheongju, Republic of Korea
[3] Department of Safety Engineering & Department of Big Data, Chungbuk National University, Cheongju, Republic of Korea, jhwon@chungbuk.ac.kr

Abstract. This study analyzed the trends in major fall accidents occurring at small-scale construction sites in South Korea. In 2020, small-scale construction sites accounted for more than 70% of work-related fatalities at all construction sites, of which 34.9% were caused by falls. Measures to prevent fall accidents at small-scale construction sites should be implemented to substantially reduce fatalities at construction sites. In this study, the trend of fall accidents at small-scale construction sites was analyzed using LDA topic modeling to suggest a government direction for preparing a safety management policy. The results show that fall accidents mainly occur during painting work outside apartments, piping work, factory roof panel work, etc. Therefore, the government should focus on monioring sites derived from topics or preparing a safety education program for site managers.

Keywords: Fall accidents, small-scale construction site, LDA topic modeling

1 Introduction

The construction industry is one of the most dangerous [1]. Construction workers account for approximately 7% of the total number of workers in all industries, but the number of fatalities accounts for a much higher proportion, including South Korea (hereafter Korea) [2]. In the last 10 years, the work-related fatality rate in the construction industry has been higher than that in the manufacturing industry in Korea.

As a result of analyzing the number of workplaces and work-related fatality statistics by construction cost in 2020, the number of small-scale construction sites (hereafter small sites) with a construction amount of less than 5 billion won was 301,271, accounting for 91.5% of the total construction sites (329,279), and the proportion of fatalities was approximately 72.3% (331 out of 458 people) [3]. In other words, small-scale construction sites are areas where industrial accidents occur intensively within the construction industry. In particular, the work-related fatality rate at small sites is about 4.43 times higher than that of construction sites worth more than 12 billion won, where dedicated safety managers are mandatory under the

Occupational Safety and Health Act, and 34.9% of them were found to have been caused by falls. The Korean government has implemented several policies to prevent industrial accidents at small sites. However, the effect of preventing industrial accidents has been insufficient because accident prevention policies are the same regardless of the construction cost.

To fundamentally reduce fatalities at small sites, the government should prepare safety measures suitable for scale; in particular, it is necessary to intensively manage falls. Therefore, this study analyzes the trend of fall accidents at small-scale construction sites and suggests a direction for safety management policies that the government should prepare.

2 Method

2.1 Data collection

The industrial accident cases in Korean reported to the Korea Occupational Safety and Health Agency (KOSHA) were analyzed to determine the trend of accidents at small-scale construction sites. Among the industrial accident data provided, all information regarding personal or site identification was removed from KOSHA in advance. The data provided were 983 cases of work-related fatalities at construction sites with a cost of less than 5 billion won from 2017 to 2021, and only accidents classified as falls were derived and analyzed in this study. Finally, in this study, 586 accident cases were analyzed: 140 in 2017, 143 in 2018, 114 in 2019, 113 in 2020, and 76 in 2021.

2.2 Data preprocessing

The collected raw accident data consisted of sentences and paragraphs composed of text and consists of facts and opinions investigated by the accident investigator. In addition, information such as the location of the site, personal information of the victims, and the hospitals where the victims were taken were presented. This means that the content of each sentence cannot be used for analysis. Thus, in the initial stage, to derive the analysis results, the step of extracting them into individual words that could be analyzed was carried out. The NetMiner 4.5 program was used for the preprocessing step.

Among the first extracted words, stopwords that frequently appear in general sentences, including conjunctions, prepositions, adverbs, and surveys, and keywords unrelated to the trend of accidents such as "death," "hospital," "accidents," "worker," and "regional names" were removed. In addition, keywords with a TF-IDF Threshold value of 0.5 or less and keywords with the number of letters of less than three characters were removed.

2.3 Topic modeling

Latent Dirichlet Allocation (LDA)-based topic modeling and social network analysis were conducted to discover specific subject domains using accident keywords that completed data preprocessing. The NetMiner 4.5 program was used for analysis. LDA-based topic modeling is a technique that determines the degree to which several keywords appear simultaneously in each accident in other accidents and determines the keyword as a topic if it is determined that the keyword group represents the representation of the accident [4]. Thus, it is determined that this technique is suitable for deriving accident trends. Social 2-mode social network analysis was conducted to derive the relationships between keywords for the derived topics.

3 Results

The optimal number of topics was determined using perplexity and coherence scores for falls at small sites. The optimal number of topics was five; the main keywords are listed in Table 1. The results of the network analysis of these topics are shown in Figure 1.

Topic 1 and Topic 4 are falls that occur at building construction sites such as apartments and multi-family houses, and Topic 1 are consist of keywords related to painting work using hanging scaffolding or waterproofing work such as "apartment", "paint", "outer wall", "rope", "rooftop", and "waterproof". Topic 4 is an accident that falls into the opening or scaffold when disassembling the form, and the keywords are "scaffolding", "work plate", "building", "housing", and "external".

The Keywords for topic 2 are "mobile elevated work platform (MEWP)", "subcontractor", "piping", and "boarding". Topic 2 are falls that occurred during piping construction using the MEWP at subcontractors.

Topics 3 and 5 indicate falls that occur at factory construction sites. Topic 3 keywords are "steel frame", "factory", "panel", "roof", "crane", and "bolt", which consist of falls that occur during roof panel construction using cranes. Topic 5 is a fall accident that occurs when working to install or remove facilities such as solar panels on the roof of a factory, and the main keywords were "roof", "factory", "facility", "ladder", "stair", "removal", and "solar".

Table 1. Topic modelling result of falls on small-scale construction sites

Number	Domain	Keywords
Topic 1	Wall painting or waterproofing work of apartment buildings	Apartment, painting, exterior wall, rope, rooftop, maintenance, waterproof, replacement
Topic 2	Piping work using MEWP	MEWP, subcontractor, piping, boarding
Topic 3	Roof panel installation of factory	Steel frame, factory, panel, roof, crane, bolt
Topic 4	Disassembling the form of building	Scaffold, Work plate, building, housing, external
Topic 5	Facilities installation or removal of factory	Roof, factory, facility, ladder, stair, removal, solar

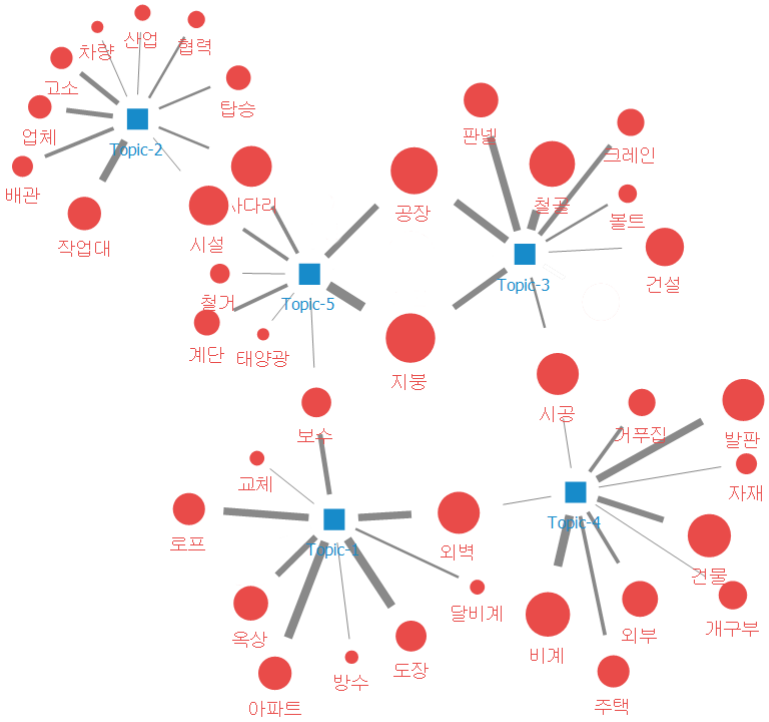


Fig. 1. Topic network of falls on small-scale construction site

4 Conclusion

In this study, the trend of work-related fatality accidents caused by falls at small-scale construction sites in Korea was analyzed. Topic modeling and network analysis were conducted to determine the accident trends.

It was found that fall accidents at small sites mainly occurred during housing and factory construction. It was found that several accidents occurred during housing construction, where ropes were cut in the process of wall painting of apartments or waterproofing works using a hanging scaffold. Wall painting or waterproofing works are performed during the construction completion stage or maintenance stage of the building, and these are performed in the absence of a manager. Life ropes must be installed when working with hanging scaffolding and must be fastened to two or more fixed points to prevent accidents during these tasks. However, these safety rules were not observed owing to the absence of a manager. In addition, there were still cases of arbitrary misuse, such as connecting work scaffolding with wires when disassembling the forms.

It was found that falls occurred mainly when working on roofs during the factory construction. As a result of analyzing the detailed causes of the accident cases, it is mainly caused by the lack of the most basic safety measures, such as wearing personal protective equipment (PPE) such as safety belts and installing safety handrails; the fundamental reason is that both managers and workers ignore safety rules to finish the work quickly.

Falls at small-scale construction sites are caused by the absence of managers or by indifference to the safety of workers and managers. Therefore, the government needs to intensively supervise the sites, and it is necessary to create and operate an education program for managers to comply with the minimum safety rules to prevent fundamental falls at small construction sites.

Acknowledgments. This study was supported by the Academic Service Contract of the Occupational Safety and Health Research Institute in 2022.

References

1. Pinto, A., Nunes, I.L., Ribeiro, R.A.: Occupational risk assessment in construction industry – Overview and reflection. Saf. Sci. vol. 49, no. 5, pp. 616--624. (2011)
2. Sunindijo, R.Y., Zou, P.X.W.: Political skill for developing construction safety climate. J. Constr. Eng. Manag. vol. 138, no. 5, pp. 605--612. (2019)
3. Ministry of Employment and Labor: Analysis of industrial accident status. (2022)
4. Blei, D. M., Lafferty, J. D.: Topic models, In Srivastava, A N., & Sahami, M. (eds.). Text Mining: Classification, Clustering, and Application, pp. 71--94. Boca Raton: CRC Press. (2009)

A Study on Clustering Analysis of Judgment

Eun-Young Park[1], and Sun-Young Ihm[2,*]

[1] Dept. of Visual Design, Hyupsung University, Korea, pey54@naver.com
[2] Dept. of Computer Engineering, Pai Chai University, Korea, sunnyihm@pcu.ac.kr
[*] Corresponding Author

Abstract. In this paper, we collected judgment sentences and extracted nouns to analyze criminal data. Next, we vectorized the case, performed clustering analysis, and derived features through grouping of cases.

Keywords: Bigdata, Clustering, Judgment

1 Introduction

Recently, as interest in big data increases, attempts to utilize big data in various fields of society are increasing, and this trend is expected to accelerate due to the development and increase of big data analysis technology. Such efforts and attempts to analyze social phenomena and solve related social problems using big data are also being witnessed in the fields of police science and criminology. The use of crime big data can guarantee effectiveness as well as improvement of efficiency in existing police activities and crime prevention and investigation, so the need for research is sufficient [1]. In this paper, we performed a similarity analysis after collecting judgments that do not contain information that can identify individuals and contain specific details about the case.

2 Analysis of similarity of judgment data

In this study, we collected judgment data for intrusion and related text analysis. The data of the 2019 judgment was set as the target of collection, and we searched and collected the sentences of cases where judgments were pronounced in 2019 as a search term for 'intrusion' in the Korean court judgment search system. We selected 364 cases of collected judgments as research subjects and proceeded with analysis.

We first performed preprocessing on the collected data for analysis. To do this, we first divide the sentence data into words based on spaces. Next, morphemes are analyzed using KoNLPy [2], a Python package for Korean information processing. Morphological analysis is to understand the structure of various linguistic properties such as morphemes, roots, prefixes/suffixes, parts of speech, etc. In this study, nouns are extracted from judgment data to analyze the similarity between cases. Through the

Twitter object of KoNLPy, only nouns with 'Noun' tag are separated and extracted from the text. Fig. 1 shows case information from which only nouns are extracted.

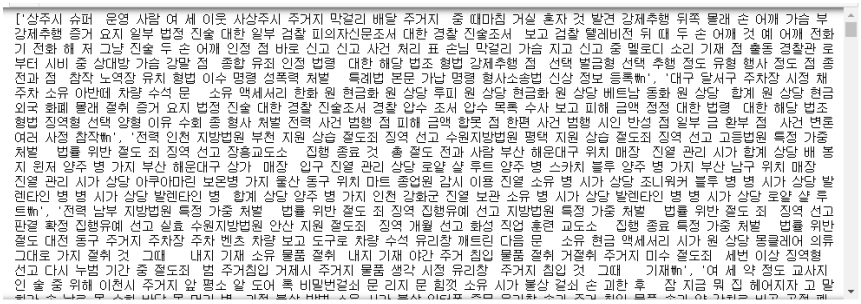


Fig. 1. A noun extraction result of Judgment.

For text document clustering, vectorization of each document is first required [3, 4]. Therefore, we vectorize each event so that it can be analyzed based on the frequency of the number of words. Next, similar cases were clustered through clustering analysis. Clustering analysis is a method of grouping data without a classification class among similar data. We used the kmeans clustering algorithm, and Fig. 2 shows the results of clustering.

	class	data
0	0	상주시 슈퍼 운영 사람 여 세 이웃 사상주시 주거지 막걸리 배달 주거지 중 때마...
1	0	대구 달서구 주차장 시정 채 주차 소유 아반떼 차량 수석 문 소유 액세서리 한화...
2	1	전력 인천 지방법원 부천 지원 상습 절도죄 징역 선고 수원지방법원 평택 지원 상습 ...
3	2	전력 남부 지방법원 특정 가중 처벌 법률 위반 절도 죄 징역 집행유예 선고 지방...
4	0	여 세 약 정도 교사지인 술 중 위해 이천시 주거지 앞 평소 알 도어 록 비밀번걸쇠...
...	...	...
360	0	인터넷 사이트 진짜 것 분 연락 글 연락 절도 범행 친분 범행 가담 것 권유 명 ...
361	0	대전 동구 층 운영 의류 점 매장 판매 소유 시가 상당 의류 착용 그대로 물품 절취...
362	2	전력 대전 지방법원 산지 야간 건물 침입 절도죄 징역 집행유예 선고 유예 기간 ...
363	1	창원시 의창구 앞 노상 주차 그랜저 차량 시정 운전 석 문 차량 내부 운전 석 옆 ...
364	1	야간 건물 침입 절도 서울 동작구 빌딩 층 소재 운영 구장 이전 종업원 일 미리 소...

Fig. 2. A clustering result.

As a result of the analysis, it was confirmed that the class 0 group showed the characteristics of habitual theft, the class 1 group showed the characteristics of special theft, the class 3 showed the characteristics of shop theft, and the class 4 showed the characteristics of the vehicle theft. Class 2 was identified as a theft offense that did not fall under the other classes.

3 Conclusion

In this paper, we collect judgment data and perform clustering analysis on intrusion cases. If the analyzed results are used, the identification of criminals and means of crime can be facilitated even in new cases, so we expect that the investigation will be conducted effectively and efficiently.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2021R1C1C2011105).

References

1. Kwon, Y.: Study on the Application and Legal Limits of Big Data for Crime Prevention and Investigation. Law Preview. 17, 179-198 (2017)
2. Park, E.L., Cho, S.: KoNLPy: Korean natural language processing in Python. Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology, pp.133-136. (2014)
3. Agudelo, G., Parra, O., Velandia, J.B.: Raising a Model for Fake News Detection Using Machine Learning in Python. Proceedings of the 17th Conference on e-Business, e-Services and e-Society (I3E), pp.596-604 (2018)
4. Singh, A.K., Shashi, M.: International Journal of Advanced Computer Science and Applications. 10, 305-310 (2019)

Measurement of Center Point Deviation for Detecting Contact Lens Defects

Ginam-Kim, Sung Hoon-Kim, In-Joo, Kwan Hee-Yoo

Dept. of Computer Science, Chungbuk National University
1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, Republic of Korea
{kgn4192, sidsid84, jooin95, khyoo}@chungbuk.ac.kr

Abstract. In this study, we investigate the distance measurement deviated center to detect contact lens defects. To this end, after detecting a circular region of interest (ROI), the deviation measurement from the center point of the original image is measured. When only the Hough Circle Transform algorithm was used for ROI detection, the results of the Gaussian Blur and Canny Edge Detection algorithms were compared and analyzed. It was observed that the method using all three algorithms mentioned above is able to detect more accurate ROI.

Keywords: Contact lens, Hough Circle Transform, Gaussian Blur, Canny Edge Detection

1 Introduction

Product defect detection is essential for quality control in manufacturing industry. The occurrence of defects leads to a huge waste of resources, increases the cost of the enterprise, and causes considerable harm to human life and safety. Representative examples of defect detection studies include automatic thresholding for defect detection by Ng [2], a review of pavement defect detection methods by Cao et al. [3], and fabric defect detection by Chan et al. [4]. There is a by Fourier analysis. Defects can also occur during the manufacturing of contact lenses. The contact lens injection process in this study uses the sandwich method [5][6] to color the dye between the lens layers. In this process, defects such as the center defect due to detachment of the center point of dye coloring, the Colorpoor defect in which coloring cannot be properly performed, the Inkcut defect, and the Line defect due to scratches may occur. To date, research on contact lenses has focused on contact lens detection and wearability classification such as the study by Raghavendra et al. [7], contact lens iris recognition such as the study by Choudhary et al. [8], and wearable contact lens type classification such as studies by Doyle et al. [9] and Kimura et al. Researchers have focused more on classification and detection with the lens on the iris, such as iris recognition by hyperparameter tuning, as reported in [10]; however, computer vision-based contact lens defect detection study has not yet been conducted. To date, the inspection of contact lenses for defects has relied mostly on manual labor. To avoid time-consuming and labor-intensive manual

inspection and improve resource efficiency and worker and consumer safety, this study focuses on the central point deviation among the various defects that can occur in the contact lens injection process. We attempt to remedy this through fault detection using distance measurements. This study uses the Hough circle detection [2] algorithm for this purpose; however, if the original image is used as is, the change in the gradient direction of the edge pixels according to the image causes errors in the common parameters. There is a possibility. Therefore, after minimizing the circle detection error through preprocessing, the deviated center-point distance of the contact lens is measured.

2 Background

2.1 Hough Circle Transform

The Hough Circle Transform [11] finds a circle if a circle in the image is described as

$$(x - a)^2 = r^2 \quad \textbf{(1)}$$

where *(a, b)* are the coordinates of the circle center and r is its radius. As shown in Figure 1, we use the values of the channels of the image in 2D space to determine the position of the center point in the accumulation array of slope normals that differentiate the points with respect to the coordinate system based on the threshold. Afterwards, to return the radius in the transformed coordinate system, an algorithm is used to detect the position. The Hough Circle Transform has built-in Canny Edge Detection.

2.2 Canny Edge Detection

Canny Edge Detection [12] applies the Sobel kernel horizontally and vertically to determine the gradient in each direction. When the horizontal gradient is defined as G_x and the vertical gradient is defined as G_y, the edge gradient detects the slope and direction of the edge, as shown in Figure 2 in the image coordinates.

$$Edge\_Gradient\ (G) = \sqrt{(G^2_x + G^{2)}} \quad \textbf{(2)}$$

$$Angle\ (\theta) = tan^{-1}(G_y / G_x)$$

2.3 Gaussian Filter

A Gaussian distribution is a normal distribution with a symmetrical probability distribution centered at the mean, and a Gaussian filter [13] is a filtering method that uses a specific-size filter mask generated by approximating the Gaussian distribution function to blur the image. In other words, it reduces the noise of the image such that it does not react sensitively during image processing. The following formula is used for the Gaussian distribution function in two-dimensional space, where x is the distance from the origin of the horizontal axis, y is the distance from the origin of the vertical axis, and σ is the standard deviation of the Gaussian distribution.

$$G(x, y) = (1/(2\pi\sigma 2))e^{((-x^2+y^2)/2\sigma^2)} \quad (3)$$

3 Dataset

The image for obtaining the deviated center point distance is the same as that in Figure 3 and has a size of 800×800×24. In this study, 103 pieces of data were used to measure the deviated center-point distance.

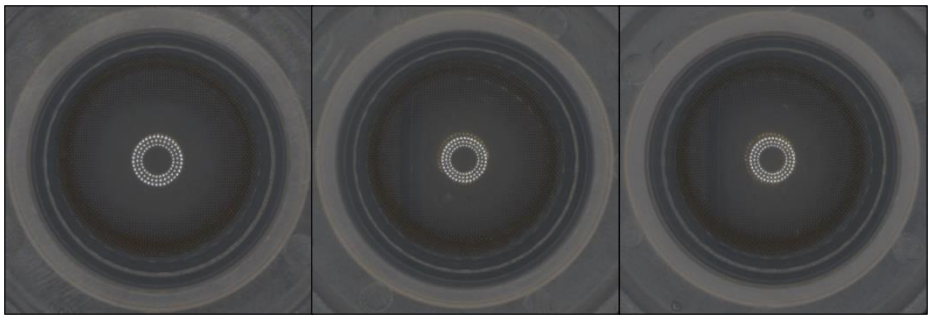


Fig. 3. Sample image of dataset

4 Proposed Method

In this study, to measure the deviated center point distance, after converting the RGB channel image to grayscale, we first apply Gaussian Blur to reduce the noise of the image and make it less sensitive. The canny edge detection algorithm is then used to convert the gradient to the same gradient threshold to be measurable. Finally, after the printed lens area is detected using the Hough circle transform algorithm through the above preprocessing, the deviated center point distance is measured using the following

formula, where x and y are the coordinates of the center point of the image, and width and height are the coordinates of the center point of the deviated center.

$$Error\ Distance(x, y) = \sqrt{((x - width)^2 + (y - height)^2)} \quad \textbf{(4)}$$

Figure 4 illustrates the step-by-step process mentioned above.

