

Shot Category Detection based on Object Detection Using Convolutional Neural Networks

Deokkyu Jung*, Jeong-Woo Son*, Sun-Joong Kim*

*Smart Media Research Group, ETRI, Daejeon, South Korea

jdk30621@etri.re.kr, jwson@etri.re.kr, kimsj@etri.re.kr

Abstract— Shot category detection is a method that extracts classified shots which comprised of several sequenced frames. These are based on object detection, which relates feature extraction of image processing. Shot category detection is helpful to utilize the neural network. Neural networks are widely generalized to our society. Artificial neural networks applied in many kinds of researches and developments, which provide convenient technologies for fundamental knowledge of deep learning. In terms of artificial neural networks, we introduce shot category detection based on object detection using convolutional neural networks (CNN). This paper also proves that CNN is efficient for supervised learning.

Keywords— Artificial Intelligence, Machine Learning, Deep Learning, CNN, Object Detection, Shot Category Detection.

I. INTRODUCTION

Shot [6] is one part of the movie scene that separated by constant frame. It consists of the featured character or object, and focuses proportion of one's component inside of screen. Shots are classified as picture's angle, which is important to emphasize for certain character depends on the motions and colors [7]. However, there has not been objected detection [8] using shot category. Among many kinds of shots, it has to classify these materials. Detection of shot category is also necessary to verify these shots. The purpose of shot category detection is finding out the different features among object detections.

It is easy to validate the shot detection such as simple shot produced by human camera. The simple procedure [3] of shot detection is not easy to classify for the basic categories. Simplified image cannot reflect on the object detection. The simplified image has lower resolution in terms of performance. Therefore, we want to build movie shot categories constructed by basic shots. With these, we detect shot categories for precise computer learning. These shots are important to treat concretely for object detection.

Shot category consists of several ranges of image size through the object detection [4], which is affected by certain extraction of image recognition. Shot corresponds to the object of characteristic point. Shot category requires constant sizes of image that can easily detect on input image rather than curved segmentation. Therefore, we want to utilize object detection of shot category.

Our result is that we made more than 90% of shot category detection, based on nine different categories using 10,000 datasets of tagging. The process of CNN is explained in section II, shot category detection is located in section III, the experiment of our work is detailed in section IV, and the last of this paper presents a conclusion in section V.

II. RELATED WORK

In the previous work [9], SVM based object category detection explains transfer learning to detect several images for regularization. However, the comparison between origin and transferred model is depicted as small amounts of image detection in one category, which is deficient to implement shot category detection. Classes are focused on the minimization of model's angle rather than the learning of object detection.

The find-grained recognition [10] is similar for our object detection. The boundary among pixels presents image segmentation, which normalizes the range of image feature. In the matter of shot category detection, the segmentation is improper to match our category selection.

CNN consists multilayer formats to process thousands or millions of neural networks [1]. Each layer has hidden layer to handle small amounts of layers. CNN reflects a human's perception how to memorize the specific things. Literally speaking, convolution makes complicated operations how machines learn. That means precise image detection is significant for convolution processes.

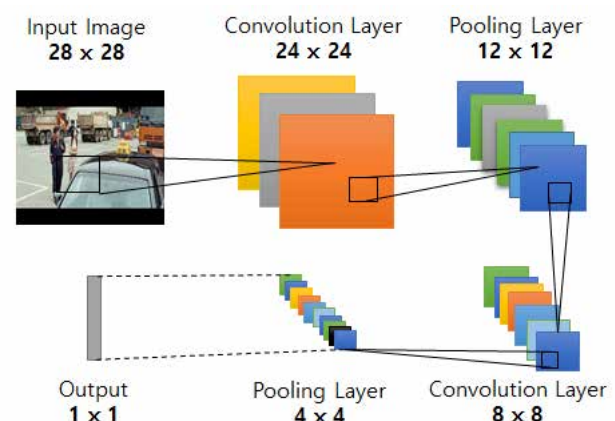


Figure 1. Feature extraction of CNN.

In Figure 1, the procedure of CNN is explained. Suppose the input image is 28x28. Through the 5x5 kernel, the input image is perceived by several feature maps for filtering hidden layers, which is made into 24x24 convolution layer. Convolution extraction is basically proceeded by neural activation function. These nodes are overlapped as same sizes of nodes. The capacity of convolution layers has to be lessened. In order to decrease some convolution layers, there is pooling layer for subsampling. Pooling also gives the input image correction. As the pooling works proportional rate of pixel, it consists of 12x12 pooling layer. These convolutions and pooling methods proceed one part of global feature. Also, the prior feature maps are proceeded again as feature extraction, which is made by 8x8 convolution layer. After that, the subsampling goes through 4x4 pooling layer. Thus, the last feature maps consist 1x1 output that fully connects layer. The output is expressed by means of the one of the classes.

III. SHOT CATEGORY DETECTION

Shot category detection is important for classes design and composition of object detection. In this section, we explain the conception of the shot, the detail of nine shot category classes, object detection of shot category and the utilization of shot category to detect by CNN.

