Pitching Classification and Habit Detection by V-Net

Sota Kato and Kazuhiro Hotta
Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan

Keywords: Habit Detection, Video Classification, Pitching Classification, V-Net, Grad-CAM.

Abstract: In this paper, we propose a method that is classified pitching motions using deep learning and detected the habits of pitching. In image classification, there is a method called Grad-CAM to visualize the location related to classification. However, it is difficult to apply the Grad-CAM to conventional video classification methods using 3D-Convolution. To solve this problem, we propose a video classification method based on V-Net. By reconstructing input video, it is possible to visualize the frame and location related to classification result based on Grad-CAM. In addition, we improved the classification accuracy in comparison with conventional methods using 3D-Convolution and reconstruction. From experimental results, we confirmed the effectiveness of our method.

1 INTRODUCTION

In recent years, the field of image recognition has greatly developed, and it is expected the practical use in various tasks such as automatic operation and factory automation. Such technologies can also be applied to the analysis of detailed human movements. The purpose of this paper is to classify pitching motions using deep learning and to visualize their habits, and to realize objective judgments rather than human subjective judgments.

In order to analyse human motion, it is important to visualize the reason why learning judged. In deep learning, there is Grad-CAM (R.R. Selvaraju and M. Cogswell, 2017) which is a method for visualizing the important location for classification. Grad-CAM is often used in still image classification. However, since video classification includes frames, Grad-CAM is not useful for video classification because only important location in all frames is obtained. We want to know both important frame and location in a video. For this reason, it is difficult to apply Grad-CAM to conventional video classification methods.

To address the problem, we propose a video classification method using V-Net (F. Milletari and N. Navab, 2016). The number of channels at final layer in V-net can be the same number of frames as an input video, so it is possible to visualize the important frame and location with some devices. In addition, the input video is reconstructed from the output of V-Net.

Figure 1: Example of habit detection. Left shows an input image and right shows the result of habit detection.

Figure 1 shows the example of visualization. In experiments, we classify the pitching of three types of balls. If a video is classified well, we visualize the reason why V-Net classified the video into the class. The important location and frame are related to the habit of pitching of the person. As a result of classification of three kinds of pitching motions using V-Net, we achieved higher classification accuracy than 3D-Resnet (S. Ji, W. Xu, 2013) which has been used in conventional video classification. In addition, from the result of visualization, we confirmed that CNN can detect important frame and location for classification.

The organization of this paper is as follows. First, section 2 describes related works. Section 3 explains the proposed method. Section 4 describes the details of data. Section 5 shows experimental results. Finally, in section 6, we describe the conclusion and future works.

Kato, S. and Hotta, K.
Pitching Classification and Habit Detection by V-Net.
DOI: 10.5220/0009347106150621
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 615-621
ISBN: 978-989-758-402-2; ISSN: 2184-4321

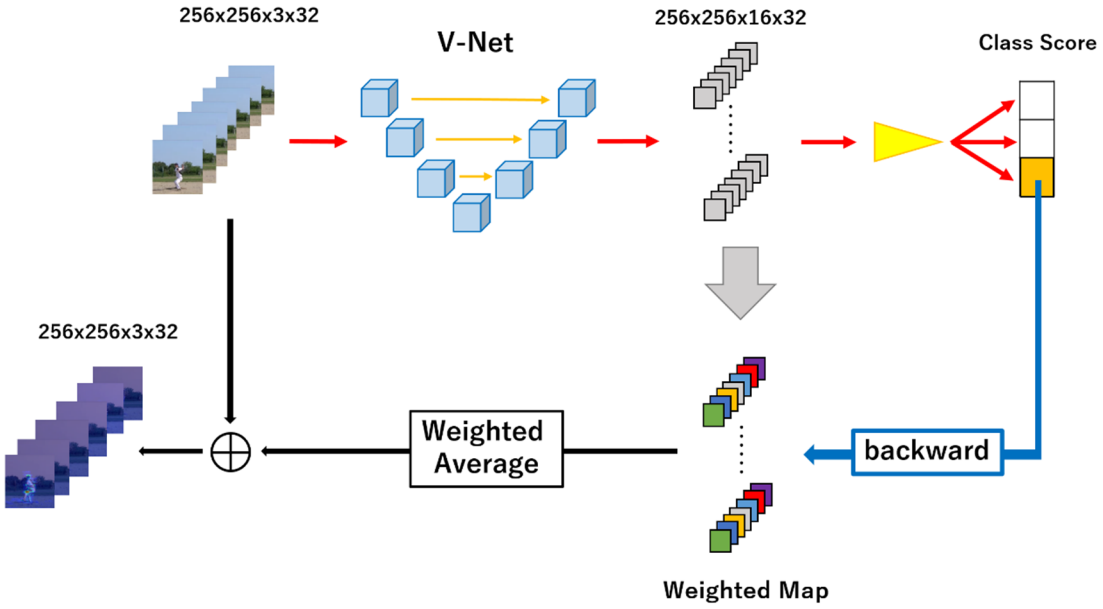

Figure 2: Overview of our proposed method.

