CNN-based Visual/Auditory Feature Fusion Method with Frame Selection for Classifying Video Events

Giseok Choe[*], Seungbin Lee and Jongho Nang
Department of Computer Science and Engineering, Sogang University
Seoul, South Korea
(gschoe, mercileesb, jhnang)@sogang.ac.kr
*Corresponding author: Giseok Choe

Received October 3, 2018; accepted January 17, 2019; published March 31 2019

Abstract

In recent years, personal videos have been shared online due to the popular uses of portable devices, such as smartphones and action cameras. A recent report[1] predicted that 80% of the Internet traffic will be video content by the year 2021. Several studies have been conducted on the detection of main video events to manage a large scale of videos. These studies show fairly good performance in certain genres. However, the methods used in previous studies have difficulty in detecting events of personal video. This is because the characteristics and genres of personal videos vary widely. In a research, we found that adding a dataset with the right perspective in the study improved performance. It has also been shown that performance improves depending on how you extract keyframes from the video. we selected frame segments that can represent video considering the characteristics of this personal video. In each frame segment, object, location, food and audio features were extracted, and representative vectors were generated through a CNN-based recurrent model and a fusion module. The proposed method showed mAP 78.4% performance through experiments using LSVC[2] data.

Keywords: Multimedia, Computer Vision Systems, Aritifical Intelligence, Video Classification

A preliminary version of this paper was presented at APIC-IST 2018, and was selected by the conference review process. This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2017-0-01772. Development of QA system for video story understanding to pass Video Turing Test) and Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2017-0-01781.Data Collection and Automatic Tuning System Development for the Video Understanding)

http://doi.org/10.3837/tiis.2019.03.033 **ISSN : 1976-7277**

1. Introduction

In recent years, personal videos have been shared through YouTube or Flickr due to the popular uses of smartphones and action cameras. A recent report[1] predicted that 80% of the Internet traffic will be video content by the year 2021. Accordingly, content-sharing companies perform video event detection by managing a large scale of videos to provide services for users. However, video event classification conducted by watching videos via humans can take considerable time and human resources. To resolve this problem, computer vision researchers have continuously conducted studies on the classification of main video events automatically. Personal videos may have severe noise due to shaking and lighting depending on the expertise of shooters or shooting device performance, and videos have a different duration. Videos are thus more difficult to handle than single images due to the temporal relation between frames. In recent years, studies using a deep neural network (DNN), which has played a major role in solving various problems in the computer vision area, have been conducted to analyze videos using complex features.

Most studies have involved experiments with short-duration video data, and videos have been analyzed using object-oriented features extracted from a convolutional neural network (CNN) trained with ImageNet[3]. However, personal videos were collected from various categories, and various sets of information included place, food, and voice etc. that can be used to detect main events. Thus, this study samples segment frames from various duration images and features of objects, places, foods, and voices from a variety of viewpoints are extracted to analyze the main events. In addition, the main events are detected by encoding features that have a sequential structure into fixed-length features. The LSVC 2017[2] dataset, which was used in the Large-Scale Video Classification 2017, one of the largest video datasets in the world, was used in the experiment for performance evaluation. The proposed method in this study achieved a performance improvement of 9% compared to that of the existing single-feature-oriented mode, and 78.4% of performance in the mean average precision (mAP) can be achieved.

