

Edge Computing System applying Integrated Object Recognition based on Deep Learning

Kwihoon Kim*, Sewon Oh**

*KNUE, 250 Taeseongtabyeon-ro, Gangnae-myeon, Heungdeok-gu, Cheongju-si, Chungcheongbuk-do, 28173, Korea

** ETRI, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Korea

kimkh@knue.ac.kr, sewonoh@etri.re.kr

Abstract— With the development of personal broadcasting such as YouTube, the demand for editing the video he filmed himself is steadily increasing. Traditional media video editing solutions help edit the results of the filming. However, there is a problem that requires a lot of editing time, as people usually edit videos that are from tens to hundreds of times longer than the final result of a personal broadcasting after filming. To overcome this problem, this paper automatically classifies images with specific scenes in the whole media image editing process, and secondly proposes automatic media editing solution technology in which people intervene. In particular, personal broadcasting focuses on the use of images that include characters, specific objects, and cue sign gestures among the entire. While the existing deep learning techniques such as faces, objects and gestures are advanced, integrated recognition technologies that simultaneously deal with special requirements for editing videos are still in the early stages of research. In this paper, the automatic composite recognition technology for editing video based on deep learning is proposed. The proposed technology was implemented with python and tensorflow software based on edge computing equipment. Using actual youtube videos, it took 0.1 second to process five-person recognition, 63-food recognition, or cue sign recognition using clapping or V poses at the same time. The recognized results are divided into timestamps of the entire movie, recognition results, and locations of objects on the screen, and are output to the json file. In addition, this solution was developed on an edge computing in order to increase real-time reliability. We expect to provide automatic video editing based on perceived json results as well as shorter editing times based on this implementation.

Keywords — Automatic media editor, Object recognition, Context recognition, Deep learning, Edge computing

I. INTRODUCTION

These days, as smartphone cameras are developed and personal broadcasting platforms such as YouTube and Africa are active, many attempts to do personal broadcasting have increased. However, in order to broadcast personally, since they have to edit the videos they have shot, more people learn to edit videos on their own, and interest in videos is increasing. Even in the hopes of students in the future, YouTube creators were created, and even workers with jobs are dreaming of becoming YouTubers. However, video editing and filming are not easy for private broadcasting.

In particular, for video editing, there are professional editing tools such as Premiere Pro, After Effects, Vegas, Final Cut, and Editus. It is difficult to use for personal use. In addition, a free tool allows you to shoot and edit even in a mobile environment, and easily edit videos on a PC. However, in order to edit the video, it takes a lot of time and storage space to cut out unnecessary parts while watching all the recorded videos, and to combine necessary parts.

This paper attempts to implement a deep learning-based video face, food object, and pose recognition solution based on edge computing for various images. Currently, there are various deep learning-based open sources, but there is currently no integrated recognition solution for automatic media image editing.

This paper is about an integrated automatic object and pose recognition technology for a media image editing solution. In particular, as an example, after recognizing face objects, food objects, and specific poses occurring in media images at the same time, it includes a function to log the recognized results in real time every second. Therefore, in order to make an automatic video editing solution, this paper focuses on real-time realization by integrating various recognition technologies, and includes the following two key element technologies.

- Deep learning based integrated recognition edge data processing technology for automatic media image editing
- Deep learning based integrated recognition deep learning technology for automatic media image editing

II. RELATED WORK

A. Research on face recognition based on deep learning

Recently, deep learning algorithms with high accuracy have been actively applied to face recognition technology. In the existing facial recognition algorithm, recognition was performed using a combination of high features defined by humans such as Harr-like features. There is a limitation in

finding the optimal feature shape and combination in the way that a person defines features in advance. In contrast, the deep learning algorithm finds the optimal feature shape and combination for the situation by itself, and shows higher performance than the existing algorithm.

The MTCNN algorithm is a deep learning algorithm that finds the location of the face and the five main points of the eyes, nose and mouth based on CNN. Three deep learning structures with different characteristics are sequentially applied to estimate the position of a human face in the form of a box, and five main points of the face are predicted inside the box. By utilizing this, it is possible to change collectively the face image of various postures to the posture desired by the user [1].