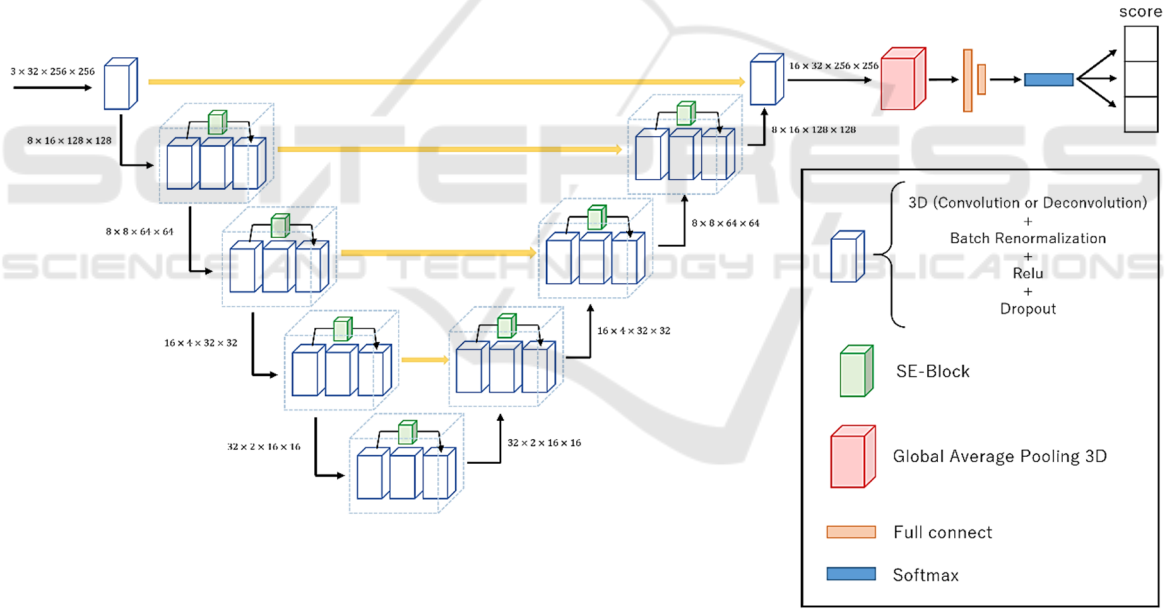

Figure 3: Network architecture of V-Net.

2 RELATED WORKS

2.1 Video Classification Method

In conventional video recognition methods, feature points such as STIP (Space-Time Interest Points) (I. Laptev and T. Lindeberg, 2003) and DT (Dense Trajectories) (H. Wang and A. Klaser, 2011) were used. Therefore, video classification methods using CNN are the trend (D. Tran1 and L. Bourdev, 2015) (S. Ji and W. Xu, 2013). Among them, the video recognition methods using 3D-convolution is paid attention because 2D-convolution cannot extract sequential features. Higher accuracy is achieved by CNN with 3D-convolution. One of the representative video classification method using 3D-convolution is 3D-Resnet (S. Ji and W. Xu, 2013). 3D-Resnet is the extended version of Resnet for 2D-images (K. He and

X. Zhang, 2017), and it is greatly improved by using recent huge video databases.

2.2 V-Net

V-Net is a segmentation method for volumetric medical images. U-Net (O. Ronneberger and P. Fischer, 2015) is the famous method for segmentation of 2D images. V-net is the extended version of U-net and it used 3D-Convolution to extract volumetric features. It was proposed as a segmentation method for volumetric images, but here we use it to classify pitching motions. Thus, the number of frames in the final output of V-net is set to the same frames as an input video. There is the possibility to visualize both important frame and location in a video.

2.3 Grad-CAM

Grad-CAM is a method that can visualize the location related to classification result. Conventional visualization methods for deep learning are guided backpropagation (R.R. Selvaraju and M. Cogswell, 2017) and deconvolution (M. D. Zeiler and R. Fergus, 2014) but these methods visualize and emphasize the part that reacted in image. They are not the part that contributed to classification.