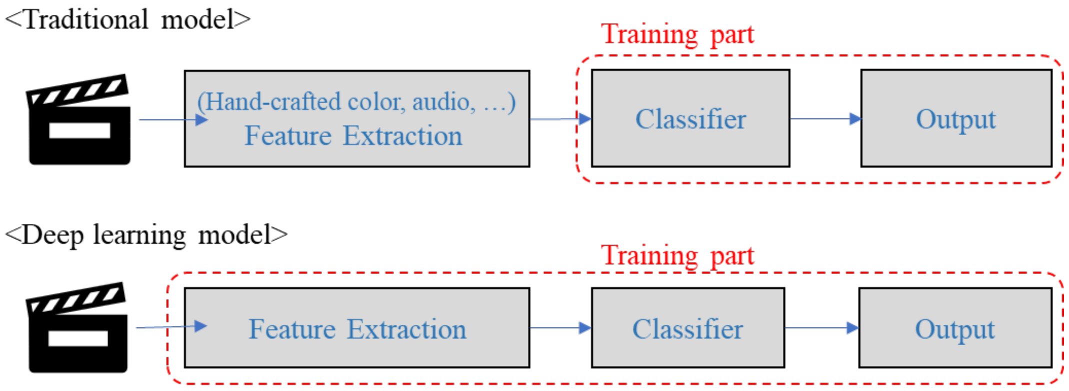
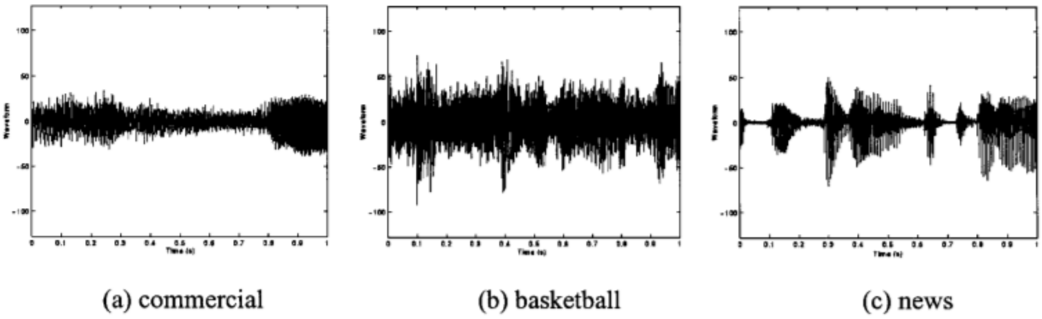


Fig. 1. Structures of traditional feature and DNN based models

2. Related Work

Various approaches have been studied to solve the video event classification problem, which has been considered one of the main problems in computer vision. The final output is returned after performing feature extraction through the pre-determined indexing method and training the classifier model. In contrast, a more recent deep learning model trains not only classifiers but also convolutional filters, which are suitable for classification using a loss function. **Fig. 1** shows the structures of the two models. Next, in Section 2.1, traditional models are explained followed by a detailed explanation about studies related to a recent DNN in Section 2.2.



(a) commercial (b) basketball (c) news

Fig. 2. An example of video classification using audio information

2.1 Traditional Feature-based Research

Studies prior to DNN extracted hand-crafted features of users for classification. In the text-based research displayed in videos, a study[4] was conducted to categorize the news subject into politics, society, and entertainment using subtitles in videos. A study[5] on using audio information classified videos into discourse, news, commercial, and sports, as shown in **Fig. 2**. A study[6] using visual information analyzed videos by extracting a variety of low-level features in frames. There have been studies dealing with transitions, objects, and motion in color, texture, and shot boundary. In particular, motion may degrade event detection performance, as it can make noise using camera movements in personal videos. In a study[7], motion compensation was applied to prevent this noise. In addition, noise was eliminated around persons who were focused on in events, and features, such as the trajectory and histogram of gradient (HoG), were extracted for better performance. These studies have significant effects on recent studies on DNN-based video analysis.

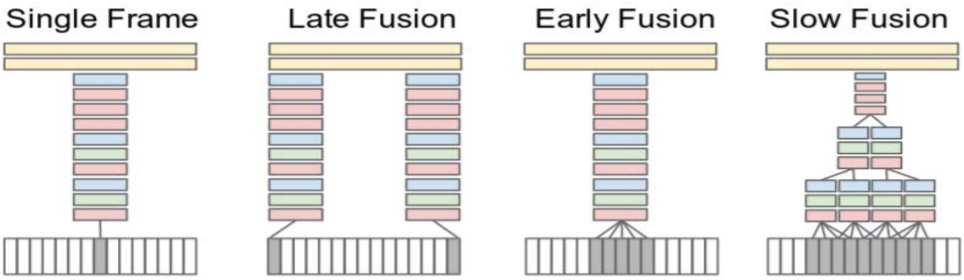


Fig. 3. An example of the fusing frame method used in [8]