Recently, a deep learning-based algorithm such as DeepID is a technology that determines whether two face images are the same person. After receiving the face image as an input of the DeepID algorithm, a real vector value is output as the result. After that, the similarity of the face can be determined by calculating a distance value between vector values. The deep learning algorithm does not define the image feature shape in advance, but learns the case of the same person and not the same person by using hundreds of thousands of pictures, and the algorithm creates the optimal feature shape necessary to recognize a human face by itself. The learned algorithm can then compare only real vector values to determine whether they are the same person [2]

B. Research on object detection based on deep learning

As deep learning techniques based on the latest Convolutional Neural Networks (CNN) are applied to images, the performance of object recognition technology has dramatically improved. The early image-based deep learning algorithm was a technology that received a single image as an input and classified the entire image into which class, and the probability of belonging to each class was presented as an output value of the algorithm. The figure below shows the initial image-based deep learning algorithm [3].

Next, a deep learning algorithm was developed to present the positions of multiple objects and their respective classes in a single image. This is commonly called a box-based object recognition algorithm, and as the output value of the algorithm, the position values in the form of boxes for multiple objects and the probability that each box belongs to a specific class are presented [4].

In addition, research on weight reduction of deep learning is actively underway in recent years. Geoffrey Hinton et al. presented a technique that can track the performance of the original model even if the size of the model is reduced in a neural net environment. In order to simulate the model, the activation of the last stage was made to follow the teacher's output value while adjusting the parameter of the softmax value [5].

XNOR.ai, an American company that started at the Allen Institute of AI, a non-profit artificial intelligence research center in the United States, converted the matrix used in the existing CNN into a binary matrix and then performed a logic gate operation. Logic gate operation is the simplest and fastest operation that a computer can do, and if you implement deep learning through it, it can perform very quickly. Currently, the paper only shows the results based on ImageNet, and in the case of the multi-objective method, the results are provided as videos, but the algorithm has not been fully disclosed.

C. Research Trends Related to Deep Learning-Based Behavior Recognition

Various methods for estimating a person's pose have been introduced over the years, and it has been developed from estimating one person's pose to a method of estimating multiple poses. In order to estimate the pose, human skeleton information is often used, and each coordinate of the skeleton is called a joint or keypoint, and the entire pose is finally estimated by estimating the effective connection between the two parts.

OpenPose is one of the most used deep learning-based bottom-up approaches for estimating a person's pose, which first detects everyone's keypoints in the image. Next, by assigning key points to separate individuals, multilateral skeleton information can be extracted [7].

III. PROPOSED AUTOMATIC VIDEO EDITING SOLUTION

A. Automatic video editing solution and scenario

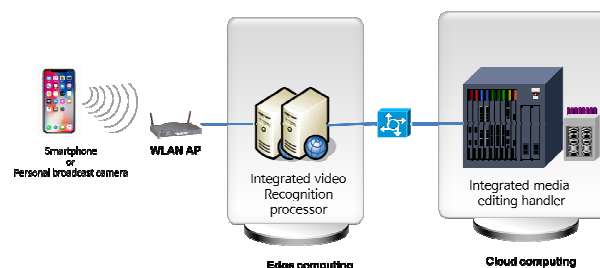


Figure 1. Overall configuration of automatic video editing solution

Figure 1 shows the overall configuration of the automatic video editing solution. The processing scenario of the automatic video editing solution is as follows.

- Step 1: The integrated video recognition processor in the edge computing device sets the face, food object, finger V motion, and applause motion to be selected in advance.
- Step 2: When a video is taken with a smartphone or a personal broadcasting camera, the video is transmitted to the edge computing device.
- Step 3: The integrated video recognition processor in the edge computing device sends the required tagging information and the original image from the original image to the cloud computing device based on the pre-set face to be selected, food object, V motion, and applause motion. .
- Step 4: The integrated media editor in the cloud computing device edits the original video based on the tagging information or performs additional editing desired by the user.
- Deep learning based video face detection: Recognition of a specific face in a video with a large number of people, and recognition of a large number of (5 people) faces in a video with a large number of people, real-time processing (within 0.1 seconds)
- Deep learning based video food object detection: detection of objects, object detection to extract food eating scenes, object detection to extract products appearing in the food room, recognition of around 100 objects simultaneously
- Deep learning based video pose detection: finger V sign gesture recognition, applause gesture recognition
- Integrated recognition setting GUI: GUI-based video object recognition selection function. Selective combination of the above three recognitions and tagging function for each detected scene by time, object type, object name, object coordinates, etc.