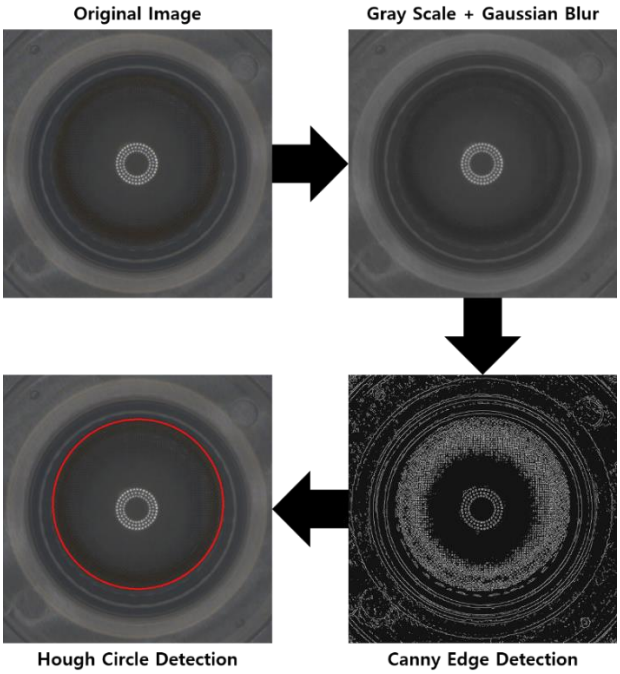


Fig. 4. Step-by-step image processing result and ROI detection process with the proposed method

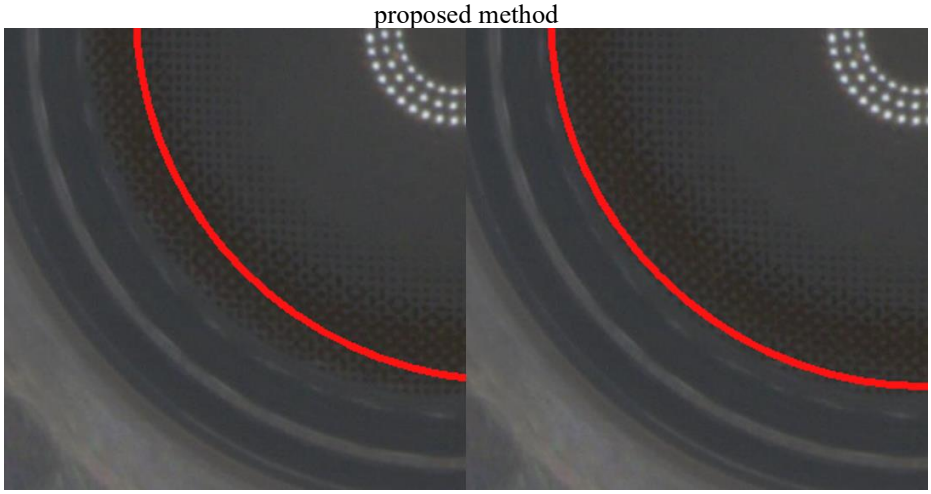


Fig. 4. Left) Result when using only the Hough circle transform, right) Result when using the proposed method

5 CONCLUSION

In this study, we used the Gaussian Blur, Canny Edge Detection, and Hough Circle Detection algorithms to measure the distance from the center point of the contact lens. H. The C algorithm is basically C. E algorithm is incorporated, but C.E.
In future studies, the authors will attempt to study the measurement of parameter-independent contact lens deviated center point distance through machine learning after carrying out labeling based on the algorithm used in this study.

Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2022-2020-0-01462) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation) and by the Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (No. P0022332, Digital data platform for material development)

References

1. Wang, T., Chen, Y., Qiao, M., Snoussi, H. : A fast and robust convolutional neural network-based defect detection model in product quality control : IJAMT, vol. 94, pp. 3465–3471.(2018)
2. Hui-Fuang Ng : Automatic thresholding for defect detection : ICIG, pp. 532-535 (2004)
3. Wenming Cao, Qifan Liu, Zhiquan He : Review of Pavement Defect Detection Methods, IEEE Access, vol.8, pp.14531-14544 (2020)
4. Chi-Ho Chan, G.K.H. Pang : Fabric defect detection by Fourier analysis : IEEE, Transactions on Industry Applications, vol.36, pp.1267-1276 (2000)
5. 김명환 : A PROCESS OF SANDWICH FOR COLOR COATING CONTACT LENSES : Publication No. KR100647133B1
6. 김명환: Coating method for cosmetic color contact lenses : Publication No. WO2011019100A1
7. R. Raghavendra, Kiran B. Raja, Christoph Busch, : ContlensNet: Robust Iris Contact Lens Detection Using Deep Convolution Neural Networks : WACV, pp.1160-1167 (2017)
8. Meenakshi Choudhary, Vivek Tiwari, Venkanna U. : An approach for iris contact lens detection and classification using ensemble of customized DenseNet and SVM : Future Generation Computer Systems, Vol.101, pp.1259-1270 (2019)

9. James S. Doyle, Patrick J. Flynn, Kevin W. Bowyer : Automated classification of contact lens type in iris images : ICB (2013)
10. Gabriela Y. Kimura, Diego R. Lucio, Alceu S. Britto Jr., David Menotti : CNN hyperparameter tuning applied to iris liveness detection : VISAPP, Vol.5, pp428-434 (2020)
11. Duda R.O. and Hart P.E. : Use of the Hough Transformation to detect lines and curves in pictures : Comm. of the ACM, vol 15, no. 1, pp. 11-15 (1972)
12. John Canny : A Computational Approach to Edge Detection : IEEE Transactions on Pattern Analysis and Machine Intelligence : vol.PAMI-8, pp.679-698 (1986)
13. IT Young, LJ Van Vliet : An adaptive Gaussian filter for noise reduction and edge detection : IEEE Nuclear Science Symposium and Medical Imaging Conference, vol.3, pp. 1615-1619 (1993)

Artificial Intelligence Techniques in Mental Healthcare: A Systematic Mapping Study

Ngumimi Karen Iyortsuun[1] , Soo-Hyung Kim[2],
Hyung-Jeong Yang[3], and Aera Kim[4],

[1-4] Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea
{kareniyortsuun@gmail.com},{ shkim, hjyang, arkim}@jnu.ac.kr

Abstract. Mental illnesses have been a menace to our societies today, and the means of confronting these illnesses is of utmost importance. To combat this menace, Machine learning researchers have been developing various models for the early diagnosis of mental disorders over the years. In this study, we conducted a systematic mapping study to show the most prevailing trends in the application of Artificial Intelligence techniques, including Machine Learning and Deep Learning, for the diagnoses of various mental disorders such as Schizophrenia, Anxiety, Bipolar disorder, Depression, Post Traumatic Stress Disorder (PTSD), Anorexia Nervosa, and attention deficit hyperactivity disorder (ADHD). In addition, we also briefly talk about the data availability challenges faced by researchers in this area of research.

Keywords: Mental health diagnosis, Healthcare, Machine learning, Deep learning, Data.

1 Introduction

Mental disorders are generally a significant disruption in a person's thinking faculty, emotional stability, or physical well-being. Mental disorders are pervasive, with over 970 million people suffering from one form of mental disorder worldwide [1]. There are various categories of mental disorders, including Psychotic disorders such as Schizophrenia, Anxiety disorders, Bipolar disorder, Depression, Post Traumatic Stress Disorder (PTSD), eating disorders such as Anorexia Nervosa, and attention deficit hyperactivity disorder (ADHD). Due to the high risks and negative societal implications of these disorders, such as suicidal ideations and suicide, the availability of early measures to combat these disorders is of keen importance.

Machine Learning, a type of Artificial intelligence (AI), has been introduced to assist mental healthcare providers in diagnosing and decision-making to curb these disorders. Also, with the introduction of Deep learning (DL) [2], outstanding performance has been seen in various data-rich application scenarios, including healthcare [3]. Although applying these algorithms opens up numerous opportunities, using self-reported data from subjects brings about legal and ethical concerns regarding data anonymization.

In this study, we attempt to answer two fundamental questions; 1. What are the recent approaches used by ML researchers in the diagnosis of mental disorders? 2. What is the situation of data availability in this research area?

2 System Study Mapping

We carried out a systematic mapping study, as illustrated in Fig. 1. The purpose was to identify studies that show a trend in developing mental health diagnoses of the most prevalent mental illnesses in our societies today. These articles were then classified and reviewed based on seven different mental disorders.

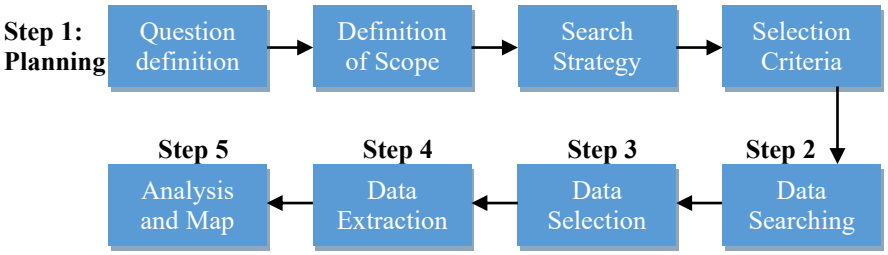


Fig. 1. Systematic Mapping Study workflow

3 Result

Depression and Anxiety, according to the world health organization (WHO), are the two most disturbing mental health disorders plaguing the world today [1]. According to a 2021 study, the rate of depression in South Korea increased from about 645,607 in 2010 to over 1 million in 2020, with a majority of these patients in their twenties [4]. This spike in numbers could be due to the outbreak of COVID-19 in 2019.

Bipolar is another type of mental disorder that creates an unusually abrupt shift in mood, whereas Anorexia nervosa impacts a person's self-image. It produces a preoccupation with weight loss and an excessive dread of gaining weight, leading to extreme actions that substantially negatively affect their lives. ADHD, on the other hand, is a neurodevelopmental disorder mainly affecting children and adolescents, but studies reveal that it can also impact adults [5]. Finally, PTSD is typically the result of a past traumatic experience. It is characterized by irrepressible thoughts about such occurrences. Unfortunately, most mental disorders cannot be cured but may be controlled primarily via psychotherapy and medication [6].

Unlike other medical conditions, there is no specific test for mental disorders. Therefore, professionals use various methods such as physical tests, clinical checklists of symptoms for Schizophrenia, Patient Health Questionnaires (PHQ) [7] for Depression, and psychological evaluations with the Diagnosis and Statistical Manual of Mental Disorders (DSM-5) handbook.

3.1 Question 1

To answer the first fundamental question of this study, researchers have recently been looking more into the application of DL algorithms due to their ability to transform data through layers of nonlinear units providing new practical knowledge from complex data. Some of the trending DL algorithms include Deep Feedforward Neural Networks (DFNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Autoencoders. These methods have shown superior results when implemented in healthcare applications. Li et al. proposed using a CNN to diagnose mental disorders automatically, and their research yielded excellent results [8]. This is not to dispute the fact that traditional Machine learning methods have also significantly contributed to this area of research.

Support Vector Machines (SVM) and Random Forests (RF) have proven to be the most applied ML methods showing encouraging results. Jo et al. [9] and Schultebraucks et al. [10] implemented these methods in their study. Incorporating both ML and DL approaches has also shown promising results, as demonstrated by Ranganathan et al. [11].

Table 1. Summary of [8-11]

Model	Precision	Recall	F1-Score	Accuracy	AUC	ERDE-50
CNN [8]	99.76%	99.74%	99.75%	99.72%	99.75%	-
i. Global RF	-	-	-	68.0%	68.0%	-
ii. XGBoost [9]				66.3%	65.6%	
i. RF	-	-	-	-	78.0%	-
ii. SVM [10]					88.0%	
i. Neural Machine Translator	0.48	0.26	0.34	-	-	0.007
ii. SVM Classifier with SGD optimization using TF-IDF [11]						

Convolutional Neural Networks(CNN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), Area Under the ROC Curve (AUC), Early Risk Detection Error (ERDE)

3.2 Question 2

One of the most challenging aspects of mental health diagnosis with AI techniques has been the availability of large multimodal datasets. This is due to the high cost of data sourcing, as it usually involves human participants; hence, using small datasets in this area is common. This challenge is characterized by data volume, velocity, variety, variability, and veracity [12]. DL models require the use of large datasets to allow the scrutinization of parameter space. However, most of the datasets researchers use are not made publicly available. Therefore, researchers turn to online sources like YouTube, Reddit, and Twitter to gather unimodal data such as Textual data. In an

attempt to demonstrate the feasibility of employing biomarkers in anxiety and depression diagnosis, Hilbert et al. [13] met the limitation of data sample size, which might have severely influenced the outcomes of their research.

Data availability is not the sole issue; data quality has also impacted model performance. Preprocessing methods may improve data quality; however, an excessive amount of data preprocessing could result in less data (due to deletion) which could lead to a bias and low-performance accuracy of the model.

Table 2. Datasets for various mental health predictions

Dataset	Data Modality	Application
Audio-visual Depressive Language corpus- AVEC 2013 (AViD Corpus) [14]	Audio/Video	Depression
Distress Analysis Interview Corpus (DAIC) [14]	Audio/Video	Anxiety, Depression, PTSD
Penn-dataset [15]	Video/Images	Schizophrenia
Spanish Anorexia Dataset (SAD) [16]	Text	Anorexia Nervosa
Turkish Audio-visual Bipolar Disorder Corpus [17]	Audio/Video	Bipolar Disorder

4 Conclusion

In this study, two fundamental research questions were considered on implementing Deep learning and Machine learning approaches to diagnose various mental disorders. Overall, our study shows that using DL methods in mental health diagnosis can offer excellent results, such as the implementation of CNN by Li et al. [8], where their approach showed a performance of at least 99.72% on all evaluation metrics used (Table 1). Also, this study shows that ML alone could be a valuable means of understanding these disorders, as proven by Jo et al. [9] and Schultebraucks et al. [10]. Furthermore, accessing multimodal datasets for this purpose is a challenge; therefore, more interest should be laid on developing accessible large multimodal datasets as this could give researchers ample opportunities to try different methods and develop better-performing models applicable to real-world systems.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT). (NRF-2020R1A4A1019191) and also supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub). Corresponding author is Soo-Hyung Kim.

References

1. (WHO), W.H.O. *Mental Disorders*. 2022 8/18/2022]; Available from: https://www.who.int/news-room/fact-sheets/detail/mental-disorders.
2. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
3. Durstewitz, D., G. Koppe, and A. Meyer-Lindenberg, *Deep neural networks in psychiatry.* Molecular psychiatry, 2019. **24**(11): p. 1583-1598.
4. So-min, K. *Over 1 million depressed patients... Most are in their 20s*. 2021 [cited 2022 27 October]; Available from: https://www.donga.com/news/Society/article/all/20210406/106260124/1.
5. Hansa D. Bhargava, M. *Attention Deficit Hyperactivity Disorder in Adults*. 2021 8/26/2022]; Available from: https://www.webmd.com/add-adhd/adhd-adults.
6. America, M.H. *Mental health treatments*. 2022 [cited 2022 27 October]; Available from: https://mhanational.org/mental-health-treatments.
7. Costantini, L., et al., *Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): A systematic review.* Journal of affective disorders, 2021. **279**: p. 473-483.
8. Li, Z., et al., *Deep learning based automatic diagnosis of first-episode psychosis, bipolar disorder and healthy controls.* Computerized Medical Imaging and Graphics, 2021. **89**: p. 101882.
9. Jo, Y.T., et al., *Diagnosing schizophrenia with network analysis and a machine learning method.* Int J Methods Psychiatr Res, 2020. **29**(1): p. e1818.
10. Schultebraucks, K., et al., *Pre-deployment risk factors for PTSD in active-duty personnel deployed to Afghanistan: a machine-learning approach for analyzing multivariate predictors.* Molecular psychiatry, 2021. **26**(9): p. 5011-5022.
11. Ranganathan, A., et al. *Early Detection of Anorexia using RNN-LSTM and SVM Classifiers*. in *CLEF (Working Notes)*. 2019.
12. Stewart, R. and K. Davis, *'Big data' in mental health research: current status and emerging possibilities.* Social psychiatry and psychiatric epidemiology, 2016. **51**(8): p. 1055-1072.
13. Hilbert, K., et al., *Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: A*

multimodal machine learning study. Brain and Behavior, 2017. **7**(3): p. e00633.
14. Gratch, J., et al., *The distress analysis interview corpus of human and computer interviews*. 2014, UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.
15. Hamm, J., et al., *Dimensional information-theoretic measurement of facial emotion expressions in schizophrenia.* Schizophrenia research and treatment, 2014. **2014**.
16. Úbeda, P.L., et al. *Detecting anorexia in Spanish tweets*. in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 2019.
17. Çiftçi, E., et al. *The turkish audio-visual bipolar disorder corpus*. in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. 2018. IEEE.

Attention-based Deep Neural Network for Predicting Fetotoxicity

Myeonghyeon Jeong[1], Sangjin Kim[2], Yewon Han[2], Jihyun Jeong[2], Dahwa Jung[2], Inyoung Choi[2], Sunyong Yoo[1,*]

[1] Department of ICT Convergence System Engineering, Chonnam National University, Gwangju, 61186, Republic of Korea
[2] Korean Medicine Informatization Team, NIKOM, Osong Building, 14, Jeongdong-gil, Jung-gu, Seoul, 04516, Republic of Korea

Abstract. Pregnant women sometimes need to take medications, but this can pose potential risks to the fetus and mother. Therefore, it is essential to classify drugs that are harmful to pregnant women or fetuses. However, assessment of drug fetotoxicity is expensive and time consuming. *In silico* approaches can identify compounds that may present a high risk to the fetus for a wide range of drugs and compounds at the low cost and time. In this study, we constructed attention-based deep neural network to predict potential fetotoxicity of drugs. Structural and physicochemical information of each drug was used as input and known fetotoxicity was used as output to train the model. We confirmed that the model gives high attention to specific molecular substructure when predicting fetotoxicity. Our study can be used as a pre-screening tool for predicting fetotoxicity as it provides key molecular substructures related to fetotoxicity of compounds.

Keywords: Virtual screening, Fetotoxicity, Machine learning, Interpretable model

1 Introduction

The fetotoxicity of drugs is toxic effects on a fetus of drugs that crosses the placental barrier. It may compromise maternal health and appear as fetal malformations, altered growth and in utero death. The pregnant women may need medication at any time, so that predicting fetotoxicity of drugs is important. However, predicting drug toxicity through *in vivo* or *in vitro* experiments is very costly, labor intensive and time consuming. Machine learning approaches can be leveraged to overcome these challenges and screening for toxicity to compounds that exist in vast chemical space. In this study, we develop the interpretable fetotoxicity prediction model based on attention-based deep neural network (DNN) for virtual screening fetotoxicity of drugs.

* corresponding author: Sunyong Yoo (syyoo@jnu.ac.kr)

2 Materials and Methods

We curated the data for predicting fetotoxicity models from the TGA (therapeutic goods administration of Austrailian government), KIDS (Korea institute of drug safety and risk management), and COCONUT (compound combination-oriented natural product database with unified terminology) datasets. The data collected from each dataset contains the name of the drug and the level of fetotoxicity risk of the drug. Prediction fetotoxicity model requires structural information of the drug, so we generated and added SMILES (simplified molecular-input line-entry system) information, which for describing the structure of compounds. According to the previously classified level of fetotoxicity risk level in each dataset, we reclassified into "fetotoxic" and "non-fetotoxic" labels in the curated dataset.

We developed the machine learning models such as attention-based DNN, logistic regression, support vector machine for prediction and analysis fetotoxicity, and compared the quantitative performance of various machine learning models. Furthermore, in order to confirm that the developed model is actually interpretable, we drawn the molecular substructures of compounds which highly correlated with fetotoxicity by analyze the attention score.

3 Results

Accuracy, AUROC, precision, recall, and F1 score were used as indicators to compare the quantitative performance of various machine learning models. The attention-based DNN model showed the highest performance indicators in accuracy (= 0.7528), recall (=0.6421), and F1 score (=0.6489), the random forest showed the highest precision (=0.756), and the extra tree showed the highest AUROC (=0.81). Figure 1 shows the three molecular substructures analyzed to be highly correlated with fetotoxicity in compounds predicted as fetotoxicity.