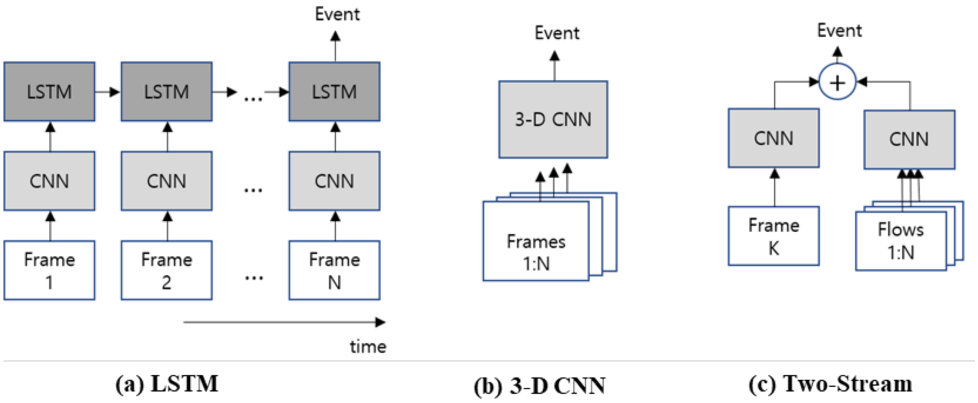


Fig. 4. Structures of the DNN-based video analysis method

2.2 DNN-based Research

Traditional feature-based works are difficult to use when studying personal videos whose themes are diverse because they analyze videos based on pre-defined criteria. To solve this problem, studies using a CNN, which revealed a significant improvement in the computer vision field, have been conducted. In one study, the researchers[8] proposed a method of fusing features between frames after selecting input frames using various methods, as shown in **Fig. 3**, to deal with temporal dependency using a CNN. Although the results of that study were not better than those of [7], it was regarded as a represented study using a CNN and became the reference study for subsequent CNN-related research.

There have been studies in consideration of temporal characteristics to overcome the shortcoming of the difficulty in analyzing sequential frames using a CNN. **Fig. 3** shows a video analysis method using a neural network of various structures. These studies have object-oriented feature extraction from each frame using a CNN trained with ImageNet in common. However, various methods are used to deal with features of continuity as shown in **Fig. 4**. **Fig. 4(a)** shows a video analysis method [9] using long short-term memory (LSTM), which has a recurrent model. The feature information in the video is compressed through features between adjacent frames using an LSTM structure. and videos are classified. Although the performance was better than that of [7], it was worse than that of [7]. **Fig. 4(b)** shows the use of three-dimensional (3D) convolution to overcome a drawback of existing two-dimensional (2D) convolution, which was difficult to use for the analysis of the continuous structure of video frames, although 2D convolution is easier for understanding single frame locality. It showed better performance than that of [9] or [7], as it considered spatial information and temporal dependency. When it was fused with features in [7], its performance was improved further. However, it required a large amount of resources to train the complex structure. One study [10] using a structure in **Fig. 4(c)** showed better performance than that of **(a)** and **(b)** that determined classification on the basis of frames only by adding motion information, such as optical flows. However, it also required a large amount of resources in a process that obtained optical flows.

Table 1. Previous works for video event classification

Paper publication	Algorithm	mAP
Wang et al.[7]	Improved dense trajectory(iDT)	0.859
Karpathy et al.[8]	Spatio-temporal CNN	0.654
Donahue et al.[9]	CNN + LSTM	0.829
Tran et al.[11]	Spatiotemporal 3D-CNN + iDT	0.904
Wang et al.[10]	Two-stream + iDT	0.915

Table 1 compares the performances of the studies described in Sections 2.1 and 2.2. The data used in training and validation were UCF-101[12]. A detailed explanation of the data is presented in the next section

3. Dataset

The performance verification was conducted using LSVC 2017 video datasets used in the large-scale video classification challenge in this study experiment. UCF-101 was mainly used for the video event classification study. As presented in **Table 2**, UCF-101 consists of 13,000 data records with a total of 101 events, such as “Apply Eye Makeup,” “Apply Lipstick,” and “Playing Cello,” in daily activities. Its average video length is seven sec. and videos are trimmed to display only main events. In contrast, LSVC 2017[2] consists of 155,000 data records with a total of 500 events. Its average length is 186 sec., which is longer than that of UCF-101, and videos are untrimmed. Its video length is longer and uses original versions without editing. Thus, it is a more difficult benchmark dataset than UCF-101.