B. Integrated video recognition processor conceptual diagram and target function

Figure 2 shows the conceptual diagram of the integrated video recognition processor. By applying various deep learning technologies such as object detection by receiving video from personal broadcasting, only scenes in which specific people/objects appear, specific motions (V-shaped finger movements, applause, etc.) are extracted and the reconstructed video is output as output. to provide. Each input/output is as follows.

- Input: Original private video
- Output: Through the application of integrated video recognition technology, etc., the recognized tagging information is output based on the face, food object, and pose recognition (finger V motion, clapping motion) that the user wants to select.
 - Printed tagging information: time, recognition type (face, food, pose), recognition name, and box coordinates of the recognized result (top left: x_{min} , y_{min} , bottom right: x_{max} , y_{max})
 - Each recognition setting can be set to “AND” or “OR”.

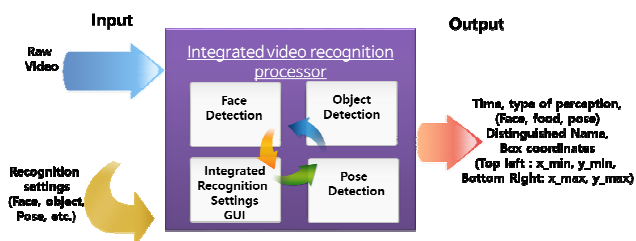


Figure 2. Concept of integrated video recognition processor

The target functions in this paper are as follows.

C. System structure diagram

This technology is a solution that provides integrated recognition results for automatic video editing.

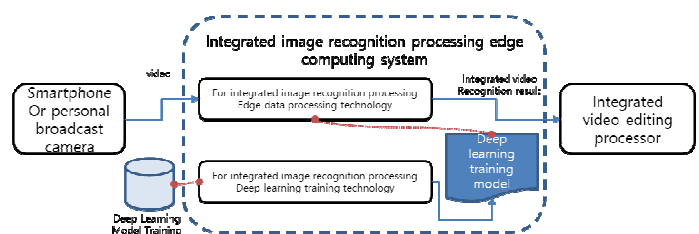


Figure 3. System structure diagram

The proposed technology is largely divided into “D. edge data processing technology for integrated image recognition processing” and “E. deep learning technology for integrated image recognition processing”.

D. Edge data processing technology configuration for integrated image recognition processing

The edge data processing technology for integrated image recognition processing consists of a base class module in charge of driving the entire system, common library modules supporting it, and individual recognition processing modules. The common library module is composed of Const, Data container, Guicomp, and log, and each recognition processing module is composed of mtcnn, deepid, tracker, deepfood, and deeppose.

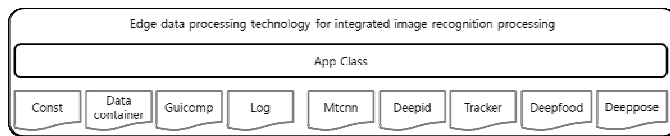


Figure 4. Module configuration of edge data processing technology for integrated image recognition processing

- Deep learning-based video face detection: Through the Mtcnn module, the position of the face is recognized. Based on the location of the face, the Deepid module finds the face most similar to the existing registered face. In order to prevent the face from moving and not being recognized on the video, it compensates for face recognition by using the Tracker module.
- Video food object detection based on deep learning: The Deepfood module uses existing object recognition technologies such as Yolo to identify the type and location of food. However, in the case of food when learning, the accuracy recognition rate for similar foods is low, so similar foods are grouped into the same type. For example, Galbitang and Seolleongtang improve accuracy by grouping them into the same food.
- Deep learning-based video pose detection: The pose recognition for extracting the applause scene identifies the position of the hand through the pose module. Judging by applause, seeing that the spacing between hands decreases and increases repeatedly. In order to extract the shape of the finger V sign, the finger V sign and position are identified using an existing object recognition technology such as Yolo.