Grad-CAM is improved the problem and visualized locations that have a large influence on the probability score of each class using the weighted average of feature maps. The weight is a coefficient that expresses the magnitude of the change that occurs in the probability score when a small change is made to a certain image location in the feature map. In other words, an image location with a large effect on class judgment has a probability score. The derivative of it is also large.

Equation (1) is the weight for the k-th filter of class **c**. The probability score $\boldsymbol{y^c}$ of class **c** is differentiated with respect to the value $\boldsymbol{A_{ij}^k}$ of pixel (i, j) in the k-th feature map, and the gradient is obtained. We averaged each feature map over all pixels. The larger the weight, the more important the feature map $\boldsymbol{A_k}$ for class c. In equation (2), the weighted average of the k-th filter is calculated in equation (1), and the output of ReLU function (max{x,0}) is the heat map. By taking the average of the slopes in equation (1), it is possible to clearly differences between classes.

$$a_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

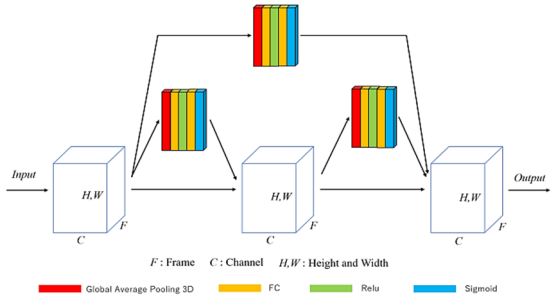

Figure 4: Architecture of SE-block.

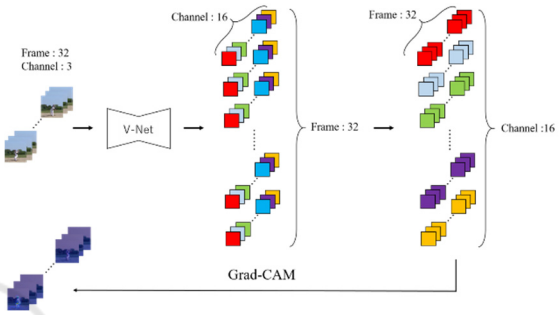

Figure 5: How to detect habit of pitching.

$$GradCAM = ReLU\left(\sum_k a_k^c A^k\right) \tag{2}$$

Grad-CAM is effective for visualization of CNN for 2-D images. It is not applicable for video classification because Grad-CAM visualizes only important location in an image. Since we want to know both important frame and location in a video. Thus, we use the V-net so that the number of outputs corresponds to the number of input frames. However, there is no guaranty that frames at final layer correspond to the frames on input video. Thus, we use reconstruction between the output and input frames. After those devices, we apply Grad-CAM to visualize the habit of pitching.

3 PROPOSED METHOD

Figure 2 shows the overview of our proposed method. An input video of 256×256 pixels and 32 frames is fed into the V-Net. As described previously, the number of channels of output layer is set to the same number of frames in an input video. Concretely, the output of V-Net is 256×256 pixels and 32 frames which is the same as the input video. In addition, the reconstruction is performed between the output layer

Figure 6: Example of throwing video frame.

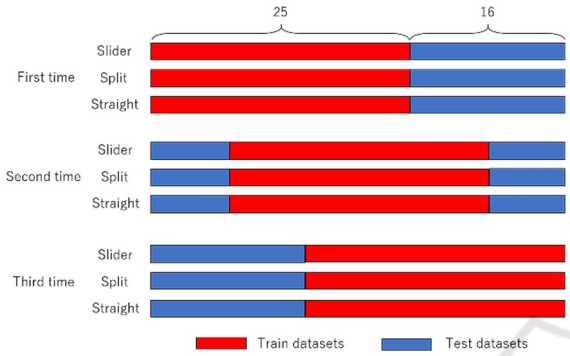

Figure 7: Cross Validation method.

and input frames in order to detect frame and location related to classification.

To classify the input video, global average pooling and fully connection layer are used. The highest probability of output is backward to V-Net and is calculated important features affected final probability score. Grad-CAM is generally used final convolution features of classification network and generated visualize heat map using these features. Similarly, we used final convolution features of the V-Net.

3.1 Details of Network Architecture

Figure 3 shows the details of the V-Net we used. V-Net consists of 7 blocks. In the encoder part, 3D-convolution, batch renormalization (S. Ioffe, 2017), Relu function and dropout are used three times for one block. The same processing is used in the decoder part as the encoder.