Table 2. Comparison of UCF-101 and LSVC 2017 datasets

Data	Class	trimming	Total Data	Mean length
UCF-101[12]	101	O	13,000	7 sec.
LSVC 2017[2]	500	X	155,000	186 sec.

Table 3. Added dataset for various perspectives

Data	Class	Avg. data per class	Total number of images
ImageNet[3]	1,000	1,200	1.2M
Place-365[13]	365	4,300	1.6M
Food-101[14]	101	155,000	101K

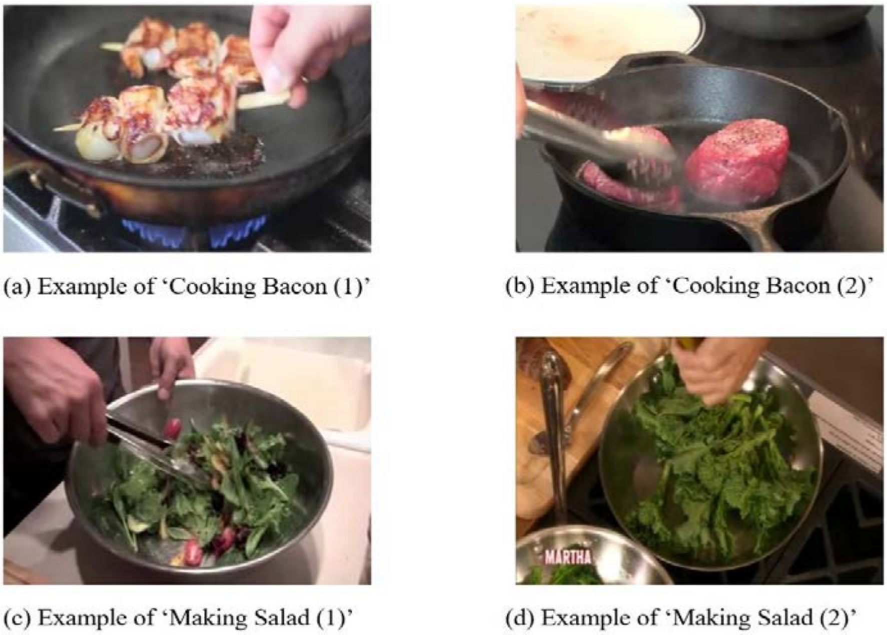

(a) Example of 'Cooking Bacon (1)'
(b) Example of 'Cooking Bacon (2)'
(c) Example of 'Making Salad (1)'
(d) Example of 'Making Salad (2)'

Fig. 5. Examples of visually similar frames

The frames shown in **Fig. 5** are contained in the LSVC dataset. In **Fig. 5, (a)** and **(b)** or **(c)** and **(d)** are similar frames in terms of visual information but explain different events, or **(b)** and **(d)** are the same event but their frames are significantly different. In response to this type of problem, a dataset for a particular perspective can be helpful. The data in **Table 3** are the datasets added for this reason.

4. Fused Feature-based Event Detection System

Although features in videos can be extracted from various perspectives, existing studies extract features via a CNN trained with object-oriented features using ImageNet. The visual information is the most important information among many modalities displayed in videos. However, since videos have significant noise due to various external factors, such as the shooting environment, shooting device performance, and the shooter's skill, it is difficult to classify events using only low-level features. In addition, main events cannot be represented with only information at a specific time. Considering the above points, this study designed the structure shown in **Fig. 7**.