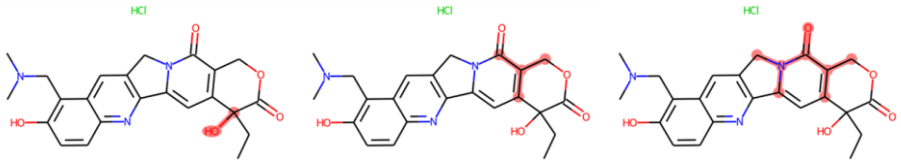


Fig. 1. Molecular substructures analyzed to be highly correlated with fetotoxicity

4 Conclusion

In this study, we confirmed the machine learning model that not only predicts fetotoxicity, but can analyze the molecular substructures highly correlated with fetotoxicity. The developed model will be able to play a key role in the research design phase of *in vitro* or *in vivo* experiments demonstrating the fetotoxicity of drugs.

Web-based Automated Neural Architecture Search Studio

Dong Jin[1,2], Ri Zheng[1,2], HeLin Yin[1],
Yeong Hyeon Gu[3,*], Seong Joon Yoo[1]

[1] Department of Computer Science and Engineering,
[2] Department of Convergence Engineering for Intelligent Drone,
[3] Department of Artificial Intelligence,
Sejong University, Seoul, Korea

{justdong, zrchuangri}@sju.ac.kr, {yhl0608, yhgu, sjyoo}@sejong.ac.k

Abstract. Currently, research related to automated machine learning, such as neural architecture search and hype-parameter optimization, is being actively conducted. However, since most of the existing research is published in the form of code or libraries, it is difficult to use for people who are not familiar with machine learning. To solve this problem, this study proposes a web-based neural architecture search studio that can be easily used by non-expert users who are not familiar with machine learning. With the proposed studio, non-expert users can also create deep learning models that achieve excellent performance on a given dataset.

Keywords: AutoML, neural architecture search, hyper-parameter optimization, user interface

1 Introduction

Automated Machine Learning (AutoML) is a process that automates repetitive machine learning model development tasks that require a lot of human and temporal resources. Recently, various studies related to AutoML have been conducted. However, there is not enough research on AutoML tools that can be easily used by non-expert users unfamiliar with machine learning. In this study, we designed and developed a Web UI-based automated neural architecture search studio that helps non-experts easily develop deep learning models with good performance. The proposed studio connects a user-friendly user interface with a neural architecture search and hyper-parameter optimization algorithm, enabling the training of deep learning models in an end-to-end manner using the web page. The proposed studio may allow non-experts to develop high-performance deep learning models.

* Corresponding author

2 Related works

Various studies on neural architecture search and hyper-parameter optimization have been conducted, and libraries providing various search and optimization methods have been published. Previous studies have provided various automated neural architecture search tools such as Auto-Sklearn [1], H2O [2], AutoKeras [3], and AutoGloun [4]. However, since they are also provided in the form of libraries, there is a problem that it is difficult for people who are not familiar with deep learning and coding to use the research results of automated neural architecture search well. Tools such as Cloud AutoML and SageMaker provide a way to handle the automated neural architecture search process through the Web user interface, but they are difficult to use and not user-friendly.

3 Proposed AutoML Studio

This study proposes a system that performs an automated neural architecture search for a given dataset and optimizes hyper-parameters for the searched neural network. The user interface of the proposed system consists of a model search UI, model retrain pop-up, and hyper-parameter search pop-up. Fig. 1. shows the system architecture of the proposed studio. Model search UI provides a user interface that allows you to search the model architecture through the neural architecture search algorithm and check the performance of each searched model. The model retrain pop-up provides the ability to retrain the model from scratch by selecting one of the model architectures searched through the model search UI. And hyper-parameter optimization can be performed through the hyper-parameter search pop-up window, which provides a function to find the best hyper-parameters for a model architecture found in the model search UI.

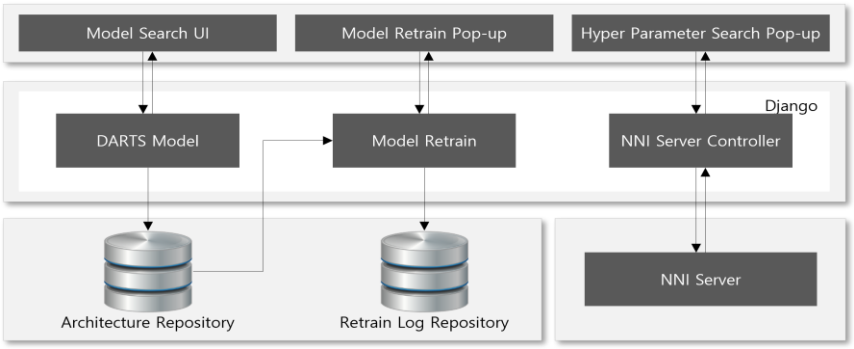


Fig. 1. System architecture

This study used a differentiable architecture search (DARTS) [5] algorithm as the system's neural architecture search algorithm. The DARTS algorithm is a neural

architecture search algorithm that defines a cell identical to the recurrent cell design of ENAS [6]. It enables learning using gradient descent by considering all possible operators that can be applied between nodes within the cell in the form of a mixed operation. Moreover, the neural network structure is explored by repeating bilevel optimization. The system supports random search, grid search, and TPE [7] as a hyper-parameter search method. Hyper-parameter search methods are implemented using the features provided by neural network intelligence (NNI) [8]. The neural architecture search algorithm and hyper-parameter optimization method can be easily used through a user-friendly web-based user interface. This allows even the users without relevant knowledge to develop a model that achieves high performance in an end-to-end manner through the proposed system.

4 Conclusion

This study proposed an AutoML studio that supports DARTS neural architecture search algorithm and hyper-parameter optimization methods such as random search, grid search, and TPE. Since the proposed AutoML studio has a user-friendly user interface, non-expert users without deep learning-related knowledge can efficiently perform neural architecture searches and examine the results. The proposed AutoML studio is expected to help experts and non-experts easily develop high-performance models.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00739, MLOps Platform for Machine learning pipeline automation).

References

1. Feurer, Matthias, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. "Auto-sklearn 2.0: Hands-free automl via meta-learning." arXiv preprint arXiv:2007.04074 (2020)
2. LeDell, E., & Poirier, S. (2020, July). H2o automl: Scalable automatic machine learning. In Proceedings of the AutoML Workshop at ICML (Vol. 2020)
3. Jin, Haifeng, Qingquan Song, and Xia Hu. "Auto-keras: An efficient neural architecture search system." In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1946-1956. 2019
4. Erickson, Nick, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. "Autogluon-tabular: Robust and accurate automl for structured data." arXiv preprint arXiv:2003.06505 (2020)
5. Liu, Hanxiao, Karen Simonyan, and Yiming Yang. "Darts: Differentiable architecture search." arXiv preprint arXiv:1806.09055 (2018)
6. Pham, H., Guan, M., Zoph, B., Le, Q., & Dean, J. (2018, July). Efficient neural architecture search via parameters sharing. In International conference on machine learning (pp. 4095-4104). PMLR

7. Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. "Algorithms for hyper-parameter optimization." Advances in neural information processing systems 24 (2011)
8. Microsoft. Neural Network Intelligence (1.7), 1 2021. https://github.com/microsoft/nni

AutoCache: Efficient Execution of UDF through the Detection of Cached Variables for the Analytical Analysis on Federated Databases

Md Arif Rahman and Young-Koo Lee*

Kyung Hee University, Korea
{ma.rahman,yklee}@khu.ac.kr

http://www.khu.ac.kr

Abstract. Recently, User Defined Functions (UDF) have become widely popular for the execution of analytical queries on the database system. UDF execution in federated relational databases is challenging due to the enormous data transfer of relational data from federated sites and the management of the program at the host site. To execute a UDF, data for each statement (query) must be obtained from remote locations and preserved on the host machine. UDFs for diverse workloads use dependent read-only queries that use relations from multiple data sources. Caching data for all queries is extremely costly. Conversely, not caching any data for any queries may increase response time due to frequent data accesses. To address this issue, we propose a technique, *AutoCache*, for determining the UDF variables required for caching. Our technique is to cache the required relation variables by materializing them. Our experiment shows that *AutoCache* outperforms the state-of-the-art research.

Keywords: UDF Optimization, Cached variables, Federated Databases

1 Introduction

A federated relational database is a set of databases that collaborate but do not replicate data. Consequently, this system does not require the expensive data integration. In a federated environment, all relevant data is transferred to the host system before any query is executed. Therefore, in a federated environment, the time required to transfer data has a significant impact on the performance of query execution [6–8].

These days, UDFs are commonly utilized for querying in analytical workloads. A UDF can store many analytical queries. There are several advantages to using a UDF rather than a standard declarative query because it incorporates the best of both querying and programming logics. So, programmers are

* Corresponding Author

getting more and more interested in UDF during analytical processing from databases. [1–3].

A UDF may contain both independent statement that can be processed alone and dependent statement that requires data from other statements to process. In the federated database system, certain statements may retrieve data directly from the base tables stored in certain locations (in the host or at remote sites). During the execution of a UDF, the result of a statement is often stored in a non-cached table variable, regardless of its dependencies with other statements. Not caching variables has several advantages. It saves time taken to store the results in memory and to remove the stored value at the completion of UDF execution. However, if these non-cached variables for remote relations are retrieved within a loop, the repeated data transfer from a remote site to the host machine may result in significant performance decrease. Rather than sending data for each iteration, caching all variables may solve this problem. However, caching all variables may result in a decrease in performance due to the time required to store all data and delete all variables following UDF execution. Existing state-of-the-art static UDF analysis research [1, 2] optimizes a UDF at the host site by determining whether or not to cache the variables after receiving the results from remote sites for all the variables in the UDF. Nonetheless, they are unable to optimize the UDF by identifying a subset of the UDF's variables that would benefit greatly from being cached and so optimize the UDF.

Here, we offer an intuition for our problem. There are four relations in a UDF called example_udf: $tabA, tabB, tabC$, and $tabD$. We assume that all nodes are stored at the host site with the exception of $tabC$, which is stored in the remote site. here are multiple statements contained within the UDF. The table variables s_1, s_2, s_3, and s_4 are responsible for storing the results of associated statements. The loop accesses the dependent table value s_2 that is dependent on s_1. Current method either caches all of the variables that require a lot of space or none of them, leading to frequent remote data access throughout each iteration within a UDF. In this research, we are specifically interested in caching a subset of a UDF's useful variables.

Example of a UDF

```
CREATE VIEW AS SELECT FROM R.tabC;
CREATE PROCEDURE example_UDF(IN i INT, IN k INT, OUT result INT...)
AS BEGIN
...
  s1 = SELECT ... FROM tabC...;
  s2 = SELECT ... FROM s1 JOIN tabA...;
  s3 = SELECT ... FROM tabB JOIN tabD...;
  WHILE (condition...)
    s4 = SELECT ... FROM s2 JOIN s3 ...;
  END WHILE;
...
END;
```

In this research, we present a novel caching technique within a UDF named *AutoCache* that caches useful variables by recognizing them based on data and control flow dependency between statements and optimizes an execution plan for federated databases.

The scope of the study is to optimize UDF with multiple dependent statements in a given federated relational database where data location is known.

The following summarizes our contribution:

- We introduce a novel caching technique within a UDF named *AutoCache* that recognizes interesting statements that retrieve data from the base relation on a remote site, materializes them for caching, and optimizes the execution plan for the federated databases.
- We show experimental evidence that *AutoCache* outperforms existing methods.

2 Proposed Techniques

In this section, we describe our proposed technique.

2.1 Identification of Interesting Variables

We first identify the interesting variables within a UDF that are useful for caching. We employ a UDF's control and data flow dependency graphs. We determine the variables that acquire data directly from the base relation by examining the graph. We find the variable containing a statement that retrieves data from remote base relations among these variables, presuming that the relation setup of the federated database is already known.

Figure 1 illustrates the control and data flow dependencies of the example UDF. We determine from the graph that the s_1, s_2, and s_3 variables contain statements that retrieve data from the base relations. As we know that the *tabC* relation is on the remote site, we consider s_1 to be our interesting variable.

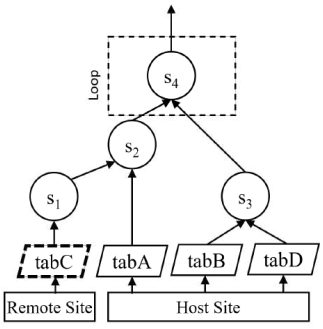


Fig. 1: Control and Data Flow dependencies of our example UDF

2.2 Caching Interesting Variables

After identifying the variables of interest, we cache them by materializing the results of the corresponding statements. By materializing, the result of the statement is physically stored and the variable can be accessed like a local table.

In our example UDF, the result of the statement that is assigned to the s_1 variable is materialized.

2.3 Generating Execution Plan

This is the final outcome of our methodology, in which we generate an optimized plan for a federated database by rewriting some codes for materialization. We presume the current plan of the system creates the execution plan without caching. Consequently, we replace only the parts of the code that include variables of interest.

In our example UDF, we substitute the first statement with the following statement.
Replace the first statement with the following statement

```
CREATE TEMP TABLE s1 AS SELECT ... FROM tabC...;
```

2.4 Pseudocode for *AutoCache*

In Algorithm 1, we present a pseudocode for *AutoCache*. Our methodology receives as input the UDF and remote node with associated base relations, and as output the execution plan for the federated environment. *getGraph()* retrieves the control and data flow graphs of the UDF in the algorithm. *getBaseNode()* identifies the query that retrieves data straight from the base relation. *getInterestingVar()* then determines the variable that contains the statement that retrieves data from the remote base table. *getExistingPlan()* obtains the existing execution plan through traditional methods without caching. *getPlan()* finally generates the plan by rewriting the matching statement with interesting variables.

3 Experiment

We present the experiments demonstrating *AutoCache*'s superiority to contemporary approaches.

3.1 Experimental Environment

We utilize Python version 3.8 to compile our code. Using PostgresSQL (accessible for free at https://www.postgresql.org) on four servers equipped with 2.10GHz 45-core Intel Xeon (R) Silver 4216 processors and 512GB of RAM per machine, we present our experimental findings.

Algorithm 1: Algorithm for *AutoCache*

Input: UDF (U), Remote Node(R_n)
Output: Execution Plan of UDF ($plan$)
ExecutionPlan(U, R_n)**:**
    $cfg \leftarrow$ getGraph(U)
    $baseNode \leftarrow$ getBaseNode(cfg)
    $interestVar \leftarrow$ getInterestingNode($baseNode, R_n$)
    $existingPlan \leftarrow$ getExisitingPlan(U)
    $plan \leftarrow$ getPlan($existingPlan, interestVar$)

Function: getInterestingVar($baseNode, R_n$)

---

getInterestingVar$(baseNode, R_n)$**:**
    FOREACH i *in* $baseNode$**:**
        if i *in* R_n **then**
            $interestedVar \leftarrow append(i)$
        end
    RETURN $interestedVar$

Function: getPlan($existingPlan, interestVar$)

---

getPlan$(existingPlan, interestVar)$**:**
    FOREACH s_i *in* $existingPlan$ *//every statement in* $existingPlan$**:**
        if $interestVar$ *in* s_i **then**
            $r \leftarrow rewrite(s_i)$
            $newPlan \leftarrow append(r)$
        else
            $newPlan \leftarrow append(s_i)$
        end
    RETURN $newPlan$

We create a synthesized workload based on the TPC-DS benchmark that consists of five fabricated UDFs. We store all of the necessary relations for the workload on three of the four remote nodes and consider the fourth node to be the host machine.

When implementing Python code to PostgreSQL server, we utilize the Python-supplied psycopg2 package. We communicate with the remote source via remote view on the host system.

3.2 Performance Analysis

In this section, we show the experimental findings that demonstrate the performance of *AutoCache*.

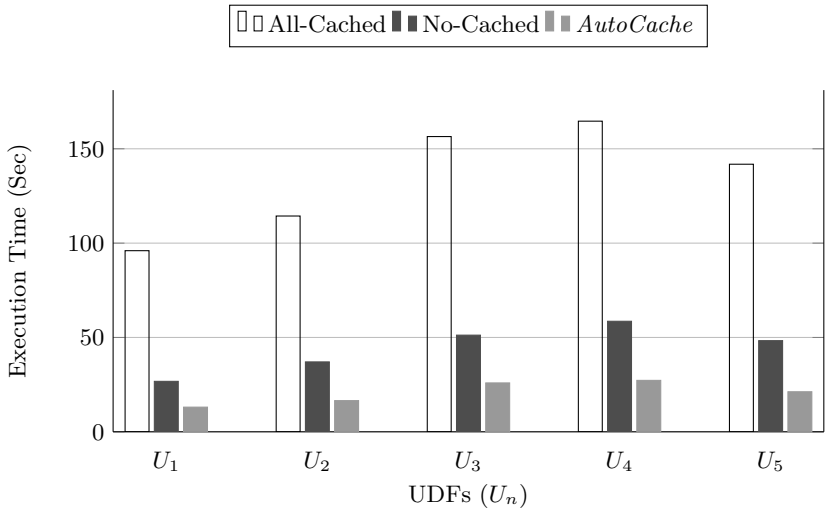


Fig. 2: Performance analysis of *AutoCache*

Figure 2 displays a comparison of *AutoCache*'s performance to that of state-of-the-art system. The horizontal axis depicts the UDFs of the simulated workload, while the vertical axis depicts execution time in seconds. We compare our method to the state-of-the-art static analysis of UDF optimization [1, 2] in both cases: caching all variables (All-cached) and not caching any variables (No-cached). We observe that our method achieves a 634 percent improvement over a no-cache system and a 104 percent improvement over an all-cache system, respectively, for the U_1 evaluation.

We gain the improvement over the no-cache environment because a dependent statement retrieves a variable from the remote site from within a loop. This enhancement would be bigger if the iteration contained additional dependant statements that all retrieved data from remote locations. On the other hand, we get an improvement over an all-cached environment because caching all the data for each statement and deleting them after UDF execution is complete requires additional time. Since there are hundreds of statements within a UDF, we would achieve greater improvement in this instance.

We have reached the conclusion that our *AutoCache* is superior to the most advanced techniques.

4 Related Works

In order to prove the novelty of our work, we surveyed a wide range of relevant literature, focusing on studies that addressed UDF optimization, Federated query processing, and query caching in distributed relational database systems. We also investigated the present practices of certain major companies utilizing relational databases (i.e., Oracle, IBM, SAP etc.)

DBridge [4, 5], and some publications have been published on the topic of UDF optimization. Using a similar declarative language, Aggify [3] showed that imperative constructions may be evaluated over row-wise results. This paper's focus was on replacing cursor loops with equivalent pipelined computation. Froid [1] leveraged the preexisting subquery optimization approaches to transform iteratively inefficient UDF execution strategies into set-oriented plans with significantly better performance. Research on predicate pushdown for cross-statement optimization has focused on decorrelating UDFs [2].

The topic of federated query processing garners a significant amount of attention in [6–8]. When deciding whether or not to offload a code in modern RDBMSs like SAP HANA, Oracle DBIM, and IBM DB2, the amount of data contained in a table is one of the factors that is taken into consideration. If there is a very large quantity of data being processed, the execution will take place in the source where the data is being stored.

We investigated query caching in relational databases that were distributed over multiple nodes. The work presented in [9] focuses on shortening the amount of time required for the execution of code by directly accessing records stored in cache memory and saving space in cache memory by removing material that is not used frequently. The work described in [10] focuses on analyzing the temporal differences in query submissions and making use of these variations so that the speed of the result caching can be improved. The work presented in [11] focuses on optimizing multiple queries by mixing in-memory cache primitives with them. However, not a single one of them addresses the issue of query caching within a UDF.

5 Conclusion

The complexity of the computations performed by the database engine grows daily as a direct result of the rising volume of data. Caching the results of queries within a UDF is becoming difficult due to the fact that caching unnecessary data may result in performance reduction.

During the course of this investigation, we created a novel variable cached selection technique that we called *AutoCache*. This technique investigates the idea of caching important queries within a UDF in order to maximize performance across federated relational database systems.

Lastly, we demonstrate that *AutoCache* outperforms existing systems when it comes to executing procedures at remote locations. Based on this, we infer that the *AutoCache*'s main strength is in minimizing data transfer in a federated environment by caching interesting relations locally on the host machine. However, the method only works in a federated setting, so interesting relations may already be replicated on the remote or host machines and we won't have any benefit there.