In addition, SE-block (J. Hu and L. Shen, 2018) is added to all blocks. Figure 4 shows the configuration of the SE-block. SE-block is composed of global average pooling (M. Lin and Q. Chen, 2014), fully connect layer, Relu and sigmoid functions. SE-block is a kind of channel attention. By assigning weights to each channel, the classification accuracy can be improved.

We use 3D-global average pooling (M. Lin and Q. Chen, 2014) whose size is 16×16×16 to reduce the dimension. The averaged features are fed into 2 fully connection layers and to output three classes.

We used softmax cross entropy loss for classification and mean least squares loss for reconstruction. Reconstruction is expected to make the outputs of V-net correspond to input frames. The equation of softmax cross entropy loss is shown in equation (3), and the equation for reconstruction loss is shown in equation (4). $P(x)$ is the correct label, $q(x)$ is the probability by softmax function, x_n is the input frame and $\hat{x}(x_n)$ is the predicted value. The final loss function is equation (5).

$$Cross\ entropy\ loss = -\sum_{x} p(x) \log q(x) \tag{3}$$

$$Reconstruction\ loss = \sum_{n=1}^{N} \|x_n - \hat{x}(x_n)\|^2 \tag{4}$$

$$Loss = Cross\ entropy\ loss + Reconstruction\ loss \tag{5}$$

3.2 Habit Detection

Figure 5 shows how to detect the habit of pitching by our method. The output of our V-Net is channel × frame × size (height and width). Since Grad-CAM is calculated by the weighted sum of each channel, we cannot use Grad-CAM directly. Therefore, the same channel which corresponds to each frame is collected by switching channels and frames as shown in Figure 5. For example, the feature map which corresponds to the first frame is shown as red. By using 3D-convolution, some feature maps are obtained, and the feature maps which correspond to the first frame are concatenated as shown in Figure 5. Then we can apply Grad-CAM to the feature maps, and the habit of pitching is detected.

4 DATASET AND EVALUATION METHOD

A baseball player throws three types of balls; straight, slider, and split. We captured videos of it with a fixed camera. Figure 6 shows the example of the pitching videos. The video size is 256×256 pixels, and the number of frames is 32 frames. 41 videos are captured for each ball type. 25 videos are used for training and 16 videos are used for test.

In experiments, we evaluate our method using cross validation. Figure 7 shows Cross Validation we used. Training data and test data were randomly divided. Experiments are performed three times, and

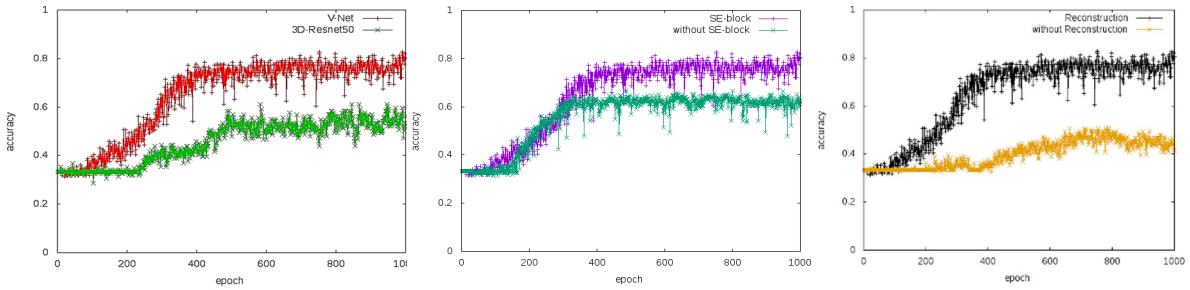

Figure 8: Results of cross validation. Left column shows the comparison between conventional method (3D-Resnet50) and proposed method (V-Net), Middle column shows the comparison our method with/without SE-block, and right column shows the comparison between our method with/without reconstruction loss.

Figure 9: Input frame (top) and reconstructed image (bottom).

the average accuracy is used as evaluation measure. 3D-Resnet50 (S. Ji and W. Xu, 2013) which is the famous video classification method is also evaluated as a comparison. We trained our method and comparison method with full scratch manner.

5 EXPERIMENTS

Figure 8 shows the graphs of classification accuracy. The left column shows the comparison between the conventional 3D-Resnet50 and our proposed method with SE-block and reconstruction. The red line shows the proposed method, and green line shows 3D-Resnet50. This graph shows that the proposed method improved more than 20% in comparison with the conventional 3D-Resnet50. The number of convolutions used in 3D-Renet is 50 layers, but our proposed method is used only 23 convolutional layers. The size of proposed network is about half in comparison with 3D-Resnet. Our proposed method gave better accuracy with a small number of convolutional layers.