E. Composition of deep learning training technology for integrated image recognition processing

Deep learning training technology for integrated image recognition processing is largely composed of a learning module for food recognition and a learning module for V sign recognition. However, deep learning models for face recognition and pose recognition use an existing trained model, and there is no need for new training, so separate training modules are not configured.

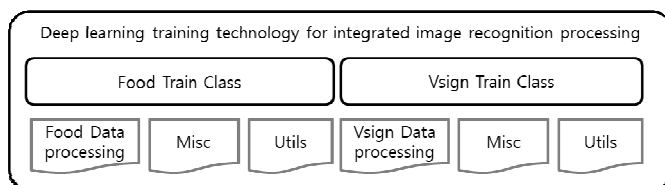


Figure 5. Module configuration of deep learning training technology for integrated image recognition processing

F. Integrated image recognition result

Figure 6 shows the result of integrated image recognition. The result log is displayed in one file. In each frame log, recognized names and coordinates for faces, foods, and poses by time are shown. At this time, if multiple faces appear, specify multiple faces in the face tag. Food and pose are also specified in the same way.

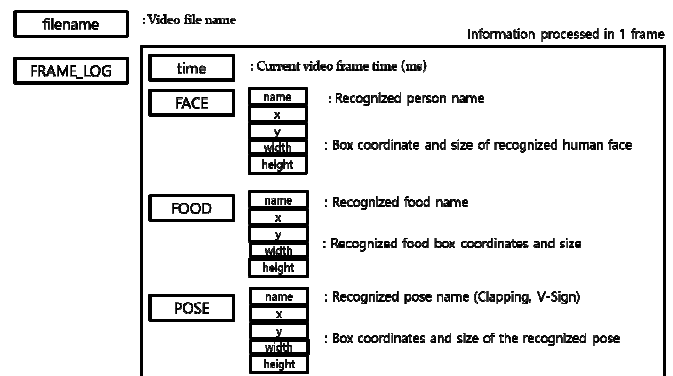


Figure 6. Integrated image recognition result

IV. CONCLUSIONS

This paper proposes a deep learning-based integrated recognition service for automatic image editing. Through this paper, we provide a basic technology that can easily edit only the desired video scenes by recognizing face recognition, food object recognition, and pose recognition in a real-time video stream. By applying this paper, personal broadcasters will be able to reduce video editing time.

ACKNOWLEDGMENT

This work was supported by the National Research Council of Science and Technology (NST) grant by the Korea government (MSIP) (No.CRC-15-05-ETRI).

REFERENCES

- [1] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.
- [2] Yi Sun, Xiaogang Wang, Xiaoao Tang, "Deep Learning Face Representation from Predicting 10,000 Classes," Proc. the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.
- [3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.
- [4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.
- [5] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, "Distilling the Knowledge in a Neural Network," NIPS 2014 Deep Learning Workshop
- [6] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Network", 2016
- [7] <https://github.com/CMU-Perceptual-Computing-Lab/openpose>



Kwihoon Kim is currently a professor in the Department of Artificial Intelligence Convergence Education, Korea National University of Education (KNUE), South Korea. He received the B.S, M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea in 1998, 2000 and 2019, respectively. He worked in LG DACOM 2000~2005. From 2005 to 2020, he was a Principle Researcher with Electronics and Telecommunications Research Institute (ETRI). He is an editor and rapporteur of ITU-T SG11 since 2006. His interested fields are Fog/edge computing, Internet of Things, 5G/IMT2020, deep learning, machine learning, reinforcement learning, GAN and knowledge-converged intelligent service.



Sewon OH is currently a Principal Researcher working for Knowledge-converged Super Brain(KSB) Convergence Research Department, Electronics and Telecommunications Research Institute(ETRI), South Korea. He received the B.S.(1999) and the M.S. degrees(2001) from Pohang University of Science and Technology (POSTECH), and received the Ph.D.(2018) from Chungnam National University (CNU), Daejeon, South Korea, respectively. Since 2001, he has been involved in several research projects on software platform (such as RFID Event Management, USN Middleware, IoT Platform) which integrates legacy applications with various data resources and sensors. His recent interested areas include anomaly detection, machine learning application, and knowledge-converged intelligent service.