Acknowledgments. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center

support program(IITP-2021-2015-0-00742) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

References

1. Ramachandra, K., Park, K., Emani, K.V., Halverson, A., Galindo-Legaria, C. and Cunningham, C., 2017. Froid: Optimization of imperative programs in a relational database. Proceedings of the VLDB Endowment, 11(4), pp.432-444 (2017)
2. Park, K., Seo, H., Rasel, M.K., Lee, Y.K., Jeong, C., Lee, S.Y., Lee, C. and Lee, D.H., 2019, June. Iterative query processing based on unified optimization techniques. In Proceedings of the 2019 International Conference on Management of Data (pp. 54-68) (2019)
3. Gupta, S., Purandare, S. and Ramachandra, K., 2020, June. Aggify: Lifting the curse of cursor loops using custom aggregates. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 559-573) (2020)
4. Chavan, M., Guravannavar, R., Ramachandra, K. and Sudarshan, S., 2011, April. DBridge: A program rewrite tool for set-oriented query execution. In 2011 IEEE 27th International Conference on Data Engineering (pp. 1284-1287). IEEE (2011)
5. Emani, K.V., Ramachandra, K., Bhattacharya, S. and Sudarshan, S., 2016, June. Extracting equivalent sql from imperative code in database applications. In Proceedings of the 2016 International Conference on Management of Data (pp. 1781-1796) (2016)
6. Josifovski, V., Schwarz, P., Haas, L. and Lin, E., 2002, June. Garlic: a new flavor of federated query processing for DB2. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (pp. 524-532) (2002)
7. Qin, A., Yuan, Y., Tan, D., Sun, P., Zhang, X., Cao, H., Lee, R. and Zhang, X., 2017, April. Feisu: Fast Query Execution over Heterogeneous Data Sources on Large-Scale Clusters. In 2017 IEEE 33rd International Conference on Data Engineering (ICDE) (pp. 1173-1182). IEEE (2017)
8. Jiang, H., Gao, D. and Li, W.S., 2008. Improving parallelism of federated query processing. Data Knowledge Engineering, 64(3), pp.511-533 (2008)
9. Hassan, C.A.U., Hammad, M., Uddin, M., Iqbal, J., Sahi, J., Hussain, S. and Ullah, S.S., 2022. Optimizing the Performance of Data Warehouse by Query Cache Mechanism. IEEE Access, 10, pp.13472-13480(2022)
10. Kucukyilmaz, T., 2021. Exploiting temporal changes in query submission behavior for improving the search engine result cache performance. Information Processing Management, 58(3), p.102533 (2021)
11. Michiardi, P., Carra, D. and Migliorini, S., 2021. Cache-based multi-query optimization for data-intensive scalable computing frameworks. Information Systems Frontiers, 23(1), pp.35-51 (2021)

Image Retrieval with GrabCut and feature matching

Dayoung Park and Youngbae Hwang

[1] Dept. Of Control and Robot Engineering, Chungbuk National University, Cheongju 28644, South Korea. wnidy100@chungbuk.ac.kr, ybhwang@cbnu.ac.kr

Abstract. Recently, content-based image retrieval (CBIR) has been used in various fields such as web portals, online shopping malls, and training dataset collection for user growth and efficiency. For a database consisting of approximately 16k fruit and vegetable images, we present an image retrieval method using GrabCut based foreground extraction and invariant feature matching. The rank for retrieval is determined based on the total distance of all matched descriptors. Experimental results show 80 percent of accuracy for query images.

Keywords: CBIR, GrabCut, feature matching, SIFT

1 Introduction

Content-based image retrieval (CBIR) is performed based on the similarity of features of the image. Recently, CBIR has been used in various fields such as web portals, online shopping malls, and training dataset collection, which result in user growth and efficiency increasement. We propose a method for performing CBIR on extracted foreground of an input image. Our method uses GrabCut [1] for foreground extraction and SIFT [2] based feature matching for the Fruit dataset. We show that the proposed method achieves 80% accuracy for query images.

2 Methods

For an input query image, when the user selects a bounding box corresponding to the target object, the GrabCut algorithm is applied to the area to separate the background and the foreground. Additional scribbles can be assigned to indicate remained or removed regions by user intention. The SIFT algorithm is then used to extract the descriptors of the database images and a query image created by synthesizing the extracted foreground on a white background. To obtain image similarity between a query image and one of database images, the Euclidean distance of all descriptor pairs is calculated by the brute force method to find feature pairs with the minimum distance. Finally, images in the database are ranked in the order in which the total distance of all the matched descriptors is smaller. The geometric verification such as homography estimation with RANSAC is not applied to consider non-rigidity of the target object.

3 Experimental results

To evaluate the proposed image retrieval method, the database was constructed with 16,854 images of 100x100 image resolution which consists of 33 classes for fruits and vegetables. The dataset was collected from Kaggle, and each image includes a segmented object taken from various angles. Fig. 1(a) shows samples of the database. As a query image, two target objects of a pineapple and a banana, in the 267x180 image were used. The foreground segmentation results for the target objects are shown in Fig 1(b).

Fig. 2. shows retrieved images when the number of keypoints is 40. The retrieval result of the pineapple was 100 percent positive, but when the retrieving for the banana, the retrieval result was 60 percent positive, because the rest are incorrectly retrieved for a corn as a banana. Our method can retrieve images with the same category and various angles.

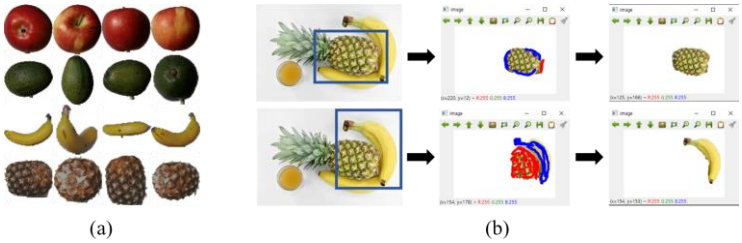


Fig. 1. (a) Examples of the database and (b) applying the GrabCut algorithm for query image.

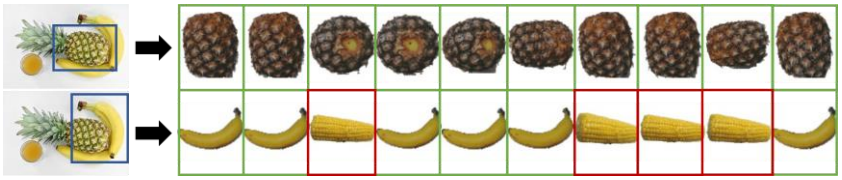


Fig. 2. Top-10 image retrieval results for regions marked with blue bounding boxes. Positive and negative images are depicted with green and red border, respectively.

Acknowledgement

This work was partially supported by the Grand Information Technology Research support program (IITP-2022-2020-0-01462), by Institute of Information & communications Technology Planning & Evaluation (IITP) (No.2022-0-00970).

References

1. Rother, Carsten, Vladimir Kolmogorov, Andrew Blake.: "GrabCut" interactive foreground extraction using iterated graph cuts. In: ACM transactions on graphics, 309-314 (2004)
2. Lowe, David G.: Distinctive image features from scale-invariant keypoints. In: International journal of computer vision 60.2, 91--110 (2004)

Prognosis Prediction using Multimodal Deep Learning in Diffuse Large B-Cell Lymphoma Patients

Sy-Phuc Pham[1,*], Sae-Ryung Kang [2,*], Hyung-Jeong Yang[1,+], Deok-Hwan Yang[2,+], Soo-Hyung Kim[1], Guee-Sang Lee[1]

[1]Chonnam National University, Gwangju, South Korea
[2]Chonnam National University Hwasun Hospital, Gwangju, South Korea
phamsyphuc123@gmail.com, campanella9@naver.com
{hjyang, drydh, shkim, gslee}@jnu.ac.kr

Abstract. In this day and age of precision medicine, determining the prognosis of cancer based on various inputs requires highly advanced computational techniques. In this paper, we introduce a comprehensive multimodal deep learning approach for autonomous patient risk prediction in Diffuse Large B-cell Lymphoma. Positron Emission Tomography scans, in conjunction with the tabular clinical information, are utilized in our approach. Our method achieves positive results when combining multiple feature representations using Keyless Attention mechanism. Our approach achieves 0.7502 C-index and 0.1554 Integrated Brier Score, better than the results of unimodal using clinical information with 0.722 C-index and 0.1658 Integrated Brier Score. Besides, the proposed loss function helps the multimodal work more effectively than using the traditional loss function in survival task.

Keywords: Lymphoma, prognosis prediction, deep learning.

1 Introduction

Approximately 30%–40% of all instances of lymphoid neoplasms in adults are classified as Diffuse Large B-Cell Lymphoma (DLBCL) [1,2]. Predicting therapy success and identifying DLBCL patients who are not likely to be cured is therefore crucial. The well-known international prognostic index (IPI) is an effective diagnostic tool that can be used to categorize patient risks [3]. However, many patients were not treated because of conditions including late relapse or primary refractoriness. Therefore, it is still very important to work on refining the survival prediction model.

We are shifting from a population-based model of healthcare to one that is more focused on individual patients. Computational approaches are essential for analyzing

[*] Equal contribution.
[+] Corresponding author.

the massive amounts of biomedical data now available for individual patients because of advances in imaging technology. In particular, the use of Deep Learning (DL) has shown remarkable promise in a wide variety of contexts, most notably in the diagnosis and prognosis of cancer [3]. Predicting a patient's prognosis in oncology is a difficult but essential endeavor. Clinical outcome prediction, including mortality and cancer recurrence, is a foundation for many choices in healthcare. Survival analysis is a branch of statistics concerned with modeling the elapsed time before an event of interest happens. Time-to-event prediction is one of the subfields that falls under survival analysis. When carrying out a survival analysis, both the raw data and the amount of time that has passed since the occurrence of the event are taken into consideration. Moreover, it incorporates all samples for which the occurrence was not witnessed prior to the previous interaction (e.g., if a patient cannot be located for clinical follow-up at any stage prior to their death). This information is stored as a binary label that indicates whether or not the occurrence was observed. Right-censored samples are those in which the occurrence was not observed.

Cox proportional hazards (CPH) linear modeling is the standard method for dealing with survival data [4]. More than two decades ago, Faraggi and Simon presented a non-linear modification of the CPH algorithm [5]. DLBCL survival prediction study's findings are founded on both genetic and clinical information [6,7,8]. In addition to the original approach, which relied on unimodal input data, multimodal DL has also been used to predict cancer patients' chances of survival.

Here, we improve upon the Faraggi-Simon approach, a full-stack Multimodal DL model for predicting DLBCL survival. Compared to prior methods, our proposed method incorporates a wider variety of input data modalities. A combination of Positron Emission Tomography (PET) scans and tabular clinical data modalities is included in the input data. Our approach uses a combination loss as the loss for the multimodal and uses Keyless Attention in the multimodal fusion layer for medical data fusion. The contributions of this study are as follows:

- We proposed an end-to-end multimodal deep learning architecture for DLBCL, which can combine three modalities with ease and efficiency.
- We use Keyless Attention to find the essential feature representations and make the input for the survival prediction module.
- We utilize combined loss for the survival prediction task to make the data fusion easier between feature representations.

The content of this paper is organized as follows. Section 2 presents the proposed method, which we used for the survival prediction task. In section 3, we discuss the experimental results. Finally, the conclusions and future work are presented in section 4.

2 Methods

In this section, we introduce our approach for multimodal deep learning survival prediction including feature extraction, fusion layer, and survival prediction module. The overall architecture of the proposed method is illustrated in Figure 1.

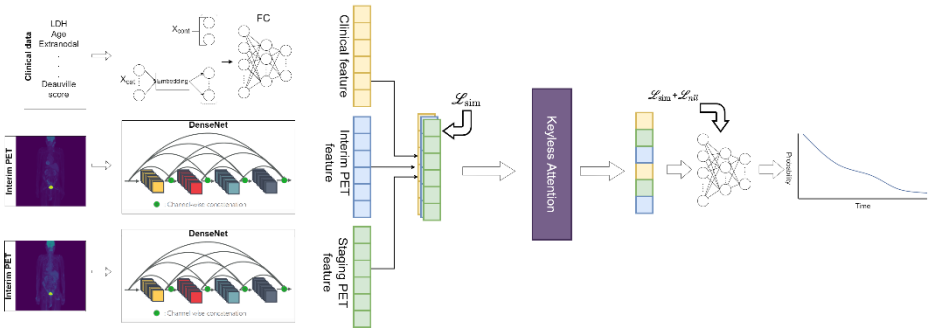


Fig. 1. Overview of the architecture of the proposed model for the survival prediction of DLBCL patients.

2.1 Data modality feature extraction

To facilitate the extraction of features from the various input data types, we built a unique sub-model for each. Categorical embeddings are used for the categorical data in clinical data. The embedding outputs are then passed into a fully-connected layer, where they are concatenated with the continuous variable. Transfer learning is used for the PET scans sub-model. DenseNet [9] convolutional neural network (CNN) architecture is applied and used. All training parameters are first specified in the CNN's lower layers using a fully-connected layer, allowing for further fine-tuning of the remaining parameters. Each specialized model generates a feature representation vector that is then used by the target modality. The feature representation $T \in R^a$ produce $T \in R^{a \times b}$, with $b = 3$ and $a = 128$.

2.2 Fusion technique

In order to perform the multimodal fusion process, an attention mechanism is used, with the aforementioned feature vectors serving as input. For medical data such as PET scans and clinical information, we improved upon a keyless attention method in [10] by adding a non-linear function based on the hyperbolic tangent (tanh). The fusion layer learns $F \in R^{a \times b}$, with each column providing the weights for the corresponding feature vector t_j, which is the column j of the matrix T. The feature vectors are then used to create a fusion vector c by summing their elements in a weighted manner, as detailed in the Equation 1.

$$c_i = \sum_{j=1}^{n} a_{ij} t_{ij} \tag{1}$$

$$a_j = softmax(W_j t_j) \tag{2}$$

where $W_j \in R^{a \times a}$ is the matrix j containing the weights that were learned of $W \in R^{a \times a \times b}$ is the tensor. Essentially, the model can figure out how much weight to give to each of the three data modalities before merging their feature vectors.

2.2 Fusion technique

This is the hazard function defined by the CPH model [11]:

$$\lambda\,(t|x) = \lambda_0(t)e^{h(x)} \qquad h(x) = \beta^T x \tag{3}$$

Baseline hazard $\lambda_0(t)$ is a non-parametric measure. $h(x)$ is a hazard function. In order to get an accurate estimation of the risk function, we employ fully- connected networks that only have one node serving as the output. Performing a loss function when minimizing the average negative partial log-likelihood, L_{nll} is the function to use:

$$L_{nll} = -\frac{1}{N}\sum_{i:E_i}\left(h(x_i) - log \sum_{j:T_j \geq T_i} e^{h(x_i)}\right) \tag{4}$$

where time of the event, or Ti, and event indicator, or Ei, are two separate quantities. In this case, xi stands for the ith observation's data. Patients who were not censored ($E_i = 1$) are denoted by N. The original Faraggi-Simon method was designed similarly to this loss function [5].

We experimented with an additional loss that penalized dissimilarity across the data modality representations since we thought that doing so could make data fusion easier. The average cosine distance between feature representations of different modalities in the input data served as this additional loss. The total loss was calculated by adding the primary loss L_{nll} and the secondary loss L_{sim} together at a predetermined weighting factor. However, we find that include this additional loss improves performance, therefore it is baked into the final configuration of the approach we offer. The final loss has been described by this function:

$$L_{total} = L_{nll} + L_{sim} \tag{5}$$

This additional loss establishes a metric for evaluating the loss for a particular set of input tensors $x = x_1, x_2$ and a tensor label $y = 1\ or\ -1$. The function for detailing the additional loss has been shown in Equation 6.

$$L(x, y) = \begin{cases} 1 - \cos(x_1, x_2), & if\ y = 1 \\ \max(0, \cos(x_1, x_2) - margin), & if\ y = -1 \end{cases} \tag{6}$$

With constants margin are declared with the value 10^{-5}, the set present the input tensor is $Z = Z_{clinical}, Z_{iPET}, Z_{sPET}$ with $Z_{clinical}$ is the clinical feature extraction, Z_{iPET} is the interim PET feature extraction, and Z_{sPET} is the staging PET feature extraction. We calculate the additional loss as detail:

$$L_{sim} = L_{sim1} + L_{sim2} + L_{sim3} \tag{7}$$

$$L_{sim1} = \begin{cases} 1 - \cos(Z_{clinical}, Z_{iPET}), & if\ y = 1 \\ \max(0, \cos(Z_{clinical}, Z_{iPET}) - margin), & if\ y = -1 \end{cases} \quad (8)$$

$$L_{sim2} = \begin{cases} 1 - \cos(Z_{clinical}, Z_{sPET}), & if\ y = 1 \\ \max(0, \cos(Z_{clinical}, Z_{sPET}) - margin), & if\ y = -1 \end{cases} \quad (9)$$

$$L_{sim3} = \begin{cases} 1 - \cos(Z_{iPET}, Z_{sPET}), & if\ y = 1 \\ \max(0, \cos(Z_{iPET}, Z_{sPET}) - margin), & if\ y = -1 \end{cases} \quad (10)$$

3 Experimental and results

3.1 Data

We use data from Chonnam National University Hwasun Hospital (CNUHH), which is collected by medical experts. The data includes clinical information, interim PET, staging PET. Patients were tracked until death or they were no longer clinically observable. We excluded patients with missing data and clinical follow-up time and split 602 patients with the same censored rate into five-fold cross-validation. In the clinical information, we have two types for the clinical data categorical variables and continuous variable. The clinical included 12 features will be described in Table 1. The 12 clinical features were divided into two groups by the medical doctor before treatment and after treatment. The Overall Survival measures the time until death or the end of follow-up after therapy has been discontinued. For the PET scans, we use Standard Uptake Value (SUV) [12] to normalize the PET scans. Then, we apply Maximum Intensity Projection (MIP) [13] to convert the PET scan into 2D image.

3.2 Experiment details

Pytorch was used to realize the proposed model. We used Adam stochastic gradient descent optimization [14] to update network weight, with the default configurations Pytorch and a learning rate with an initial value of 10^{-2} to 10^{-1}. The number of epochs is 100, and we saved the best model in the training process. If the goal is not optimized after 15 epochs, we will terminate training early. We used NVIDIA GTX 2080Ti for the training models.

3.3 Results

We evaluated the model performance on the test set and calculated the average value for two metrics. The first metric is concordance index or C-index [15] The C-index measures the proportion of orderable sample pairings for which the predicted risk is

Table 1. Characteristics of clinical information in DLBCL patients.

Treatment status	Data field	Characteristic	Value	CNUHH(n=602)
Pre- treatment	Numeric	Age	Min-max	17-92
		LDH	Min-max	144-8402
	Categories	Performance	1	201 (33.39%)
			2	322 (53.49%)
			3	65 (10.8%)
			4	14 (2.33%)
		B symptom	0	504 (83.72%)
			1	98 (16.28%)
		Extranodal status	0	456 (75.75%)
			1	149 (24.25%)
		Stage	1	118 (19.6%)
			2	195 (32.39%)
			3	137 (22.76%)
			4	152 (25.25%)
		Spleen involvement	0	579 (96.18%)
			1	23 (3.82%)
		BM involvement	0	554 (92.03%)
			1	48 (7.975%)
		IPI score	0	81 (13.46%)
			1	152 (25.25%)
			2	134 (22.26%)
			3	131 (21.76%)
			4	76 (12.62%)
			5	28 (4.65%)
		IPI risk	1	233 (38.7%)
			2	134 (22.26%)
			3	131 (21.76%)
			4	104 (17.28%)
		R-IPI	1	81 (13.46%)
			2	286 (47.5%)
			3	235 (39.04%)
During treatment		Deauville score	1	290 (48.17%)
			2	108 (17.94%)
			3	74 (12.29%)
			4	86 (14.29%)
			5	44 (7.31%)

correct. We use Integrated Brier Score (IBS) [16] for the second metric The IBS calculates the mean squared differences between actual survival rates and expected survival rates. First, we evaluated unimodal data inputs, and we got the best result at the unimodal with clinical input. Second, we conducted an analysis of multimodal settings, using a wide variety of combinations of different types of input data. The combination of clinical information with interim PET achieved 0.7246 C-index and

0.165 IBS. When clinical data and PET scans are combined with L_{nll}, the model's performance was around 0.7346 and 0.1518 for C-index and IBS, respectively. The best result of 0.7502 C-index and 0.1554 IBS were obtained by combining three modalities with L_{total}. All the results have been described in Table 2.

Table 2. The performance of the multimodal with each modality's various combinations: clinical information, interim PET scans, and staging PET scans. The best results are in bold.

Included modalities	Average (5 folds)	
	C-index	IBS
Clinical (A)	0.722	0.1658
Interim PET (B)	0.493	0.2416
Staging PET (C)	0.6004	0.2216
A + B (L_{nll})	0.7246	0.165
A + C (L_{nll})	0.7358	0.1592
A + B + C (L_{nll})	0.7346	0.1558
A + B + C ($L_{sim} + L_{nll}$)	**0.7502**	**0.1554**

4 Conclusions

In this paper, we introduced a proposed end-to-end strategy for predicting the prognosis of DLBCL using a combination of PET scans and clinical information. The proposed method uses multiple inputs, comprising an interim PET scan combined with a staging PET scan and tabular clinical information. Our proposed method included three steps. First, extract features from tabular clinical data using an embedding layer and a fully connected layer, and extract features from a PET scan using a pre-trained of DenseNet. Second, we apply Keyless Attention mechanism for fusing feature representations. Finally, the multi-layer perceptron performs the estimation of survival time. We plan to increase the performance of our model in the future by examining the relationships between different types of data.