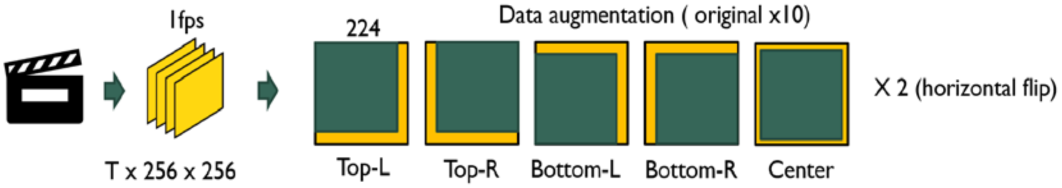


Fig. 6. A Method for data augmentation

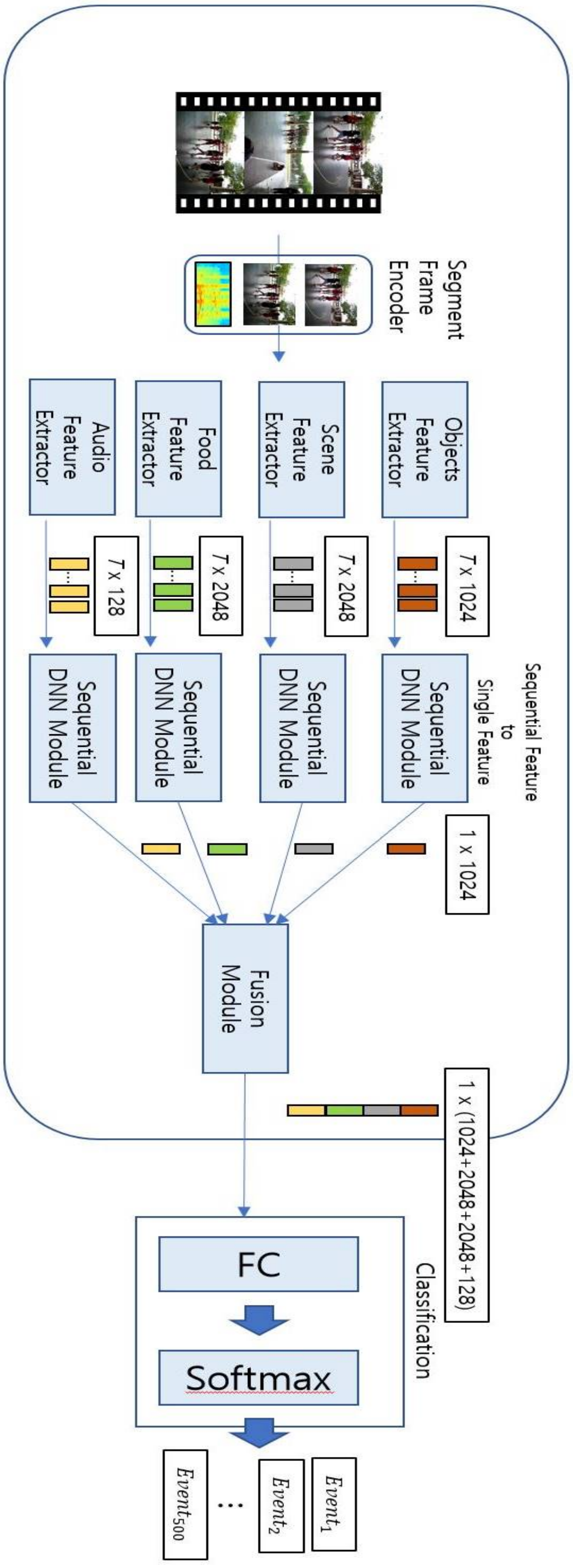


Fig. 7. Overview of the proposed model

4.1. Frame segment encoder

A large amount of data is needed to achieve superior performance with regard to training data while preventing overfitting in the DNN. However, data collection is difficult from many problem domains of daily activities and obtaining the ground truth is more difficult. To overcome this difficulty, data augmentation is applied to increase the size of training data by modifying existing data using various changes. Since it can give various changes to input data and prevent overfitting effectively to increase generalization, it has been widely used in a pre-processing procedure in the CNN study. For images, horizontal flip, random crop, and adding noise can be used to increase the number of limited data records. This study employs a method shown in **Fig. 6** to apply the data augmentation method to videos. A video that is a total of T sec. is extracted into one second. frames, which are then cropped into a fixed size of four corners in the right, left, upper, and lower sides, and the center. Then, the original data are enlarged 10 times using the horizontal symmetry and these values are averaged to reduce the noise effect. **Fig. 8** shows the algorithm of the frame data augmentation explained above.

```
function FrameAugmentation(Frames)
  frame_list = [ ]
  for each frame in Frames
    for each corner in Corners
      cropped = CropImage(frame, corner)
      flipped = FlipImage(cropped)
      frame_list.PUSH(cropped)
      frame_list.PUSH(flipped)
  return frame_list
end function
```