The middle column shows the comparison between our proposed method with/without SE-block. Purple line is our method with SE-block and green line is our method without SE-block. From the graph, we confirmed that SE-block improved the accuracy about 20%. The channel attention also works well in video classification with 3D-convolution.

The graph on the right column shows the comparison between our method with/without reconstruction loss. The black line shows our method with reconstruction loss and yellow line shows our method without reconstruction loss. From the graph, we can see that the accuracy is improved about 40% by using reconstruction loss because relationship between frames is consistent by reconstruction loss. Figure 9 shows the example of reconstructed images by the proposed method. We confirmed that reconstructed images correspond to each frame.

Figure 10 shows the visualization results of the important frame and location for classification. Red shows the most important for classification and blue shows unimportant. In the correct slider class, habit is detected at pitcher's left foot, left hand, and right elbow when he releases a ball. In the split class, the habit is around the left hand glove. The habit of straight class focuses on his back and the right arm during takeback. From these results, we see that three ball types are classified by looking at different places.

Figure 11 shows the visualization results when our method miss-classified. Top row shows the case that the correct class is slider but the video is miss-classified into split. We see that heat map responds strongly to the left hand glove that reacts to the split. Bottom row shows the case that the correct class is split but our method miss-classified into straight. We see that it reacts to the back reacting to straight and the right hand during takeback. From the result of visualization, we confirmed that our method focused on the important part to classify each ball type in the case of miss-classification.

Figure 10: Results of habit detection. From top to bottom shows slider, split and straight. The same frames are shown among three classes.

(1a)

(1b)

(2a)

(2b)

Figure 11: Visualized results of the videos when our method miss-classified. (1a) V-Net predicted slider correctly (1b) V-Net predicted split though correct class is slider, (2a) V-Net predicted split correctly (2b) V-Net predicted straight though correct class is split.

6 CONCLUSION

In this paper, we proposed pitching classification method using V-Net with reconstruction, and habit detection is performed based on Grad-CAM. It can provide higher accuracy than the conventional video classification method by using reconstruction and SE-block. By using our proposed method, we can understand important movements. Thus, our method will be useful to the analysis of human movements.

However, some results of habit detection include an ambiguous heat map or a blurred heat map. Since the improved version (A. Chattopadhyay and A. Sarkar, 2018) of Grad-CAM has been proposed, we would like to try it in the future to make visualized images clearer. Furthermore, we would like to confirm the effectiveness of the proposed method by applying our method to other video classification datasets.

REFERENCES

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization", In Proc. *International Conference on Computer Vision*, pp. 618-626, 2017.

F. Milletari, N. Navab, S.A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation", In Proc. *International Conference on 3D Vision*, pp. 565-571, 2016.

I. Laptev, T. Lindeberg, "Space-Time Interest Points", In Proc. *International Conference on Computer Vision*, pp.432-439, 2003.

H. Wang, A. Klaser, C. Schmid, C.L. Liu, "Action recognition by dense trajectories", In Proc. *IEEE Conference on Computer Vision and Patern Recognition*, pp.3169-3176, 2011.

D. Tran1, L. Bourdev, R. Fergus, L. Torresani, M. Paluri1, "Learning Spatiotemporal Features with 3D Convolutional Networks", In Proc. *International Conference on Computer Vision*, pp.4489-4497, 2015.

S. Ji, W. Xu, M. Yang, K. Yu, "3D Convolutional Neural Networks for Human Action Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume, Vol.35, pp.221-231, 2013.

K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. In Proc. Computer Vision and Pattern Recognition, pp.770-778, 2016.

O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", In Proc. *Medical Image Computing and Computer-Assisted Intervention*, pp.234-241, 2015.

M. D. Zeiler R. Fergus, "Visualizing and understanding convolutional networks", In Proc. *European Conference on Computer Vision*, pp.818-833, 2014.

S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models", In Proc. *Neural Information Processing Systems*, pp.1942-1950, 2017.

J. Hu, L. Shen, G. Sun, "Squeeze-and-Excitation Networks", In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, pp.7132-7141, 2018.

M. Lin, Q. Chen, S. Yan, "Network in network," In Proc. *International Conference on Learning Representations*, 2014.

A. Chattopadhyay, A. Sarkar, P. Howlader, V.N.Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks", In Proc. *Winter Conference on Applications of Computer Vision*, 2018.