Acknowledgements This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0021)

References

1. Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. "Cancer statistics, 2019." CA: a cancer journal for clinicians 69.1 (2019): 7-34.

2. Armitage, James O., et al. ”Non-hodgkin lymphoma.” The lancet 390.10091 (2017): 298-310.
3. International Non-Hodgkin’s Lymphoma Prognostic Factors Project. ”A predictive model for aggressive non-Hodgkin’s lymphoma.” New England Journal of Medicine 329.14 (1993): 987-994.
4. Esteva, Andre, et al. ”A guide to deep learning in healthcare.” Nature medicine 25.1 (2019): 24-29.
5. Faraggi, David, and Richard Simon. ”A neural network model for survival data.” Statistics in medicine 14.1 (1995): 73-82.
6. Merdan, Selin, et al. ”PCN10 MACHINE LEARNING PREDICTION OF SURVIVAL IN DIFFUSE LARGE B-CELL LYMPHOMA BASED ON GENEEXPRESSION PROFILING.” Value in Health 23 (2020): S23-S24.
7. Li, Lexin. ”Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information.” Bioinformatics 22.4 (2006): 466-471.
8. Mosquera Orgueira, Adri´an, et al. ”Improved personalized survival prediction of patients with diffuse large B-cell Lymphoma using gene expression profiling.” BMC cancer 20.1 (2020): 1-9.
9. Huang, Gao, et al. ”Densely connected convolutional networks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
10. Long, Xiang, et al. ”Multimodal keyless attention fusion for video classification.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
11. Cox, David R. ”Regression models and life-tables.” Journal of the Royal Statistical Society: Series B (Methodological) 34.2 (1972): 187-202.
12. Kinahan, Paul E., and James W. Fletcher. ”Positron emission tomographycomputed tomography standardized uptake values in clinical practice and assessing response to therapy.” Seminars in Ultrasound, CT and MRI. Vol. 31. No. 6. WB Saunders, 2010.
13. Kim, J., et al. ”PET-CT visualisation with integrated maximum intensity projection and direct volume rendering.” (2009): 180-180.
14. Kingma, Diederik P., and Jimmy Ba. ”Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980 (2014).
15. Harrell, Frank E., et al. ”Evaluating the yield of medical tests.” Jama 247.18 (1982): 2543-2546.
16. Gerds, Thomas A., and Martin Schumacher. ”Consistent estimation of the expected Brier score in general survival models with right-censored event times.” Biometrical Journal 48.6 (2006): 1029-1040.

Using HRNet Pose Estimator in Two-Stream Violence Detector for Crowd Situations

Toshpulatov Azamat and Yoo-Sung Kim

Department of Information and Communication Engineering,
Inha University
Incheon 22212, Korea

atoshpulatov3@gmail.com, yskim@inha.ac.kr

Abstract. A new two-stream violence detection scheme in which HRNet pose estimator is used for well-extracting human motions in crowd situations is proposed. Correct detection of violent events in video surveillance systems is essential for taking appropriate responses for them. However, in crowd situations, detecting violence becomes more difficult due to low resolutions and occlusion problems in the input videos. So effectively extracting motion information of humans in crowd situations has been crucial. In this paper, using the accurate pose estimator HRNet as a motion information extractor of humans in crowd situations is proposed. To show the effectiveness of the proposed violence detector according to the number of people involved in violence, 1200 videos are selected from RLVS(Real Life Violence Situation) dataset and sub-grouped into 4 crowd-size cases. According to the experiment results, the proposed scheme using HRNet outperforms the previous scheme using OpenPose in all 4 crowd-size cases.

Keywords: Violence detection, Two-stream, Motion information, HRNet, Crowd situations.

1 Introduction

Violence detection for intelligent CCTV has gained popularity. Motion information is important for detecting violence. To extract motion information from videos, optical flow [1] had been widely used. Recently, skeleton-based pose estimation has more popular to extract motion information. To detect skeletons of humans from videos several schemes have been proposed such as HRNet [2] and OpenPose [3]. HRNet is known as a good candidate because of detecting human pose accurately. To improve the accuracy of violence detections, instead of using only a one-stream network with the origin image frames of the input video, two-stream networks in which the motion information of humans in addition to the image frames is used together have been proposed in [4], [5], [6]. In the two-stream networks, robust split-FAST convolutional blocks [7] learning from given input videos horizontally and vertically are very useful to extract features well. However, the previous scheme [7] has a low-recognition accuracy due to the limitation of the inaccurate motion information extracted by OpenPose and C3D networks.

The main objective of this research is to deal with this issue based on spatial and temporal feature fusion by utilization of global contextual information. We propose an effective two-stream deep learning architecture that leverages skeleton-based human motion information by taking advantage of a pose estimator. That is, instead of OpenPose, one of the accurate pose estimators HRNet is used. Also, we use split-FAST convolution blocks for learning input videos horizontally and vertically to detect motion feature patterns effectively for our HRNet-based model.

Since creating the best violence detection model with high fidelity in densely populated places is crucial, we select some videos from RLVS (Real Live Violence Situation) [8] dataset to check the effectiveness of the proposed model in crowd situations. And we grouped the selected videos into 4 subcategories according to the number of involved people in the video; 1-2 people videos, 3-5 people ones, 6-10 people ones, and 11 or more people videos to imitate real-world crowd situations. Then, in the experiments, we compare the detection accuracy of the proposed two-stream violence detector using HRNet and that of the previous two-stream one with 4 different crowd-sized cases.

This paper is organized as follows. Section 2 outlines related works of two-stream violence detection in addition to utilizing methods of motion feature extraction. Section 3 presents our proposed two-steam violence detector architecture, and describes our way to subcategorize videos from the RLVS dataset. Section 4 shows the experiment results. Finally, section 5 draws a conclusion.

2 Related Works

Previous works ([4-6]) proposed two-stream violence detection schemes. In two-stream architecture, the spatial stream is accompanied by image frames and the temporal stream gets inputs of motion information extracted by OpenPose independently as shown in [Fig.1]. These schemes are known as robust [10-11]. OpenPose is the bottom-up approach that starts by detecting all body parts in an image followed by groping the detected parts into each individual person. The limitation of their generalization capability in various human action occasions in time-based motion analysis [9] is certain. With rapid changes in computer vision, it is not the recently developed and accurate pose estimator for detecting motion patterns and action recognitions in real crowd situations.

The proposed network using a motion extraction scheme by one of the accurate pose estimations like HRNet instead of OpenPose can lead to well-detecting violence events in crowded places. HRNet starts with a person detector and then estimates the body parts of each detected person. In practice, OpenPose throws missing ponts when the human joints are invisible whereas HRNet does try to estimate the location of invisible joints which can be a resource to detect violence accurately. In terms of mAP, HRNet outperforms OpenPose in crowd situations. Furthermore, in two-steam networks traditional 3D-convolutional network [12] has disadvantages in the extraction of flexible feature patterns over split-FAST convolutional blocks [7].

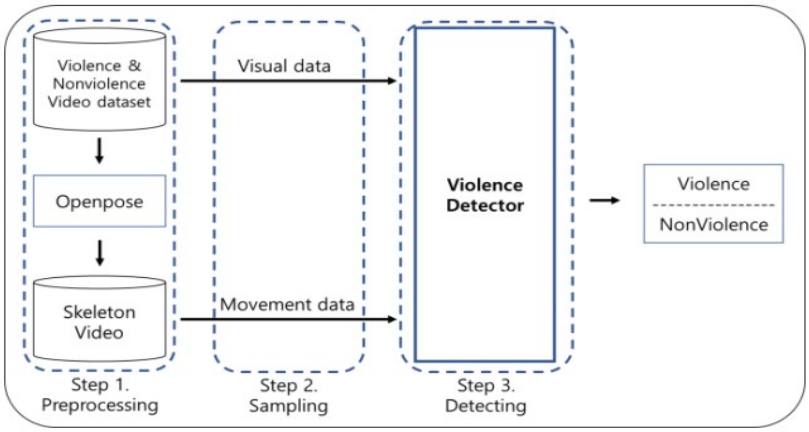


Fig. 1. A previous two-stream violence detector using OpenPose

Another study [8] uses the RLVS database with 2000 videos for developing a violence detector. After that, other studies such as [12] and others challenged the state-of-the-art models on the benchmark dataset for anomaly detection. However, it is hard to say that the RLVS dataset is captured by surveillance cameras in real-world crowd situations because of the lack of qualitative clear video frames, more specific cases (e.g., hockey fights), and the most videos capturing actions performed by mostly three or four people involved.

In real surveillance events in crowds, a large scale of humans should be involved. We selected and separated videos into 4 categories by comparing the density level of people in the video as shown in [Fig. 2].

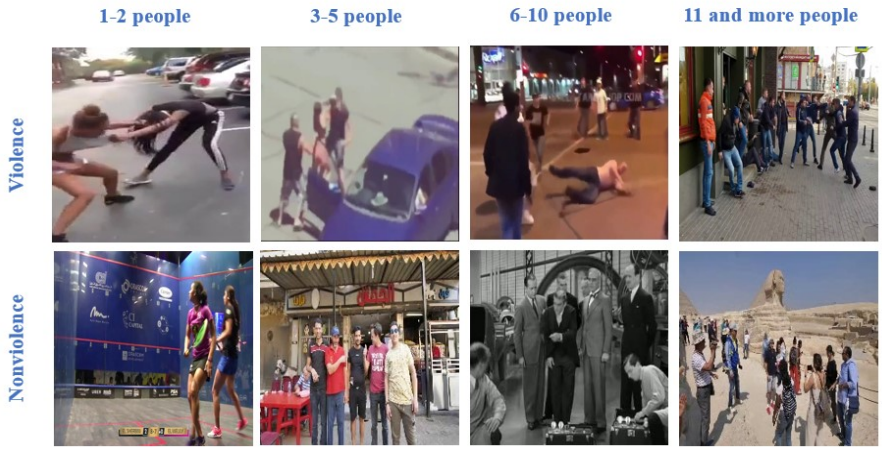


Fig. 2. Sample video frames selected from RLVS and sub-grouped into 4 cases

3 A Two-Stream Violence Detector Using HRNet Pose Estimator

3.1 Overall Architecture of the Proposed Two-Stream Violence Detector

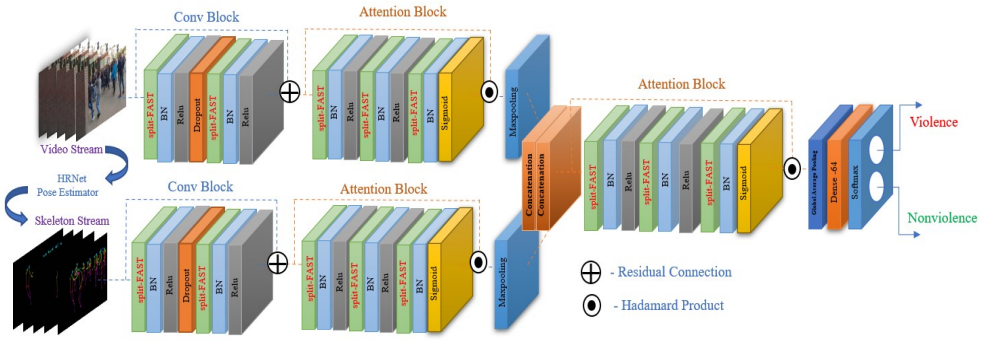


Fig. 3. Overall structure of the proposed two-stream violence detector using

Our proposed violent detector is expressed in three main parts as shown in [Fig. 3]. In the first step, the data preprocessing by HRNet can be done to extract pose estimation. This section specializes in getting only capturing human skeleton movement information from the given dataset. The second part can deal with passing through the detected skeleton frames as movement information and original frames as visual data into our third stage. The average length of video clips is 5 seconds with about 30 FPS, so using all frames of the input video gives unnecessary computation costs and a high degree of redundancy. In the experiment, we tend to use only one frame in 30 continuous frames from extracted movement data and our original data which are dealt with in the proposed second step. Violence detection taking advantage of the visual and movement data simultaneously can predict the given input video whether it is violent or nonviolent. The two-steam network can fuse spatial and temporal features very well in the end when visual and motion information of data is given separately through three-dimensional convolutions. Relying on efficient generalized and adaptable graph convolutional networks by AGC-Net [14] which has a good sense for connectivity of our motion information data for skeleton action recognition. For other detailed structures of the network, we tend to reorganize our two-stream structure in acknowledgment of the AGC-Net's convolutional blocks and the dimensions of each layer's output. Furthermore, the employed 3D convolution uses split-FAST convolution. In the network, C3D or split-FAST convolutional blocks mentioned are employed and tested by experiment. For a clear understanding of the implementation of our proposed model architecture, each stream consists of two parts, Conv Block and Attention Block. The attention block performs the task of selecting only important parts of the prediction from the feature maps entered as input.

That is to learn the weighted attention map that contains the important parts by conducting the element-wise multiplication of the weighted feature map and the original feature map. To do so, the block consists of a 3D convolution layer, and Batch Normalization(BN), and Relu activation, and these layers are stacked and make a block as one stream in our model. Then one stream for spatial and the other stream for the temporal features are fused after attention blocks by concatenation information getting through another attention block and global average pooling making our two-stream model. The visual data patterns passing through the last activation layer and the final output are obtained by the sigmoid function which output is violence or nonviolence. For the proposed model, Adam [15] optimizer is used, and the learning rate is 0.001, the resolution of input videos is 224*224 of 3 color channels, the dropout rate is 0.5.

3.2 Dataset

We need a dataset of real-world crowd situations for training the proposed two-stream violence detector and doing performance evaluation experiments.

Dataset	Violence	Non-violence
1-2 people	220	220
3-5 people	200	200
6-10 people	125	125
11 or more people	55	55
Total	600	600

Table 1. Organization of our dataset selected from RLVS

Since the original RLVS dataset is a collection of videos captured in non-constrained environments, we cleansed and chose 1200 videos to train and test the model for effective crowd management. Chosen videos were subcategorized into 4 small datasets containing 1-2 people, 3-5 people, 6-10 people, and 11 or more people datasets as shown in [**Table 1**].

4 Experiments

The first experiment is intended to show how much our proposed model using HRNet with split-FAST as the pose estimation can enhance the detection accuracies of violent events in crowd situations against the previous scheme using OpenPose with our subcategorized dataset. The weighted accuracies are obtained by taking the average, over all the classes of input videos, and the fraction of correct predictions in this class as shown in [Table 2]. In this experiment motion-based features of HRNet improve the confidence of overall weighted accuracy over OpenPose pose estimator.

Dataset	HRNet	OpenPose
1-2 people	0.89	0.84
3-5 people	0.87	0.82
6-10 people	0.87	0.81
11 or more people	0.84	0.79
Weighted Average	0.87	0.82

Table 2. Performance of two schemes using HRNet and OpenPose with split-FAST blocks

In the second experiment as seen in [Table 3], we compared the detection accuracies of our two-stream model using HRNet to that of one using OpenPose with C3D blocks. From the comparison of performances in [Table 2] and [Table 3], it is found that using the split-FAST block is better than using the C3D blocks for concentrating some specific attention areas of input videos.

Dataset	HRNet	OpenPose
1-2 people	0.83	0.81
3-5 people	0.81	0.79
6-10 people	0.80	0.77
11 or more people	0.78	0.76
Weighted Average	0.81	0.79

Table 3. Performance of two schemes using HRNet and OpenPose with C3D blocks

The proposed two-steam network not only hugely benefits from skeleton images from HRNet pose estimators over OpenPose based models also taking advantages of recognizing the characteristics of human action videos horizontally and vertically by split-FAST is greatly important to generalize unseen data as well and gives more robust performance for the detecting violence in crowd places with HRNet.

5 Conclusion

A new two-stream violence detector using HRNet with split-FAST blocks is proposed for detecting accurately violent events in crowded situations. In the proposed scheme, as an advanced posed estimator HRNet is adapted and used instead of OpenPose, and split-FAST convolutional blocks are used instead of C3D blocks for good attention to the specific areas in video frames vertically and horizontally. So the proposed two-stream network using HRNet with split-FAST blocks to leverage a highly accurate pose estimator for extracting motion information comes in useful to detect the violent behaviors in crowd situations. Preparing a dataset by selecting and sub-grouping videos into 4 classes according to the number of people in a video is done to check the effectiveness of the proposed scheme with respect to the crowd sizes. According to the experiment results, in terms of the weighted average of detection accuracies, the proposed two-stream violence detector using HRNet and split-FAST blocks showing 0.87 outperforms the previous scheme using OpenPose

and C3D blocks of 0.79. In addition, the proposed two-stream violence detector outperforms the previous scheme within all 4 crowd-size cases.

Acknowledgment

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT), Development of 5G-Based Predictive Visual Security Technology for Preemptive Threat Response under Grant 2019-0-00203.

References

1. Philip, Jobin T., Binoshi Samuvel, K. Pradeesh, and N. K. Nimmi. "A comparative study of block matching and optical flow motion estimation algorithms." In 2014 Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD), pp. 1-6. IEEE, 2014.HRNet
2. Sun, Ke, Bin Xiao, Dong Liu, and Jingdong Wang. "Deep high-resolution representation learning for human pose estimation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693-5703. 2019.
3. Osokin, Daniil. "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose." arXiv preprint arXiv:1811.12004 (2018).
4. Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems 27 (2014).
5. Zhao, Yuxuan, Ka Lok Man, Jeremy Smith, Kamran Siddique, and Sheng-Uei Guan. "Improved two-stream model for human action recognition." EURASIP Journal on Image and Video Processing 2020, no. 1 (2020): 1-9.
6. Chen, J., Xu, Y., Zhang, C., Xu, Z., Meng, X., & Wang, J. (2019). An Improved Two-stream 3D Convolutional Neural Network for Human Action Recognition. In H. Yu (Ed.), 2019 25th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing [8894962] IEEE.
7. Stergiou, Alexandros, and Ronald Poppe. "Spatio-temporal FAST 3D convolutions for human action recognition." In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp. 183-190. IEEE, 2019.
8. Soliman, Mohamed Mostafa, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. "Violence recognition from videos using deep learning techniques." In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 80-85. IEEE, 2019.
9. Zhang, Yu-Jin. "Motion analysis." In *Handbook of Image Engineering*, pp. 1127-1164. Springer, Singapore, 2021.
10. Maas, Kalvin, Erwin M. Bakker, and Michael S. Lew. "Full-Body Action Recognition from Monocular RGB-Video: A multi-stage approach using OpenPose and RNNs." PhD diss., BSc Thesis, Leiden University, 2020.
11. Rathod, Vatsal, Rishvanth Katragadda, Saurabh Ghanekar, Saurav Raj, Pushyamitra Kollipara, I. Anitha Rani, and A. Vadivel. "Smart surveillance and real-time human

action recognition using OpenPose." In *ICDSMLA 2019*, pp. 504-509. Springer, Singapore, 2020.
12. Ji, Shuiwang, Wei Xgu, Ming Yang, and Kai Yu. "3D convolutional neural networks for human action recognition." IEEE transactions on pattern analysis and machine intelligence 35, no. 1 (2012): 221-231.
13. Ding, Chunhui, Shouke Fan, Ming Zhu, Weiguo Feng, and Baozhi Jia. "Violence detection in video by using 3D convolutional neural networks." In *International symposium on visual computing*, pp. 551-558. Springer, Cham, 2014.
14. Li, Ruoyu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. "Adaptive graph convolutional neural networks." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1. 2018.
15. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

Crop leaf image classification and performance comparison using deep learning

Ki-tae Park[1], Dong-kyu Yun[1], Sang-hyun Choi[2*]

[1]Dept. Bigdata, Chungbuk National University, Cheongju, South Korea
[2]Dept. Management Information System, Chungbuk National University, Cheongju, South Korea
{dongkyu.yun, chois}@cbnu.ac.kr, {rlxogustn}@naver.com

Abstract. Recently, Korea is more aging. Also, the agricultural population in Korea is rapidly decreasing. Moreover, rapid climate change and falling grain self-sufficiency are getting worse. To solve this problem, research on smart farm-related technologies is being actively conducted. Especially, big data and image processing technologies can be used to grow crops scientifically and efficiently. In this study, images of the five crops were collected, and the collected images consisted of three datasets according to the data preprocessing method. The classification accuracy was compared using three CNN-based image classification models for each dataset. As a result, using dataset B and NASNetLarge models showed the highest classification accuracy.

Keywords : Smart Farm, Crop Leaf, Image Classification, Deep Learning

1 Introduction

Prior to the introduction of smart farm technology, it was subjective and intuitive, relying on the experiences and senses gained from growing crops, but it was quantified and objectified based on sensors and network technologies. Therefore, agricultural decisions made by repetitive trial and error and individual know-how have become more scientific and convenient with big data and artificial intelligence computers. In addition, with the development of image processing technology using big data collected while operating smart farms, growth management of crops and livestock diseases or pests can be performed in the early stages to reduce damage and increase production.

2 Data & Analytics

The leaf images of strawberries, cucumbers, corn, peppers, and rice were taken and collected directly from each farm, and additional images were collected by web crawling to match the diversity of images and the number of insufficient images.