Fig. 8. Algorithm of frame augmentation

Existing studies employed entire frames of a video using a relatively short-term UCF video. However, the average duration of LSVC videos amounts to 186 sec., and videos are untrimmed. The information displayed in part of the frames in the video may not tell the main event.

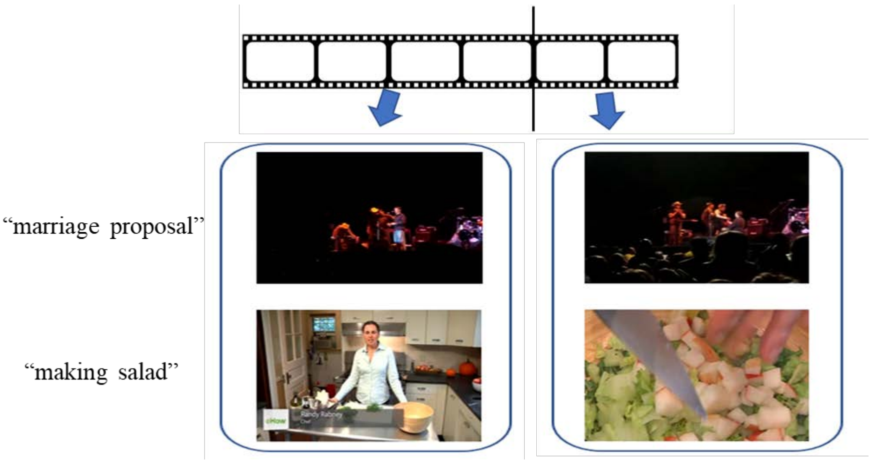


Fig. 9. Example of key frames shown in the latter part of a video

For example, the frames in the first half of the video contain features that cannot explain the main event correctly, as shown in Fig. 9. However, the main event of the entire video appears in the second-half frames. The images related to a "cooking" event involve ingredients or cooking method. Then, the second half frames display which food is cooked in the event. The event video called "Marriage proposal" displays feature frames such as "tailgate party," "music concert," and "recital" in the first half. However, the main frame of marriage proposal appears in the second half frames. Thus, a frame segment extraction method was used in which an entire video was divided into fixed lengths, as shown in Fig. 10, and the segment section was extracted randomly to generate a single frame.

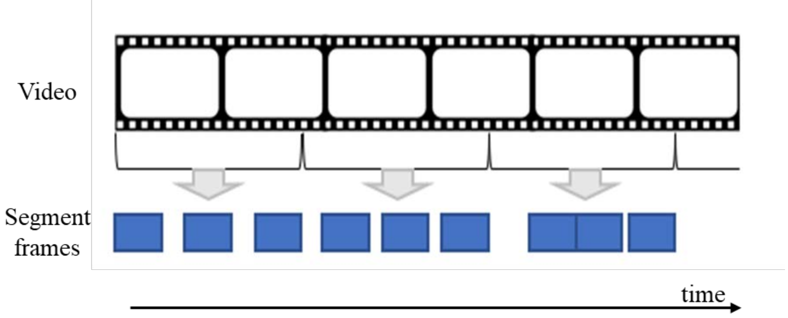


Fig. 10. A method of extracting a segment in a video

In contrast with existing studies, features were extracted on the basis of an object, place, food, and audio to utilize various pieces of information displayed in the video.

4.1. Consider various perspectives

Detailed information of the CNN used for fused-feature extraction in this study is presented in Table 4. Various CNN models were employed, and perspectives of the feature extraction were modified using the trained data. For the feature information, outputs in the last layer for feature extraction were employed and the length of vectors in each layer varied.