A. Shot

The Shot is a portion of movie scene captured by the camera, which composes different aspects of image set. Shots are normally expressed as internal significance on the screen. These are able to assort the view of character. Basically, shots are featured from one's face to chest, waist or leg. Also, shots are stressed in background or special object, where the movie angle is decided by camera's distance.

B. Shot category

Shot categories are defined with point of view depending on the camera's angle [2]. The composition of shots is focused on normal background, object, and character. The character is only for human. The examples are shown in Figure 2.



Figure 2. Shot categories

1) **Long shot:** Usually called background. This shot is basically building, mountain, road, sea.

2) **Full shot:** It takes into whole body.

3) **Medium shot:** So called knee shot. The character's head until knee.

4) **Bust shot:** It displays from face to the middle of character's chest.

5) **Up shot:** This shot has to show the character's shoulder.

6) **Waist shot:** It is focused from head to waist.

7) **Close up:** The entire shot is shown by the face.

8) **Big close up:** This is focused on not only the specific object, but also hand, lips, and eyes. It is also known as extreme close up.

9) **Intro & Credit:** As the matter of movie scene, the start and end shot such as film distributor and plots.

C. Object detection of shot category

Object detection [4] finds consistent sizes of the image rather than basic image segmentation [5]. In the matter of segmentation, featured by several elements, images are captured by color pixels. It affects image resolution regardless of specific objects. In case of multiple objects inside one picture, that segmentation cannot focus on shot category. Otherwise, Object detection utilizes fixed size of pixel, so that objection recognizes easily to make feature examples. Constant size of object detection is efficient for shot category to fill in specific parts of the image.

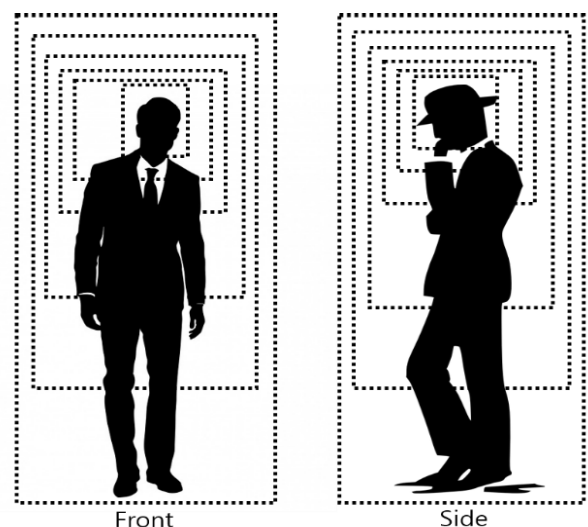


Figure 3. The basic range of object detection.

It is necessary to define the object detection range in shot category. In figure 3, close up is captured from person's hair to lower chin including round neck. Up shot has to show the same sizes in edge of shoulder. The border of bust shot ranges the center between neck and navel. Waist shot represents within wrist or waist. Lastly, medium shot is normally focused to the center of the knee including shin or thigh. Although the scope of object detection is defined in many kinds of insights, the general point is up to the angle of movie shot. No matter how the object presents front or side, object detection could be

rectangle, circle, or ellipse. Therefore, object detection is suitable for shot category.

D. Detection over CNN

Shots consist of varied angles of object detection. To detect classified shots correctly in figure 4, we want to apply convolutional neural network (CNN) for the shot category detection. Due to convolution layer, it is easy to learn correctly from our shots. CNN has deeper capacity for automatic learning in this regard. Besides that, CNN advantages 3D data by storing ability. With the fully connected layer of CNN, feature maps are constantly produced in different connected nodes. In the matter of shot category, these shots are extracted as fixed size, which is convenient for CNN's learning. Although the neural system is complicated, CNN composes deeply characterized semantic object detection. It is suitable for fixed size to detect in shot category. Therefore, we chose CNN for the shot category detection.

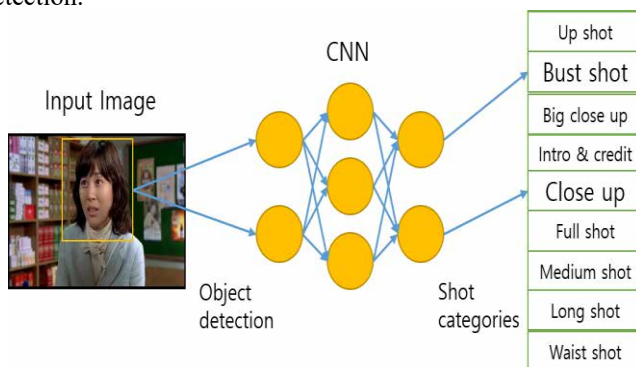


Figure 4. The procedure of shot category detection.

IV. EXPERIMENTAL RESULT

The experimental devices are listed as follows: Windows 10 64bit, Intel Xeon CPU E5-2695, 2.10GHz, 256GB of RAM. To examine shot categories, we constructed with nine classes, 10,000 shots of 57 Korean movies.

TABLE 1. TAGGING DATA.