* Corresponding author

Finally, 2,000 images per crop, total of 10,000 images were used for crop leaf classification. In addition, the collected data was preprocessed in three ways as shown in Table 1. The classification accuracy was compared using image classification models NASNetLarge, EfficientNetB7, and VGG16 on the preprocessed dataset.

Table 1. datasets were preprocessed in three ways.

Dataset	Data preprocessing	
A	1) Change image quality (50% less) 2) Image Resize (Apply differently depending on the model) 3) rescale=1./255	
B	1) Equivalent to Dataset A 2) shear_range=0.2 3) zoom_range=0.2	4) rotation_range = 40 5) fill_mode='nearest' 6) horizontal_flip = True
C	1) Equivalent to Dataset B 2) OpenCV Canny algorithm	
Train data : Test data = 9 : 1		

The overall results of classification accuracy can be found in Table 2. First, as a result of comparing the classification accuracy by dataset unit, it was confirmed that dataset B with five image augmentation techniques was high overall. Also, the classification accuracy was higher than dataset A and dataset C. Next, as a result of comparing by image classification models, NASNetLarge was confirmed that the accuracy was higher than EfficientNetB7, and VGG16. Therefore, when classifying leaf images of crops, using dataset B and NASNetLarge is the most ideal for classification accuracy.

Table 2. comparison of Image classification accuracy by datasets and models.

Dataset / Model	A (Acc,%)	B (Acc,%)	C (Acc,%)	Average (Acc,%)
VGG16	88.31 %	86.40 %	79.51 %	82.96 %
EfficientNetB7	95.53 %	96.35 %	85.49 %	92.46 %
NASNetLarge	95.12 %	96.71 %	89.62 %	93.82 %

3 Conclusion

The implications of this study are that the dataset was constructed by collecting crop images directly from farmers for use in actual smart farm sites, and the classification accuracy was verified through CNN-based three classification models. Especially, in dataset C, contour detection algorithm was used in consideration of the leaf characteristics of crops with similar contours between the same or similar species. Finally, as a result of this analysis, the average of the overall classification accuracy was about 90%, which was higher than expected. These results are expected to play a key role in the development of next-generation Korean smart farms.

End-to-end Multimodal Transformer Fusion for Video Emotion Recognition

Hoai-Duy Le, Hyung-Jeong Yang*, Soo-Hyung Kim, Guee-Sang Lee, Seok-Bong Yoo, Ngoc-Huynh Ho, Sudarshan Pant

Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea

hoaiduy1396@gmail.com, {hjyang, shkim, gslee, sbyoo}@jnu.ac.kr, {nhho, sudarshan}@chonnam.ac.kr

Abstract. Recently, with the raising of short video social media platforms such as TikTok, Facebook Reels, and YouTube Shorts, video emotion recognition has become an active research area. The main challenge in video emotion recognition is how to effectively combine the information from different data sources including video frames, voice signals, and subtitles. In this paper, we present a method to fuse multimodal information from language, image, and audio modalities for video emotion recognition. Specifically, we leverage the multi-head attention mechanism in the transformer encoder to align and integrate the features extracted from multiple modalities of the raw input video. Our method shows improvements compared to other approaches for the video emotion recognition task on two benchmarks including IEMOCAP and CMU-MOSEI.

Keywords: multimodal fusion, transformer, video emotion recognition

1 Introductions

A few years ago, the global rise of short video platforms including TikTok, Facebook Reels, and Instagram Reels led to the explosion of short video content. Online videos have become the major data users use to share their activities and interact with each other in cyberspace. Consequently, video data analysis has attracted significant attention from both industrial companies and science communities [10, 11]. One of the most popular applications of video data analysis is video sentiment analysis and emotion recognition [12].

Naturally, humans express their emotions in various manners via language, voice, and facial expression. Therefore, to truly understand human emotions, integrating information from multiple modalities is crucial. Various existing works have tried to explore effective fusing strategies for multimodal video sentiment analysis and emotion recognition. For example, Chen et al. [16] design a gate controller to filter noisy modalities before inputting fused features into a temporal attention-based LSTM (long

* Corresponding author.

short-term memory) network. Wei et al. [13] introduce the Bi-Bimodal Fusion Network (BBFN) focusing on pairwise bimodal fusion strategy by learning text-related representation from text-visual and text-audio pairs. Delbrouck et al. [14] present TBJE that leverages the Transformer encoder to reduce the computational cost and extract joint information from linguistic and acoustic modalities for emotion recognition. Also, Wei et al. [15] present a hierarchical mutual information maximization method to maximize inter-modality mutual information. These mentioned studies have proven effectiveness in improving performance compared to unimodal methods. But they still have a common limitation is that their input data is hand-crafted features. The hand-crafted features are fixed during the training process; therefore, the model performance will heavily rely on selecting extraction algorithms to obtain appropriate inputs.

In this paper, inspired by Dai et al. [1], we introduce a simple and effective Transformer-based method for emotion recognition from raw input videos. Our method is trained from raw videos in an end-to-end manner and further improves the emotion recognition model performance compared to other existing methods. Transformers [3] were first proposed in natural language processing (NLP) for the task of machine translation and since have been extended to computer vision (CV) [7,8], speech recognition [9], and many other fields [17, 18]. The multi-head attention mechanism in the transformer can be learned to discover the relevant words in a sentence or the important parts in an image that need to be enhanced to the prediction outputs. By using Transformer encoders, our model captures temporal information from the extracted high-level features in each modality and then performs multimodal fusion to predict emotion outputs. We evaluate our method on two standard benchmarks including IEMOCAP and CMU-MOSEI. The extensive experiments verify the effectiveness of our approach compared to previous methods.

The paper is organized as follows: we present the details of our method in Section 2. Then we describe the extensive experiments and experimental results on two datasets in Section 3. Finally, we conclude the paper in Section 4.

2 Methodology

In this section, we present our proposed method. The raw video clips containing video frames, audio signals, and subtitles are fed into the network as input data, and the model outputs the emotions for the whole video. We first present the overall architecture of the approach (Sec. 2.1). Then we introduce the details of each module (feature extraction in Sec. 2.2, fusion module in Sec. 2.3, and classification head in Sec. 2.4).

2.1 Overview

Figure 1 shows the architecture of our proposed method. Our approach consists of three modules including feature extraction, fusion module, and classification head. First, each modality goes through a separate feature extractor to extract high-level features. Because video is sequential data, the Transformer encoders are used to capture the

temporal information from each modality. After that, three sequences of features and three modality-specific features are derived for three modalities. Then for the fusion module, to produce the multimodal mutual representation, the Transformer encoder is adopted to learn the correlation between multimodal features from the concatenated sequence of three modalities' sequences of features. Finally, a feed-forward layer a utilized over the combination of three modality-specific features and multimodal mutual features to make the emotion prediction.

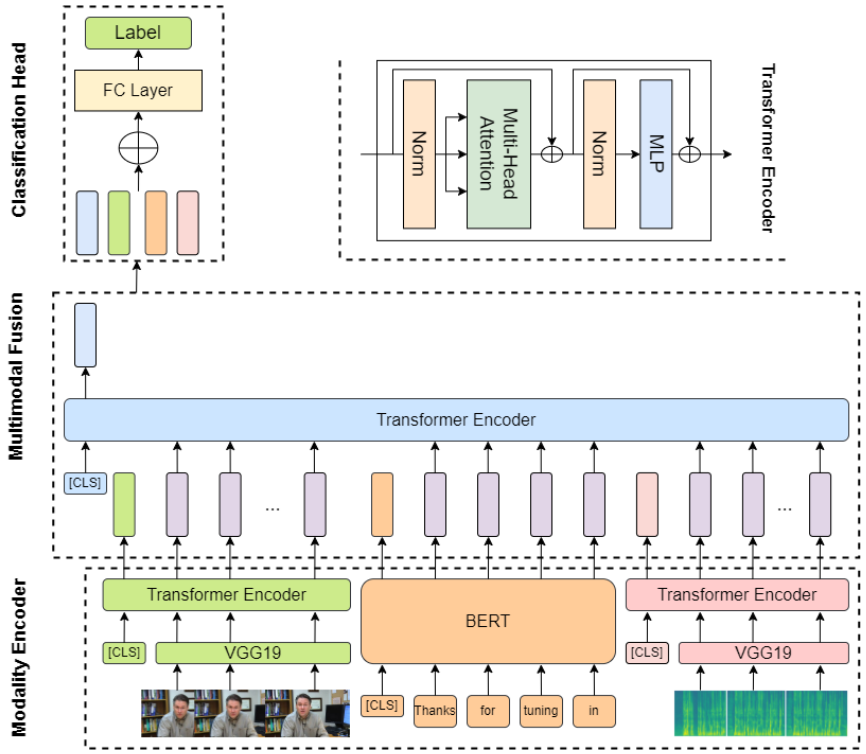


Fig. 1. The overall architecture or our proposed method.

2.2 Feature Extraction

For textual modality, we use the BERT-based model to preprocess and extract word embeddings. We take the output of the [CLS] token as the textual representation and the rest of the BERT output as the sequence of word embeddings for further steps. For audio modality, we first transform the audio signal to log-mel spectrogram and split it into a sequence of spectrogram pieces. For visual modality, rather than using full video frames, we leverage the pre-trained MTCNN [2] to detect and crop the face region in the frame. Then we utilize two individual CNN networks to extract visual features from the sequences of spectrograms and video frames. Furthermore, we use Transformer

encoders to capture the sequential information of audio and visual modalities. We also obtain the [CLS] position tokens as visual and audio representations.

Generally, we adopt the same input processing steps and feature extraction as Dai et al. work [1]. The reason for this is to emphasize the effects of multimodal fusion on the model performance and fair comparison with other works. Using more sophisticated pre-trained language models and CNN backbone may enhance the performance. In this work, we mainly focus on the fusion technique.

2.3 Transformer-based Fusion Module

Our objective is to learn the joint multimodal representation of the video data to enhance the model performance. To this end, we adopt the Transformer encoders at multiple layers to fuse the multimodal features. The multi-head attention mechanism in the transformer allows each token to attend to each other in the input sequence and aggregate information from the entire sequence. It naturally provides a mechanism to model the interaction between multimodal features and learn the joint representation of the sequence of multimodal features.

After obtaining three multimodal sequences of features outputted from the feature extraction module, we concatenate them into a single sequence. Like the ViT model [7], we prepend a specific learnable [CLS] token to the sequence as this token's output can be served as the joint representation of the whole input sequence. Next, we fed this sequence to a series of transformer encoders. We adopt the standard architecture of the transformer in our work. Given an input sequence $I \in \mathbb{R}^{n \times d}$, the transformer first projects I h times into $Q_i \in \mathbb{R}^{n \times d_k}$, $K_i \in \mathbb{R}^{n \times d_k}$, $V_i \in \mathbb{R}^{n \times d_k}$, $i \in 1, \dots, h$. Then the scaled dot-product attention function is computed for each head as:

$$Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d_k}}\right) V_i \tag{1}$$

After that, the multi-head attention function is the concatenation of all attention heads:

$$\begin{aligned} MultiheadAttention(Q, K, V) &= concat(A_1, \dots, A_h) \\ \text{where } A_i &= Attention(Q_i, K_i, V_i) \end{aligned} \tag{2}$$

Finally, we take the output of the [CLS] token as the joint multimodal embedding for further steps.

2.4 Emotion Classification Head

We concatenate three modality-specific representations with the joint multimodal representation and feed it into a feed-forward layer. Because of the multilabel classification task, we use the sigmoid activation function to predict the emotion labels and use Binary Cross-Entropy (BCE) as the loss function to train the model:

$$\mathcal{L}_{train} = BCE(y, \hat{y}) = \sum y * \log(\hat{y}) + (1 - y)\log(1 - \hat{y}) \quad (3)$$

3 Experiments

In this section, we present our experiment on two benchmark datasets (Section 3.1). We then describe our experiment setting (Section 3.2). After that, we present our ablation study and results compared to other methods (Section 3.3).

3.1 Dataset and Evaluation Metrics

We conduct the experiment on two benchmark datasets including CMU-MOSEI and IEMOCAP. Specifically, for full end-to-end learning, we adapt the reorganized version of both datasets from Dai et al., 2021 work. The detail of restructured versions is described below, and the statistics are shown in Table 1 for both datasets.

The CMU-MOSEI dataset is labeled by six emotion categories: happy, sad, angry, fearful, disgusted, and surprised. The dataset contains 20477 movie review video clips from YouTube in total which is divided into training, validation, and testing set of 14524, 1764, and 4188 data samples, respectively.

The IEMOCAP dataset contains the annotation of six emotions: angry, happy, excited, sad, frustrated, and neutral. The dataset is split at the utterance level and consists of 7380 videos in total. Each video records an actor during a dyadic conversation in English with the other. For the learning process, the dataset is randomly portioned 70%, 10%, and 20% into the training, validation, and testing set, respectively.

Table 1. Statistic of CMU-MOSEI and IEMOCAP used in the experiments.

Dataset	Modality	#sample	#training	#validation	#testing
CMU-MOSEI	{l, v, a}	20477	14524	1765	4188
IEMOCAP	{l, v, a}	7380	5162	737	1481

According to prior works, we use the standard accuracy and F1-score as evaluation metrics for the IEMOCAP dataset. For the CMU-MOSEI dataset, we use weighted accuracy and F1-score to evaluate the model. The weighted accuracy is calculated by:

$$WAcc = \frac{TP \times N/P \ + TN}{2N} \quad (4)$$

where TP means total true positive, N total negative, P total positive, and TN true negative. Due to the imbalance dataset problem, weighted accuracy and F1-score are appropriate metrics to evaluate the model performance.

3.2 Implementation Detail

We use Pytorch to implement our model. Following Dai et al. work [1], we use a pre-trained ALBERT-base from the Hugging Face library to extract textual features. We adopt the same preprocessing steps for video frames and audio signals. Specifically, the video frames are sampled every 0.5s and are fed to the pre-trained MTCNN [2] to obtain the sequence of faces. The audio signals are transformed into Mel Spectrogram and are divided into sequences of spectrograms by 400ms time window. We construct two separate VGG-19 networks and train from scratch as feature extractors for audio and visual modality. We utilize the standard Transformer encoder as [3]. We train the model for 30 epochs with Adam optimizer [4] setting the initial learning rate to 5e-5 for all parameters except ALBERT weights with 5e-6.

3.3 Experimental Results

Table 2. Experimental results on IEMCAP and CMU-MOSEI datasets.

Modality	Model	IEMOCAP		CMU-MOSEI	
		Acc	F1 score	WAcc	F1 score
Multimodal	EmoEmbs [5]	72.0	49.8	64.2	44.2
	MulT [6]	77.6	56.9	65.4	45.2
	FE2E [1]	85.65	57.14	65.84	47.03
	Transformer-based Fusion **(Ours)**	**86.12**	**60.07**	**67.36**	46.84

Table 2 illustrates the results of our method compared to other multimodal approaches on both IEMOCAP and CMU-MOSEI datasets. We compare our model with multimodal approaches received hand-crafted input data including EmoEmbs [5], MulT [6], and FE2E [1] trained from raw input data. As shown in the table, our network outperforms previous multimodal methods on both two datasets. Furthermore, the end-to-end manner performs better than the two-stage pipeline with hand-crafted input features.

Table 3. Ablation study results

Modality	Model	IEMOCAP		CMU-MOSEI	
		Acc	F1 score	WAcc	F1 score
Textual	ALBERT	82.52	53.99	64.34	45.77
Visual	CNN	80.86	53.12	57.59	38.84
Audio	CNN	53.63	37.74	57.25	38.77
Multimodal	Early Fusion	84.72	57.19	65.49	45.49
	Transformer-based Fusion **(Ours)**	**86.12**	**60.07**	**67.36**	**46.84**

In addition, we conduct an ablation study to clarify the influence of the multimodal fusion module. The results are shown in Table 3. First, the unimodal method for each modality is evaluated separately. Then we simply apply the early fusion strategy that concatenates three modality embeddings to predict the emotions. The results are much improved compared to unimodal models. Finally, we add the transformer-based fusion module, the performances are significantly enhanced. In CMU-MOSEI, the WAcc reaches 67.36% from 65.49% and the F1 score achieves 46.84% from 45.49%. The Acc and the F1 score goes up from 84.72% and 57.19% to 86.12% and 60.07% in IEMOCAP, respectively.

4 Conclusion

In this paper, we presented a Transformer-based method for video emotion recognition. We leveraged the multi-head attention in the Transformer encoder to fuse the features from different modalities. By combining fused representation and modality-specific representations, the model performance is significantly improved. Experiment results showed that our approach achieved better performance compared to other existing methods on two benchmarks including IEMOCAP and CMU-MOSEI.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (NRF-2020R1A4A1019191)

References

1. Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal End-to-End Sparse Model for Emotion Recognition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5305–5316, Online. Association for Computational Linguistics.
2. Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
4. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization.arXiv preprint arXiv:1412.6980,2014.
5. Wenliang Dai, Zihan Liu, Tiezheng Yu, and Pascale Fung. 2020a. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition.ArXiv, abs/2009.09629.
6. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P. and Salakhutdinov, R., 2019, July. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for Computational Linguistics. Meeting (Vol. 2019, p. 6558). NIH Public Access.
7. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly,

et al.An image is worth 16x16 words: Transformers for image recognition at scale.arXiv preprintarXiv:2010.11929, 2020.
8. Chen, C.F.R., Fan, Q. and Panda, R., 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 357-366).
9. Yuan Gong, Yu-An Chung, and James Glass. AST: audio spectrogram transformer.arXiv preprint arXiv:2104.01778, 2021.
10. Lei Zhang, Shuai Wang, and Bing Liu. 2018a. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4): e1253.
11. Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Ming-hao Yin. 2019. A survey of sentiment analysis in social media. Knowledge and Information Systems,60(2):617–663.
12. Ramandeep Kaur and Sandeep Kautish. 2019. Multi-modal sentiment analysis: A survey and comparison. International Journal of Service Science, Management, Engineering, and Technology (IJSSMET),10(2):38–58.
13. Han, Wei, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis." In Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 6-15. 2021.
14. Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. In Second Grand Challenge and Workshop on Multimodal Language (Challenge-HML), pages 1–7, Seattle, USA. Association for Computational Linguistics.
15. Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
16. Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A. and Morency, L.P., 2017, November. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM international conference on multimodal interaction (pp. 163-171).
17. Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval.ICCV, 2021.
18. Cho, J., Youwang, K. and Oh, T.H., 2022. Cross-Attention of Disentangled Modalities for 3D Human Mesh Recovery with Transformers. arXiv preprint arXiv:2207.13820.

Posture Prediction using Bidirectional Relevance of Audio-Visual Data

So-Hyun Park[1], Shin-Hyeong Park[2], Young-Ho Park[3,*],

[1]Department of IT Engineering, Sookmyung Women's University, Seoul, 04310, South Korea

[2]Department of Software, Sookmyung Women's University, Seoul, 04310, South Korea

[3]Big Data Using Research Center, Sookmyung Women's University, Seoul, 04310, South Korea

{shpark, yhpark}@sm.ac.kr, shpark0308@sk.com, *Corresponding Author

Abstract. Existing posture prediction methods based on audio-visual data suffer limitations because they do not consider several characteristics of a musical instrument player. To address this limitation, we propose a posture prediction model that consid-ers the bi-directional relevance of audio-visual data. For this, we first propose a data replacement strategy using a sound similarity matrix that substitutes abnormal data that can occur due to incorrect pose estimation. We then used the Bi-LSTM model to learn bi-directional relevance to predict posture. The experimental results show the effectiveness of our method.

Keywords: Posture Prediction, Audio-Visual Data, Bi-LSTM

1 Introduction

Several studies have been conducted on predicting posture using the relationship between sound and posture in the music domain. As a representative study, Shlizerman et al. proposed a method to predict posture using an LSTM deep learning network. Specifically, the proposed method improves posture prediction accuracy by learning the relationship between sound and posture. However, it is difficult to use the proposed method as a general solution for posture prediction in the music domain because it cannot consider several characteristics of a musical instrument player, such as different playing styles for different music pieces. For example, consider person A who excessively uses his arms when playing music A, and person B, whose arms are fixed when playing the same music. In this case, if only the relation between sound and posture is considered, the distance between the test data and the predicted data may fall far apart, resulting in the decreased accuracy of posture prediction. Therefore, to construct a general performance posture prediction model, it is necessary to consider both the sound-to-posture and posture-to-sound relation.

Even with a bi-directional relationship between sound and posture in posture prediction, other issues (i.e., null or outlier values due to incorrect pose estimation) may occur. Previous studies solved this issue by removing null or outlier values or

replacing them with the intermediate value by mixing the result of several pose estimation techniques. However, removing the null or outlier values results in the sparsity of the pose estimation dataset. It also makes it difficult to build a natural posture prediction model similar to that of a real performer. In addition, replacing the null or outlier values with the intermediate value by mixing the result of several pose estimation techniques may cost a lot of time.