BN-Inception[15, 16] and VGG19[17] that were trained with ImageNet were used to extract object-based features. Transfer learning was performed with Resnet152[18] trained with ImageNet into Place365[13] and Food101[14] data to extract a distinguished representation between place and food. In addition, videos contain not only various visual features but also many audio features. To employ voice information to classify main events, VGGish[19] that showed superior performance in feature extraction using a CNN in recent years was used after changing voice signals to spectrogram.

Table 4. Details of multi-modality feature extraction

Perspective	Structure	Dataset	Extraction Layer	Size
Object	BN-Inception[15]	ImageNet[3]	global_pool	1024
Object	VGG19[17]	ImageNet[3]	FC6	4096
Location	Resnet152[18]	Place365[13]	Pool5	2048
Food	Resnet152[18]	Food101[14]	Pool5	2048
Voice	VGGish[19]	Audioset[19]	Embedding layer	128

5. Experiment and analysis

The purpose of this study was to improve existing models by extracting object, place, food, and audio CNN representations from sec. unit frames and designing a DNN model in consideration of temporal dependency to classify the main events of personal videos collected from YouTube or Flickr. Experiments were conducted to detect events by only seeing 300 frames at a maximum in the first half without performing segment sampling separately.

Table 5. Performance comparison using object-oriented features

Feature size	**Recurrent model**	**Single feature extraction model**		
		VGG-19 (ImangeNet only)	**BN-Inception (ImageNet only)**	**4 CNNs (ImageNet + Place + Food + Audio)**
Video (1 x Size)	Pooling	0.648	0.702	0.723
Sequence (T x Size)	NetVLAD	0.664	0.723	**0.765**
	NetFV	0.651	0.714	**0.751**
	LSTM	0.544	0.683	0.721
	GRU	0.569	0.690	0.730

The performances of the basic model using only object-oriented visual information are presented in Table 5. The performance of BN-Inception was better than that of the model using VGG19. VGG19 performed worse than that of 1,024-dimension BN-Inception despite using 4,096-dimension information. This result indicated that the performance difference was determined by which structure of the neural network was used and from which layer extraction was made for visual information features. Furthermore, the performance of NetVLAD[20] was better than that of LSTM. This indicated that different features can be produced depending on the structure of the recurrent model.

Table 6. Performance comparison using features from various perspectives

Characteristics Size	**Recurrent Model**	**Object**	**Location**	**Food**	**Voice**
Sequence (T x Size)	NetVLAD	0.723	0.688	0.660	0.090

Table 6 presents the classification performances using a single feature among object, location, food, and audio. When only a single feature was used, a method using an object-based visual feature showed the best performance. However, location or food-based features showed worse classification performance than using object-based features. The object-based features revealed the best performance because they can extract distinguished features of objects that were relatively represented universally, while location or food could extract good features about specific events. When only audio information was used in the experiment, it was easily affected by noise. The classification of events using only audio information is not easy, even for humans. The experiment results also indicated much worse performances than using only visual information.

Table 7. Performance of the proposed model using multi-modality features

Model	**mAP**	**Combination of multi-modality features**
Model 1	0.723	Object
Model 2	0.734	Object, Food
Model 3	0.742	Object, Location
Model 4	0.750	Object, Location, Food
Model 5	0.765	Object, Location, Food, Voice

The experiment results of the multi-modality features proposed in this study are presented in Table 7. Model 1 is a case using object-based visual features. Models 2 and 3 improved performances by adding food or location features compared to that of using a single feature. As revealed in Models 3, 4, and 5, performances improved whenever a feature was added. When all four features were used, performances were improved by 4% compared to using an object-based single feature.

Table 8. Performance comparison using the proposed segment sampling

Feature size	**Recurrent model**	**mAP@val (increased performace)**	
		Model 5	**Model 6 (Model 5 + Sampling)**
T x Size (T x Size)	NetVLAD	0.765 (+5.8%)	**0.784 (+8.4%)**
	NetFV	0.751 (+5.2%)	**0.773 (+8.3%)**
	LSTM	0.721 (+5.6%)	0.745 (+9.1%)
	GRU	0.730 (+5.8%)	0.756 (+9.6%)

However, when the main event in a video is not contained in 300 frames, the event may be detected inaccurately due to the lack of information. Thus, features were extracted by the random sampling of partial sections in the following video. Table 8 presents the performances before and after sampling. Model 6 shows the performance when events were detected through segment sampling and multi-modality features used in Model 5. The proposed segment sampling showed 2 ~ 3% performance improvement regardless of the method.