Shot Categories	Training data	Test data
Long shot	695	180
Full shot	523	152
Medium shot	408	293
Bust shot	1,636	279
Up shot	1,642	308
Waist shot	598	301
Close up	1,141	173
Big close up	1,059	216
Intro & Credit	298	98

In TABLE 1, we diversified 10,000 shots into nine classes. We collected these shots individually and grouped nine classes. Next, we adjusted basic 32x32x3 (height, width, depth) image. In the global constants, we divided datasets into 8:2 rate. Finally, the important thing is the initial learning rate (0.02~0.1) to adapt to learning parameters.

As we can see in Figure 5, precision rate resulted from 90% to 99.9%. Our precision rate depended on initial learning rates and max steps. Starting from 30,000 (max step) and 0.06 (initial learning rate), it resulted approximately 99% precision. Max step 60,000 was the better than other steps, where the consequence was less than 99% precision rate on 0.1 initial learning rate. Shot categories detection is mostly higher when the initial learning rate is increased. On the contrary, too much learning rate decreases the precision rate.

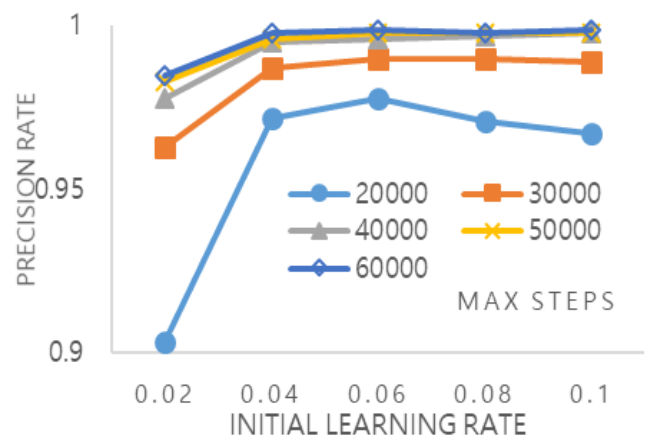


Figure 5. The precision rate of shot categories.

V. CONCLUSIONS

Overall, shot category detection is an effective way to learn various object detection using Convolutional Neural Networks. It is clear that more than 90% precision indicates the reliability in the supervised learning. The compositions of shot classes are concisely combined. We hope that movie shots are useful for detailed shot category detection. Shot categories contribute the performance of image detection, which is also affected in the deep learning neural network.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [17ZH9900, Development of programmable interactive media creation service platform based on open scenario]

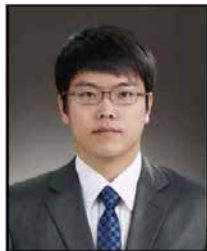
REFERENCES

- [1] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009).
- [2] Park, Dug-Chun. "Corelation of Television Drama's Rating and Image Production Techniques-Mainly on the Home Dramas." *The Journal of the Korea Contents Association* 8.9 (2008): 66-73. (written in Korean).
- [3] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." *Computer Vision and Pattern*

- Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on.* Vol. 1. IEEE, 2001.
- [4] Papageorgiou, Constantine, and Tomaso Poggio. "A trainable system for object detection." *International Journal of Computer Vision* 38.1 (2000): 15-33.
 - [5] Vilarinho, David López, et al. "Discrete-time CNN for image segmentation by active contours." *Pattern Recognition Letters* 19.8 (1998): 721-734
 - [6] Shot, [https://en.wikipedia.org/wiki/Shot_\(filmmaking\)](https://en.wikipedia.org/wiki/Shot_(filmmaking))
 - [7] Basics of Video Shooting, https://www.asu.edu/alti/ltlab/tutorials/video/basics/type3_mls.htm
 - [8] Martinez, Manuel, Alvaro Collet, and Siddhartha S. Srinivasa. "Moped: A scalable and low latency object recognition and pose estimation system." *Robotics and Automation (ICRA), 2010 IEEE International Conference on.* IEEE, 2010.
 - [9] Aytar, Yusuf, and Andrew Zisserman. "Tabula rasa: Model transfer for object category detection." *Computer Vision (ICCV), 2011 IEEE International Conference on.* IEEE, 2011.
 - [10] Angelova, Anelia, and Shenghuo Zhu. "Efficient object detection and segmentation for fine-grained recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2013.



Deokkyu Jung received his B.S in Computer Engineering from Korea Polytechnic University, Republic of Korea in 2014. He received his M.S. in Computer Science and Engineering from UNIST (Ulsan National Science and Technology), Republic of Korea in 2017. He joined ETRI, Republic of Korea in 2017, where he is currently research intern. His research areas are deep learning, media platform, and distributed system.



Jeong-Woo Son received his MS and Ph.D. degrees in computer engineering from Kyungpook National University, Daegu, Rep. of Korea in 2007 and 2012 respectively. Since 2013, he has been with ETRI, Daejeon, Rep. of Korea. He focuses on machine learning, NLP, and information retrieval.



Sun-Joong Kim received her BS degree in computational statistics and her MS degree in computer science from Chungnam National University, Daejeon, Rep. of Korea, in 1989 and 2000 respectively. In February 1989, she joined ETRI, Daejeon, Rep. of Korea, where she is currently principal researcher and director. Her research interests include convergence service control, smart TV, content knowledge mining.