This paper proposes a posture prediction model considering the **bidi**rectional **r**elevance between **a**udio-**v**isual data (BidiRAV). For this, we used the Bi-LSTM model for make a general performance posture prediction model by learning the diversity of postures. In addition, we propose a null and outlier values replication strategy by constructing an audio-visual similarity matrix. The main idea of the proposed data replication strategy is that it can fill invalid values by inspecting similar sounds and replacing them with the most similar ones. The experimental results show that our model can effectively improve posture prediction accuracy with the proposed guidance of **bidi**rectional **r**elevance between **a**udio-**v**isual data and data replication strategy.

2. Our Model

Dataset. We first collected the playing the piano video data from the Youtube platform. Specifically, we collected videos of three different piano performances, namely Bach Invention No.2, Chopin Etude Op.10 No.1, and Chopin Polonaise Op.26 No.1. The total length of the collected videos is 705 seconds. We obtained a total of 2115 images that were used for experiments.

Feature Representation. We then extracted visual features from the collected images. For this, we extracted the skeleton coordinate values of hand posture P using a pre-trained Mediapipe model. The extracted hand posture joints include a total of 20 joints, including four joints for each finger and wrist joint. Here, each joint contains three-dimensional data of x, y, and z. Further, we extracted audio features from collected videos. For this, we cut the videos into one-second videos and extracted Mel-Frequency Cepstral Coefficient (MFCC) features denoted as M from them. Here, the MFCC sampling rate was 44100.

Figure 1. An example of pose estimation result

Posture Outlier Replacement Module. Due to the continuity of audio while playing a musical instrument, it is possible to extract features without interruptions. However, null and outlier values may occur in visual features when pose estimation fails to extract the needed skeleton. For example, there are certain cases when the

pose estimation cannot detect a hand posture when the video is taken from the top (See Figure 1). We propose a null and outlier values replication strategy to solve this issue by constructing an audio-visual similarity matrix.

Posture Prediction Bi-LSTM-based model $\mathrm{BidiRAV}(P_t, M_{t+1})$ is constructed to predict posture considering the bidirectional relationship between sound and posture. By giving P_t as the input value and M_{t+1} as the result value, the bidirectional

relationship between the two modals is learned, and then posture prediction is performed.

Figure 1. An example of pose estimation result

3.Experiment

We compare our model with several posture prediction models, including Li et al. and Shlizerman et al. Since the posture was predicted using LSTM in both models, the methods of the two papers were implemented as one and compared with the proposed BidiRAV.

We evaluate the BidiRAV performance in two subtasks: (1) Changes in posture prediction accuracy according to the percentage of missing values, and (2) Changes in prediction accuracy when the bi-modal data interrelationship is considered. In addition, we measured the performance by measuring the RMSE, which calculates the error rate between the predicted value and the correct value.

We first report the experiment results related to the effect of the data replication strategy on posture prediction, as shown in Table 1. For the sake of experiments, the percentage of null and outlier values was arbitrarily adjusted to measure the accuracy of posture prediction. From Table 1, we can observe that the proposed data replacement strategy can effectively eliminate null and outlier values and improve the accuracy of posture prediction.

The second experiment examines the effect of the bi-directional relationship between sound and posture on posture prediction accuracy. As shown in Table 2, we can observe that the accuracy of posture prediction of the proposed Bidi-RAV has \

Table1. Experiment on the effect of data replacement strategy on the accuracy of posture prediction

	Null data percentage			
	0.01	0.02	0.03	0.04
Rmse	0.1871	0.2052	0.2240	0.2348

Table2. Experiment on the effect of bi-directional relevance consideration on posture prediction accuracy

Num of music	LSTM	BidiRAV (Ours)
1	0.2557	0.2496
2	0.1237	0.1211
3	0.1962	0.1871

4.Conclusion

In this paper, we proposed BidiRAV, a model that advances posture prediction accuracy by considering the bidirectional relationship between sound and posture. To do this, we first find a substitute for postural outliers in the sound similarity matrix with the data replacement strategy. Afterward, the Bi-LSTM model was used to learn bi-directional relevance to predict posture. The experimental results depict that the proposed method achieves considerable improvement compared to existing models.

Acknowledgments. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology [NRF-2021R1C1C2004282].

References

[1] Shlizerman, E.; Dery, L.; Schoen, H.; and Kemelmacher-Shlizerman, I. 2018. Audio to Body Dynamics. In Proceed-ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[2] Li, B.; Maezawa, A.; and Duan, Z. 2018. Skeleton Plays Piano: Online Generation of Pianist Body Movements from MIDI Performance. In Proceedings of theAudio to Body Dynamics. In Proceedings of the International Socie-ty for Music Information Retrieval (ISMIR).

[3] Zhang, F., Bazarevsky, V., Vakuov, A., Tkachenka, A., Sung, G., Chang, C. L., and Grundmann, M. 2020. Medi-apipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214.

Estimation of Machine Health Stability using Deep Learning

Dimang Chhol, Sunghoon Kim, Kwan-Hee Yoo

Computer Science Department, Chungbuk National University, South Korea
{dimangchhol, sidsid84, khyoo}@chungbuk.ac.kr

Abstract. Knowing the upcoming trend of machine stability in the smart manufacturing industry is necessary to produce a better quality product. After analyzing data every thirty minutes, we get essential features critical for predicting future machine health stability. This research study used various deep learning techniques, including LSTM, GRU, and LSTM, with GRU architecture.

Keywords: Machine Health Stability, Smart Factory, LSTM, GRU

1 Introduction

Machine Health assessment is essential in today's revolution industry 4.0 [2]. Knowing the essential feature that correlates with machine productivity prediction in the smart factory enables manufacturers to understand and prepare for the best product outcome. Thus our goal is to contribute with an extension of existing machine health stability with machine learning [1] with deep learning. Based on [1], the author defines Machine Health Stability as: "*Machine Health Stability is the reliability of equipment in its actual life cycle based on productivity and machine-self data*."

This paper aims to understand the product outcome's reliability and the machine's current status in smart manufacturing systems for increasing the outcome, which is essential for the company. We used deep learning models such as GRU, LSTM, and LSTM+GRU. The experimental results found that LSTM outperforms GRU, LSTM+GRU, and SVM Linear.

2 Related Studies

We review the previous study on feature selection and prediction techniques using machine learning or deep learning. Borith et al. 2020 [4] presented a technique for predicting a machine's non-active state with statistical techniques using machine learning for extraction features. They found that linear SVM performs better than other experiment models(Decision Tree, KNN, Random Forest). Pheng et al. 2022 [3] used statistical analysis and a deep learning technique to predict process quality performance in the manufacturing industry. The authors found that LSTM got higher

accuracy than ANN. Apart from this, Bakhit 2021 [1] proposed a machine-learning algorithm to predict machine health stability. The author found that SVM performs better than other machine learning methods, including Ridge, Random Forest, and K-NN. Besides that, S. Khan and T. Yairi, 2018 [5] reviewed various conference papers and journal articles from 2013 to 2017 on a different method of system health management with deep learning. They found the most suitable deep learning for fault diagnosis and Machine health monitoring is RNN, LSTM, and GRU [6], [7], [8], [9].

We improved the previous work by adapting the work of Borith et al. 2020 [4] with Pheng et al. 2022 [3] and extending from Bakhit 2021 [1] by enhancing the prediction of MHS with deep learning and improving the data process technique.

3 Proposed Methods

Based on Bakhit, 2020 [1], to calculate the MHS, we used the following formula:

$$\text{MHS} = \omega_1 \times Rate_{alarm} + \omega_2 \times Rate_{non-active} + \omega_3 \times Rate_{defective} + \omega_4 \times Rate_{NG\ product} + \omega_5 \times Probability_{fault} \quad (1)$$

Where $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5$ is the failure probability value equivalent to 0.1, 0.2, 0.3, 0.4, and 0.5, respectively. The level of machine health stability(MHS) defines that from 0.85 to 1.00 is excellent, 0.75 to 0.85 is good, 0.65 to 0.75 is average, 0.5 to 0.65 is poor, and 0.0 to 0.5 is terrible.

In Fig. 1, we demonstrate our global structure's flow chart: collecting data, extracting data, and using deep learning to estimate machine health stability.

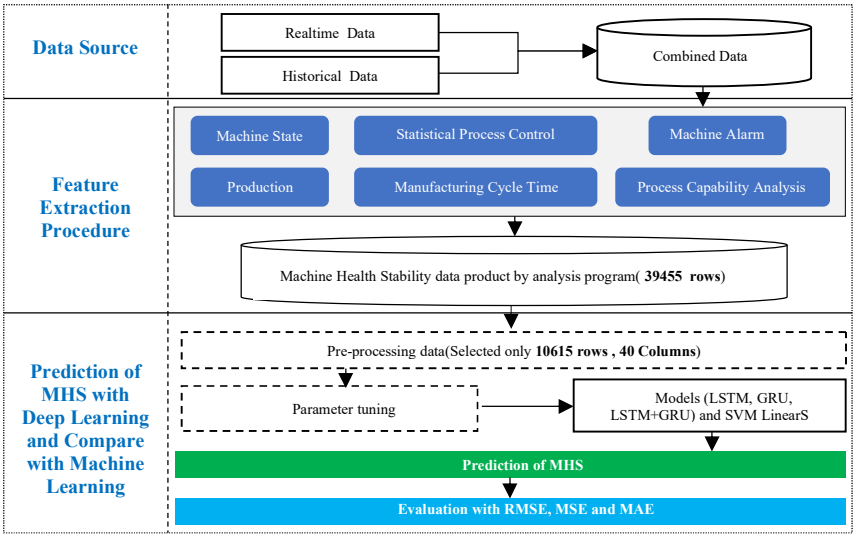


Fig. 1 Proposed Method adapted from Bahit [1], 2020, Pheng et al. 2022 [3], Borith et al, 2020 [4]

4 Experiment Result

In this research study, we gathered and analyzed data sets from January 2020 to September 2022, consisting of **39455 rows**. We selected one type of variable for our experiment, which consisted of **10615 rows** and 40 Columns. As a result, our data consist of Terrible quality 1, poor quality 95, average quality 3649, good quality 5295, and excellent quality 1575. We also chose specific variables to create a deep learning and machine learning model.

We compare the LSTM model with GRU, LSTM+GRU, and the machine learning model Linear SVM to demonstrate the proposed method. First, we split data for training 70% and testing 30%. Then We define the configuration parameters of LSTM, GRU, LSTM+GRU with activation "relu", training epochs 25, the loss function is "mse" and optimizer is "adam" . Finally, the comparison output result of our model shows in Table 1. Our experiment results show that LSTM has the highest accuracy compared to GRU, LSTM+GRU, and Linear SVM.

Table 1. Result of model comparison(test data)

Model	RMSE	MSE	MAE
LSTM	0.013778	0.000189	0.008653
GRU	0.018798	0.000353	0.009777
LSTM+GRU	0.014009	0.000196	0.009291
Linear SVM	0.016227	0.000263	0.006736

Fig. 2a shows the result of the prediction MHS with LSTM with test data, while Fig. 2b shows the result of the prediction MHS with GRU with test data, and Fig. 2c shows the result of the prediction MHS with LSTM+GRU model with test data. Finally, fig. 2d shows the result of the prediction MHS with Linear SVR model with test data.

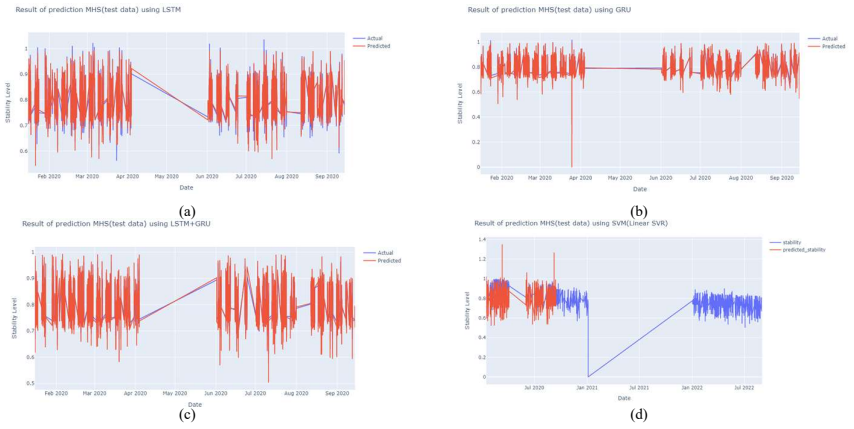


Fig. 2 Result of prediction of MHS; (a) using LSTM (b) using GRU; (c) using LSTM+GRU; (d) using SVM(Linear SVR)

5 Conclusion and Future Work

In this research study, we adapted and extended the data preparation from the previous research in the smart manufacturing industry. We work on implementing a thirty-minute analysis program, which is the main point of data preparation. It combines various statical process analyses, process capability analyses, conditional variables, defective product rate, and manufacturing cycle time to produce a fault prediction data set. In addition, we experiment with various deep learning models, showing that LSTM performs better than GRU, LSTM+GRU, and Linear SVM. In the future, we will work on defining new essential features that affect the outcome of MHS. Moreover, we plan on integrating the user interface for our current web-based system into the company.

Acknowledgments. This research was supported by the MSS(Ministry of SMEs and Startups), Korea, under the Cloud-based customized smart factory big data platform research and development support program(S3290113) supervised by the TIPA(Korea Technology and Information Promotion Agency for SMEs) by the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.

References

1. Bakhit, S.: Prediction of Machine Health Stability in Smart Factory Systems using Machine Learning. Department Of Computer Science, vol. Master. Chungbuk National University (2021)
2. Lee, G.-Y., Kim, M., Quan, Y.-J., Kim, M.-S., Kim, T.J.Y., Yoon, H.-S., Min, S., Kim, D.-H., Mun, J.-W., Oh, J.W.: Machine health management in smart factory: A review. Journal of Mechanical Science and Technology 32, 987-1009 (2018)
3. Pheng, T., Chuluunsaikhan, T., Ryu, G.-A., Kim, S.-H., Nasridinov, A., Yoo, K.-H.: Prediction of Process Quality Performance Using Statistical Analysis and Long Short-Term Memory. Applied Sciences 12, (2022)
4. Borith, T., Bakhit, S., Nasridinov, A., Yoo, K.-H.: Prediction of Machine Inactivation Status Using Statistical Feature Extraction and Machine Learning. Applied Sciences 10, (2020)
5. Khan, S., Yairi, T.: A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing 107, 241-265 (2018)
6. Yuan, M., Wu, Y., Lin, L.: Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network. In: 2016 IEEE International Conference on Aircraft Utility Systems (AUS). IEEE, (2016)
7. Gugulothu, N., TV, V., Malhotra, P., Vig, L., Agarwal, P., Shroff, G.: Predicting Remaining Useful Life using Time Series Embeddingsbased on Recurrent Neural Networks. arXiv preprint (2017)
8. Zhao, R., Yan, R., Wang, J., Mao, K.: Learning to Monitor Machine Health with Convolutional Bi-Directional LSTM Networks. Sensors 17, 273 (2017)
9. Malhotra, P., TV, V., Anand, A.R., Vig, L., Agarwal, P., Shroff, G.: Multi-Sensor Prognostics using an Unsupervised HealthIndex based on LSTM Encoder-Decoder. arXiv preprint (2016)

Predicting Outlier Particle in a Cleanroom Semiconductor using Deep Learning Techniques

Saksonita Khoeurn[1], Munirot Thon[1], Lina Maria Cuervo Diaz[1], Bunroth Sok[1,2], Jae Sung Kim[1], Wan Sup Cho[1,2]

[1] Department of Big Data, Chungbuk National University, Cheongju, South Korea
[1]{saksonita, munirot.thon, l.cuervod, comkjsb}@cbnu.ac.kr
[2]{bunroth.sok, wscho63}@gmail.com

Abstract. In semiconductor manufacturing, the number of microscopic particles in the room can influence the quality of the product. Keeping the cleanliness of the cleanroom is required to reduce the risk of product failures and losses and maintain the quality of production. Hence, predicting and managing the particles is crucial to optimizing the environment in semiconductor cleanrooms. In this research paper, deep learning models such as GRU, LSTM, and BiLSTM have been utilized to predict the outlier particles (sub-inflow particles) and fine dust anomalies to maintain cleanliness in the cleanroom. The results demonstrate that the models are efficient and capable of making a prediction for semiconductor cleanroom.

Keywords: deep learning, GRU, LSTM, semiconductor, cleanroom, time-series forecasting

1 Introduction

Semiconductors are everywhere and have grown tremendously. The cleanroom is a space designed to keep all equipment clean and free of contaminants that are necessary for industrial production and semiconductor manufacturing [1]. In order to achieve better results in semiconductor manufacturing and assembly, cleanrooms must have very high levels of environmental control because even the smallest speck of dust can damage semiconductor materials [2][3].

In order to prevent industrial production failure and downtime the massive number of airborne particles inside must be managed. Over-load amounts of particles might lead to product failure, that's why they must be removed on time, to prevent any problems that can occur during the production process [4]. Thus, prediction methods are recommended to notify the person in charge to make decisions [5].

Many academic and industry researchers have developed numerous algorithms to make predictions, so when selecting a forecasting algorithm, decision-makers must consider several factors of the prediction process, such as forecasting objectives, frequency, structure, data quality, etc. [6]. Therefore, this research strongly relies on the use of deep learning techniques like Gated Recurrent Unit (GRU), Long Short-

Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) for predicting the outlier particle in the cleanroom.

The paper is organized as follows: Section 2 describes the data and methodology of the research. Section 3 provides the results of the models based on two types of evaluation metrics, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Lastly, section 4, presents the concluding remarks.

2 Data and Methodology

The methodology used in this research is composed of various components including data collection, data preparation or data preparation, deep learning model building, and model evaluation. Fig 1 represents the flow chart for particle forecasting.

Fig. 1 Methodology Flow Chart.

2.1 Data Collection

The dataset used in this research was collected from a variety of sources, including installed sensors at the manufacturer and APIs, within the same period of one-minute intervals. The external weather data was collected using an API based on the manufacturer's location. The second set of data to be collected was the differential pressure data, which we obtained from sensors installed in each zone. Another type of sensor was used to collect the data on temperature, humidity, and particles (THP). Lastly, we collect the data for heating, ventilation, and air conditioning (HVAC) from a different kind of sensor implanted in each zone in the manufacturer. After combining all of the data mentioned, the total dataset of 40564 observations and 15 features is ready for exploratory data analysis (EDA).

2.2 Exploratory Data Analysis (EDA)

After plotting the data, we notice that particle 10 contains the most outliers in zone 100 which we will use as the target variable. An analysis of this particle was performed to see the time of the day and the total amount by each day of the week demonstrated in Fig. 2. The result shows that particle 10 obtained the most on Thursday with 237.630; meanwhile, the highest peak is from 3 pm to 4 pm during the day with a total of 141.070.

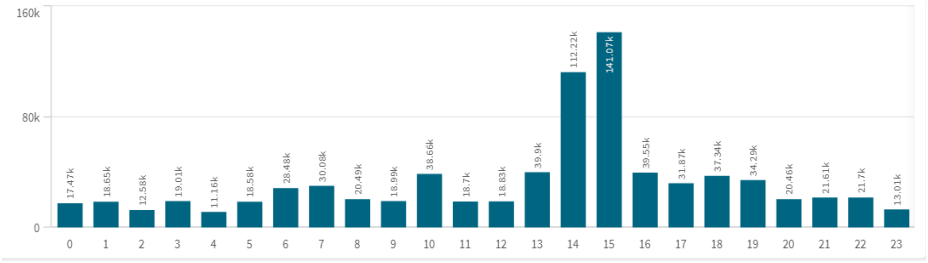


Fig. 2 Histogram of particle 10 frequency grouped by hours.

The Spearman Ranking analysis was conducted to check the correlation coefficient between the variables. It shows that the strongest correlation points according to the analysis are the steam pressure and the temperature y at the dew point temperature.

Relatively, humidity is one of the environmental conditions typically specified for cleanroom humidity control operations. Therefore, the humidity level requires extra care to guarantee that the level in the cleanroom does not alter.

2.3 Data Preparation

Based on the prior studies with the same data sources, it has been discovered that Zone 100 comprised the majority of the particle variables' outliers. Due to that, we will focus on the particle analysis based on zone 100. We selected the data from December 22nd to 28th, 2021, for this research as it contained the least missing value especially for the data of THP compared to the other time frames. The missing values were then filled with the forward fill method, which carries forward the last known value before the missing one, to fill in for that time frame.

The moving averages and lagging with 5mn, 15mn, and 60mn were converted and then added as exogenous variables among other pre-existing features, in order to perform the time-series analysis. The datasets with 9249 rows and 96 columns have been converted and split into training and testing sets with a portion of 80%–20%. As the GRU, BiLSTM, and LSTM take a 3-Dimensional input both the training and testing dataset has been reshaped before we can use it to fit into the model [7], [8].

2.4 Model Building

In this research, the time-series forecasting task was implemented using only three deep learning algorithms: GRU, LSTM, and BiLSTM of Keras's recurrent layer. Two hidden layers in each of the three models include 64 neurons and one additional neuron in the output layer. In order to make the LSTM and GRU networks more robust to changes, the Dropout function was also implemented by randomly removing 20% of the network's units [9]. The model trains 80% of the training data and the other 20% as validation data for 100 epochs and a batch size of 16.