The results showed that the performance (0.723) revealed when only an object-based feature was extracted was improved to 0.784 through multi-modality features and segment sampling.

6. Conclusion and future research

This study improved a problem that made main event classification difficult by taking the characteristics of personal video shots from various fields into consideration. The performance of a DNN can vary depending on the learning data, parameters, neural network structure, and feature extraction layer. However, performances were improved simply by connecting features extracted from other perspectives and through segment sampling if a video was long.

In a research, we found that adding a dataset with the right perspective in the study improved performance. It has also been shown that performance depending on how you extract keyframes from the video. Therefore, in future research, optical flow will be added to increase diversity of perspective, and key frame extraction will be considered according to video characteristics.

References

[1] CISCO, "Cisco Visual Networking Index: Forecast and Methodology," Feb 15, 2018; https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html.

[2] Z. Wu, Y. G. Jiang, L. S. Davis, and S.-F. Chang, "LSVC2017: Large-Scale Video Classification Challenge," 2017. Article (CrossRef Link)

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255. Article (CrossRef Link)

[4] W. Zhu, C. Toklu, and S.-P. Liou, "Automatic News Video Segmentation and Categorization Based on Closed-captioned Text," *Urbana*, vol. 51, pp. 61801, 2001. Article (CrossRef Link)

[5] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 20, no. 1-2, pp. 61-79, 1998. Article (CrossRef Link)

[6] B. T. Truong, and C. Dorai, "Automatic Genre Identification for Content-based Video Categorization," in *Proceedings of International Conference on Pattern Recognition*, pp. 230-233, 2000. Article (CrossRef Link)

[7] H. Wang, and C. Schmid, "Action Recognition with Improved Trajectories," in *Proc. of IEEE International Conference on Computer Vision*, pp. 3551-3558, 2013. Article (CrossRef Link)

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732, 2014. Article (CrossRef Link)

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625-2634, 2015. Article (CrossRef Link)

[10] L. Wang, Y. Qiao, and X. Tang, "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4305-4314, 2015. Article (CrossRef Link)

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3d Convolutional Networks," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 4489-4497, 2015. Article (CrossRef Link)

[12] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild," *arXiv preprint arXiv:1212.0402*, 2012. Article (CrossRef Link)

[13] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017. Article (CrossRef Link)

[14] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining Discriminative Components with Random Forests," in *European Conference on Computer Vision*, pp. 446-461, 2014. Article (CrossRef Link)

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. of International Conference on Computer Vision and Pattern Recognition*, 2015. Article (CrossRef Link)

[16] S. Ioffe, and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv:1502.03167*, 2015. Article (CrossRef Link)

[17] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014. Article (CrossRef Link)

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. Article (CrossRef Link)

[19] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, and B. Seybold, "CNN Architectures for Large-scale Audio Classification," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131-135, 2017. Article (CrossRef Link)

[20] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297-5307, 2016. Article (CrossRef Link)

Giseok Choe is currently a doctoral student at Sogang University. He is interested in a system that combines video broadcasting technology and AI technology. He received his MS degree from the Computer Science and Engineering Department at Sogang University.

Seungbin Lee is currently working the position of software engineer at Coupang. He is interested in how AI could improve our life, especially information retreival and deep learning. He reveived his MS degree from the Computer Science and Engineering Department at Sogang University.

Dr. Jongho Nang received his Ph.D. and M.S. degrees in Computer Science from Korea Advanced Institute of Science and Technology (KAIST), Daejon, Korea, in 1992 and 1988, respectively, and his B.S. degree in Computer Science from Sogang University, Seoul, Korea, in 1986. He has been a professor of Computer Science and Engineering Department, Sogang University since 1993. His research interests include multimedia system, deep neural network, and internet technology.