3 Results

After training and evaluation, all three selected models produced similar results based on two evaluation metrics, MAE and RMSE. Table 1 summarizes the performances of models based on the mentioned evaluation metrics. According to the result from Table 1, the LSTM model has the lowest error value in term of MAE while BiLSTM has the lowest error value in term of RMSE.

Table 1. Performances of classifier models.

Model	MAE	RMSE
GRU	27.1119	36.2790
LSTM	19.8055	29.2045
BiLSTM	20.3488	27.5656

4 Conclusion

In conclusion, the goal of this research is to predict the sequence of the particle inside the semiconductor cleanroom using deep learning algorithms. The research used two methods to evaluate the performance of the selected models. As shown in the result section, the BiLSTM model surpasses the other models with the lowest average error value of MAE and RMSE. This research provides the most effective model to detect the outlier particle in the cleanroom which is a helpful guide for future research.

References

1. Whyte, W.: Cleanroom Technology: Fundamentals of Design, Testing and Operation. John Wiley & Son Ltd, United Kingdom (2001)
2. The Value of Cleanrooms for Semiconductor Application, https://angstromtechnology.com/
3. Dobler, M., Rüb, M., Billen, T.: Minienvironment solutions: special concepts for mask-systems. In: 27th European Mask and Lithography Conference, pp. 243-257. SPIE, (2011)
4. Case Study: Particle Detection Challenges in Pharmaceutical Cleanroom, https://www.chemtronics.com/
5. Key Elements of Contamination Control, https://www.cleanroom-industries.com
6. Mahmoud, E.: Accuracy in forecasting: A survey. Journal of forecasting 3, 139-159 (1984)
7. How to Reshape Input Data for Long Short-Term Memory Networks in Keras, https://machinelearningmastery.com/
8. Yang, B., Wang, S., Markham, A., Trigoni, N.: Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. International Journal of Computer Vision 128, 53-73 (2020)
9. Dropout Regularization in Deep Learning Models with Keras, https://machinelearning mastery.com/

Data Augmentation Method for Moiré Patterns of PCB Components

Taek-Lim Kim[1], Sung-Chul Yun[2], Tae-Hyoung Park[3*]

[1] Department of Control and Robot Engineering,
[2] Industrial AI Research Center,
[3*] School of Electronics Engineering, Chungbuk National University, Cheongju 28644, Korea
{taeglem, steveyun, taehpark*}@cbnu.ac.kr

Abstract. As a method of measuring the height of a PCB component, there is a method using the Moiré method. When generating a moiré image of a PCB component, a shadow area is generated due to kinematic problems, a reflection area is generated due to a material, and the moiré pattern is damaged. An error in height measurement occurs due to the damaged pattern. Networks can be used to solve these problems. However, the defective data in the inspection equipment is insufficient, so a data augmentation method is required. This paper proposes a data augmentation method considering the moiré pattern.

Keywords: Data Augmentation, Computer Vision, Moiré Pattern.

1 Introduction

Inspection equipment for PCB components plays a variety of roles. In general, it plays a role in detecting or classifying defects. Inspection equipment using a moiré pattern is used to measure height [1]. The height of the moiré pattern can be measured using three or more phase-shifted images. Figure 1 shows two problems to be addressed in this paper. In (a), the shadow is created by the location of parts and optics. As a result, the moiré pattern is damaged, and the pattern to be measured is not measured and appears as a black area, so the height cannot be measured. Figure 1 (b) is a case in which light is diffusely reflected without being reflected and returned by the material or shape [2].

However, there is a problem in that the moiré pattern is damaged due to the shadow area or material. Deep learning can be used to restore damaged images. However, deep learning requires a large number of images. The biggest problem with deep learning for inspection equipment is creating insufficient data. In this paper, we propose a learning data augmentation method considering the moiré pattern.

1
2

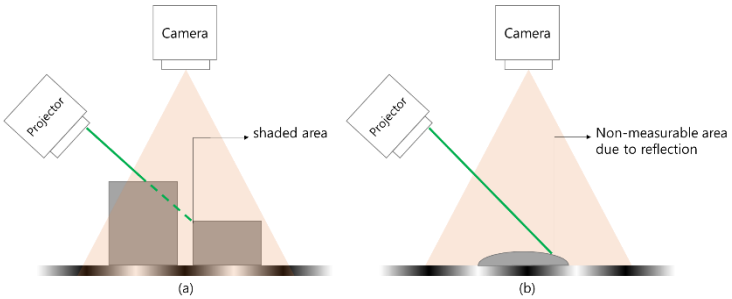


Figure 1. Reasons for Moiré Pattern Damage on PCB Components. (a) is a kinematic problem between the part and the pattern projector, (b) is a reflection problem due to the shape and material of the part.

2 Data Augmentation Considering Moiré Pattern

A typical data augmentation method is shown in Figure 2. The data augmentation method can be applied to general data, as shown in Figure 2. However, if the moiré pattern exists above, the methods (b) to (e) damage the moiré pattern, making it impossible to estimate the height. Therefore, the data augmentation method for the moiré pattern must be made differently.

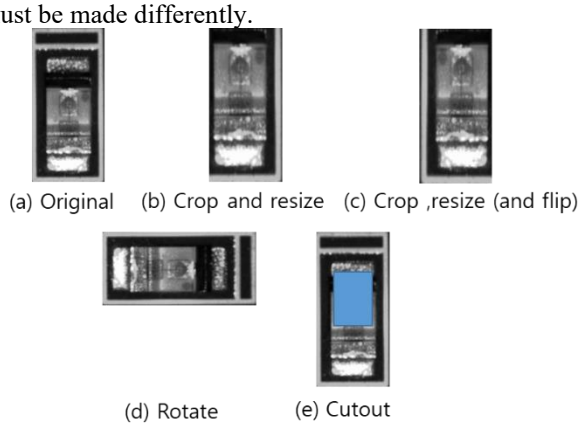


Figure 2. Typical Data Augmentation Methods.

Figure 3 is the result of projecting the moiré of the PCB component. Figure 3 is an image subtracted from consecutive shooting results when the same phase and component board are used. That is, if (a) and (b) find the same phase of the same part or the difference in data, it is output as shown in Figure 3. (c). Through (c), it can be confirmed that an acceptable difference that cannot be distinguished with the naked eye appears. Learning the difference between these delicate parts is challenging, and there is a problem with a small dataset. Data augmentation is required, but if data

augmentation is used, as shown in Figure 4 (a), which is an existing method, it does not consider the moiré pattern, which makes learning difficult. Therefore, we propose Figure 4(b). We were inspired by lidar data augmentation techniques and considered the proposed method [3].

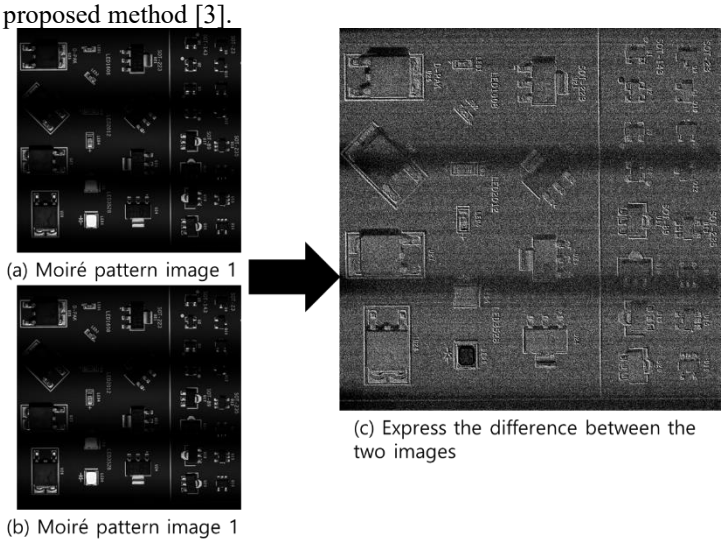


Figure 3. Moiré pattern image and difference image of PCB board.(a) and (b) are images of the same phase and the same part, and the difference between the two images is expressed in (c).

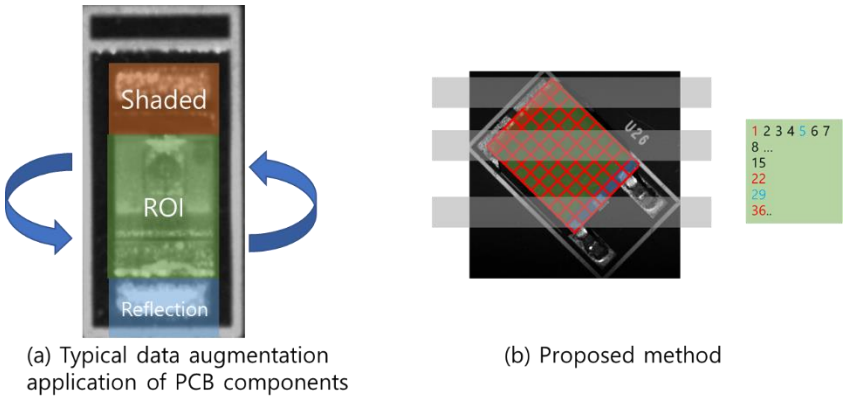


Figure 4. An example of how to apply data augmentation to a part. (a) general method, (b) set up the grid in the proposed method.

Figure 4(b) is the proposed method. First, the shadow and reflection areas are set in the area of the part, and then the ROI is set. In the set ROI, red indicates an area entirely covered by the moiré pattern, and blue indicates a reflection area. When data is switched within a grouped grid to augment learning data, data augmentation can be performed without damaging the moiré pattern.

3 Conclusion

In this paper, we present a method for effectively augmenting a PCB board's moiré pattern image data. When using this method, because the moiré pattern is considered, false images are not augmented and do not interfere with learning. However, the limitation of the proposed method is that the data augmentation effect is insignificant, and verification of various types of parts has not been done yet. In future studies, the proposed method will be improved by considering the verification of various parts of the network and a practical training data augmentation method.

4 Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00795, Development of Moire-Patten Type 3D Camera System for AI Based Analysis)

References

1. Zuo, Chao, et al. "Phase shifting algorithms for fringe projection profilometry: A review." Optics and Lasers in Engineering 109 (2018): 23-59.
2. Yen, H-N., and D-M. Tsai. "A fast full-field 3D measurement system for BGA coplanarity inspection." The International Journal of Advanced Manufacturing Technology 24.1 (2004): 132-139.
3. Wang, Chunwei, et al. "Pointaugmenting: Cross-modal augmentation for 3d object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Non-Contact Body Temperature Measurement through Facial Thermal Features: A Survey

Ulziitamir Davaadorj[1], Aziz Nasridinov[1, *]
Dept. Of Computer Science, Chungbuk National University, South Korea
Email: {tamiraa, aziz}@chungbuk.ac.kr

Abstract. Non-Contact Body Temperature Measurement (NCBTM) cameras have extensively been utilized to accurately estimate human body temperature for pandemics, including COVID-19 and the Monkeypox outbreak. NCBTM minimizes the risk of infection among medical personnel and is cost-effective and less time-consuming. Most NCBTMs calculate the body temperature by concentrating on the facial thermal features. However, the body temperature measurement focused on the face skin's surface has challenges regarding external factors that lead to inaccurate results. Therefore, this study explores getting body temperature accurately using facial thermal features. Our survey provides an overview of NCBTM methods by selecting different parts of facial landmarks, facilitating the reduction of false negatives that can be affected in measurement.

Keywords: Body temperature measurement; non-contact; thermal camera; face detection; thermal sensor

1 Introduction

Measuring body temperature is an initial practical activity to manage the spread of the epidemic and segregation of people by detecting fever in the early stage. Measurement approach can be categorized into contact and non-contact methods. The body temperature is calculated for the type of contact based on oral, tympanic, and armpit thermometers. Although this method takes human's internal temperature, it has a higher risk of infection as well as inefficient, especially during the pandemic. In contactless method, the temperature of the human skin surface is measured by using a thermal sensor. This method significantly reduces the risk of infection for the medical diagnosis stage, and it is the most efficient and fastest way. According to research (South Korea, 2020), nurses face an increased risk of acquiring infection; 44.2% of the sampled nurses reported a high risk of disease during the COVID-19 risk score by occupation [1]. This fact led to demands for the rapid development of AI-based NCBTM and brought modifying the traditional approach.

We present a survey of non-contact body temperature measurement techniques using a thermal camera. Our study discusses various methods of taking accurate body temperature based on the thermal features of facial parts in recent research. More specifically, we explore a better understanding of the issues in current studies and approaches to handling them. This paper is organized as follows. First, the perception of the non-contact body temperature measurement process and related studies are

introduced in Section 2. Next, the advantage and issues of studies and possible improvements are discussed in Section 3.

2 Survey of Non-Contact Body Temperature Measurement

There are several stages involved in image processing and temperature reading of the body. Generally, the human body's temperature in contactless is estimated by scanning the entire face area or specific frontal components, including the eye, nose, forehead, and cheek. Deep Learning or Machine learning models automatically detect the target area and transfer coordinates in a rectangular termed a Region-of-Interest (ROI). Subsequently, the body heat is computed using a combination of the obtained ROI coordinates and thermal data. The error estimation approaches (mean absolute error (MAE), or root mean squared error (RMSE)) and accuracy calculation are used to evaluate the results by comparing obtained ground truth values in contact ways such as oral, ear, and armpit. Table 1 summarizes all the characteristics of the studies with NCBTM as the main objective of this work.

Table 1. Summary of chronological literature

Study	Algorithm Used	ROI	Temperature Estimation	Performance	Thermal Camera
[2]	MobileNet-SSD	Forehead	Average temperature	MAE:0.375° C RMSE:0.439° C	FLIR2.5, 80x60 px
[3]	OpenCV	Entire face	Highest temperature	Accuracy of 98.1%	FLIR2.5, 80x60 px
[4]	Face Recognition algorithm	Inside Eye	Single point temperature	Accuracy of 85%	Thermal camera
[5]	Cascaded-YOLO V.3 (two)	Eye, nose, cheek	Prediction by DL method	Average error:0.2° C	Thermal camera
[6]	RetinaFace Detector	Forehead	Center point temperature	Error < 0.5° C	FT20, 256x192 px

Each study obtains the body temperature value by suggesting different objectives. For example, Lin et al. [2] propose detecting faces using Single-Shot-Multibox Detector (SSD) with MobileNet from thermal images and defining a forehead as ROI. After determining the ROI, the body temperature is identified by computing an average temperature of all pixel points of the forehead is calculated considering facial thermal characteristics. Radiometric calibration on the sensor is conducted with parameter updating to get the accurate temperature. Finally, they use a high-performance thermal camera to evaluate their method, and NCBTM is assessed with an overall error of 0.375° C of MAE and 0.439° C of RMSE.

Recent research performs several steps to accurately measure the body temperature using a combination of RGB and a thermal camera. The authors in [3] consider the face as ROI from the RGB image using the OpenCV algorithm, and then the ROI range is matched with the thermal image. The highest value of the entire face is considered as body temperature and then compared with the measurement that used an IR

thermometer. The result achieved an average accuracy of 98.1%. The experiment illustrates the optimal distance that can handle external factors such as ambient temperature and light intensity, marked as 0.5 meters to 4 meters.

The NCBTM of this paper [4] determines eye area as ROI in a thermal image by recognizing the face and detecting eyes. They measure the body temperature from the inner corner of the eye, which suggests preferring a single point in ROI. Additionally, this work utilizes the sensor to set the proper distance between the person and a thermal camera. An Ultrasonic sensor configures the person's space with a threshold of 30 cm to operate the process correctly, and the result achieves an accuracy of 85%.

The authors in [5] suggest using facial skin features to measure the body temperature, such as the inside eye having thin skin and the cheek having thick skin. In addition, they propose that artificial intelligence can be operated in contactless body temperature measurement. The main contribution is that a deep learning method predicts body temperature using selected ROIs. Three types of ROI are determined using an integration of two YOLOs (You-Look-Only-Once); the inside eye, nose, and cheeks. The experiment data is collected by a thermal camera in various environments, including hot, normal, cold, and after exercise. They estimate the highest temperature from the inside eye and the average temperature from the nose and cheek as body temperature. A tympanic thermometer collects the core body temperature measurement in contact type as ground truth. The environment recognition model is used to improve the performance of the body temperature prediction. The performance shows an average error of 0.2º C compared with the ground truth. In real-time experiments, the RMSE of the prediction result was obtained as 0.1053.

Several recent studies consider the method of contactless measurement of vital signs (body temperature (BT), heart rate (HR), and respiration rate (RR)) using a thermal camera and RGB camera. We mainly observe the body temperature measurement from this study [6]. RetinaFace algorithm that can detect face and face landmarks simultaneously works to select the ROI. The forehead is picked as ROI for BT calculation, and the center point on ROI is identified as body temperature value. A supplementary method is described here to control head movement according to the facial landmarks is performed to get an accurate temperature of the given point. In addition, the affine transformation method is utilized to align the frames of two cameras (RGB and thermal). The actual body data consists of an oral thermometer measurement, and the result of measuring the forehead skin surface reports a difference of less than 0.5º C degrees.

Ultimately, the environment setup is an essential factor as well. Each study configures an experimental design to get a highly accurate temperature value, such as the distance and angle of the thermal camera and the person.

3 Discussion

The research in NCBTM is dynamic and paramount regarding the eagerness for accurate measurements. In order to improve the measurement accuracy, we searched for state-of-the-art development using facial thermal features.

The authors in [5] described that the combination of several ROIs' temperatures is used for measuring the body temperature in real-time. This method brought the lowest error by comparing the remote and ear thermometer measurements. In addition, they proved that inside eye temperature has a linear relationship with core body temperature, in which they verified inside of the eyes is not affected significantly by external factors. Lin et al. [2] demonstrated that the forehead measurement could detect body temperature with a lower error by calculating the average value of an ROI. On the other hand, Yang et al. [6] illustrated that a single point temperature of the forehead region could identify body temperature, which has achieved a lower performance. The authors Yaghi et al. [4] identified obstacles hindering the accurate temperature results that should be addressed to push the technology further. For example, they used a suitable sensor to handle the distance variation in the experiment.

Recent studies mainly utilized the evaluation method of comparing the results of experiments with ground truth data. Lin et al. [2] method have several challenges when collecting body temperature data. For instance, the temperature measurement of the forehead should be conducted over various external impacts such as brightness and darkness and room temperature variation. Next, the remote measurement of body temperature must be compared with core body temperature measurement, meaning that it is difficult to define if it can obtain accurate temperature in real-time. Aufar et al. [3] mentioned that body temperature measurement is calculated by detecting the entire face and picking the highest temperature. Regarding that, the outside area of the face can be included as the ROI. The authors in [5]'s method must consider the situation when the patient is using a mask or sunglass. Because to detect the nose, cheek, and eye, the patient should take off the mask or sunglass first.

We found some considerations that recent studies used insufficient participants for the experiment, and core body temperature is not compared for evaluation. Finally, we observed that accuracy improvement could be addressed by minimizing the ROI range.

References

1. Lee, Juyeon, Kim, Myounghee, Estimation of the number of working population at high-risk of COVID-19 infection in Korea, Korean Society of Epidemiology, (2020)
2. J. -W. Lin, M. -H. Lu and Y. -H. Lin, A Thermal Camera Based Continuous Body Temperature Measurement System, 2019 IEEE/CVF International Conference on Computer Vision Workshop, pp. 1681-1687, (2019)
3. F. Aufar, M. A. Murti and M. H. Barri, Design of Non-Contact Thermometer Using Thermal Camera For Detecting People With Fever, 2021 International Conference on Computer Science and Engineering, pp. 1-5, (2021)
4. M. Yaghi, A. Alsalmani, M. Qasymeh, M. Alkhedher, M. Yaghi and M. Ghazal, Accurate and Contactless COVID-19 Fever Screening and Attendance Tracking System, 2022 2nd International Conference on Computing and Machine Intelligence, pp. 1-6, (2022)
5. C. Song and S. Lee, Accurate Non-Contact Body Temperature Measurement with Thermal Camera under Varying Environment Conditions, 2022 16th International Conference on Ubiquitous Information Management and Communication, pp. 1-6, (2022)
6. Yang, F.; He, S.; Sadanand, S.; Yusuf, A.; Bolic, M. Contactless Measurement of Vital Signs Using Thermal and RGB Cameras: A Study of COVID 19-Related Health Monitoring, Sensors (2022)

MULTIMEDIA
LIFE
STORAGE
NETWORK
DATABASE
SYSTEM
BIG DATA
SCIENCE
CLOUD
TREND
CLUSTER
ANALYSIS
BUSINESS
SOCIETY
GRAPHICS
VISUALIZATION

THE KOREA
BIG DATA SERVICE SOCIETY
한국빅데이터서비스학회

N13 404-2, Chungbuk University, Chungdae-ro 1
Seowon-Gu, Cheongju, Chungbuk 28644, Korea

Email: kbigdataservice@gmail.com

Homepage: www.kbigdata.or.kr