Proposing a classification method of human behavior focusing on orbital characteristic on feature space in a neural network

Masafumi CHIDA, Keita SATO, Nahomi M. FUJIKI and Yoshihiro HAYAKAWA

National Institute of Technology, Sendai College Hirose Campus
4-16-1 Ayashi-chuo, Sendai, 989-3128, JAPAN

*Corresponding Author: fujiki@sendai-nct.ac.jp

Abstract

We propose a method using an auto encoder multilayered neural network that is expected to have the ability of classifying and extracting small scale difference in human behavior. The video image of ping-pong swing motion was recorded by a high-speed camera and analyzed differences in player's movement corresponding to each different cluster obtained by k-means method applied to orbital data in the feature space of the network. We confirmed that the orbits in the feature space represent some kind of characteristics of swing motions and showed the possibility of automatic detection of human behavior by focusing on orbital features created in the middle layer of an auto encoder multilayered neural network.

Keywords: The feature space, A multilayered neural network, Clustering methods.

1. Introduction

A multilayered neural network is the base of Deep learning technology that is used for various purposes such as data mining, automatic data analysis, image recognition and so on⁽¹⁾. Identical mapping on a multilayered neural network may allow us to extract abstract characteristics of input data appearing in its feature space⁽²⁾⁽³⁾. Such a neural network is well-known as an auto-encoder. Identical mapping on a multilayered neural network is realized by presenting the input data to output neurons as its training data. If we reduce the number of middle layer's neurons in such an auto encoder, it is possible to convert the information contained in multidimensional input data into the low dimensional feature space defined in the middle layer of a multilayered neural network.

In 1995, Shirai et.al. showed that the applying

identical mapping learning of an auto encoder multilayered neural network and reducing the information of spectroscopic reflection rate data into 3 neurons in the middle layer of the neural network succeeded to indicate the characteristic of Munsell Hue circle represented in its middle layer⁽⁴⁾.

Their result suggests us that we might be able to extract characteristic of data which is continuously changing in time such as each flame image cut out from a video image of human behavior by using such a multi-layered neural network called an auto encoder these days.

A set of time-series data is mapped as a series of corresponding points onto feature space, so that images of human behavior are also mapped as a series of points corresponding to the time sequence of data in the feature space of an auto encoder. It is known that tracing such series of points draws an fluid orbit in the feature space. When we focus on human behavior, some of those orbits may represent a certain specific meaning corresponding to the certain kind of behaviors⁽⁵⁾. However, it is difficult to figure out visually through video images what kind of difference in actual behavior is represented by each orbit. Furthermore, a neural network processes input data nonlinearly and there is no direct translation of encoded information shown in the feature space of the network. In order to find out the meaning represented by these orbits, it is necessary to analyze video images in detail. We apply clustering methods for orbital data to classify and examine video images corresponding to these orbits and try to establish a way of analyzing human behavior.

In this paper, we chose swing actions of a ping-pong player as the target behavior that consists of limited motions in a limited area. Video images of swing motions are recorded by a high-speed camera and are used as input data of a multilayer neural network and training output data

for supervised learning as well. Furthermore, we are also observing ping-pong ball placements returned in a rally and evaluate expectation values of return area prediction.

Fig. 1. The basic structure of an auto encoder 7-layered NN and input images.

2. Extraction of specific feature by use of neural networks

In order to examine the classification ability of the auto-encoder composed of multilayered neurons, we analyzed video images of swing motions of a ping-pong player.

Picturing forehand and backhand swings in ping-pong rallies was made from the left side of a player by a high-speed video camera (Sony Action CAM HDR-AS 300). Each frame of images is a size of 1280×720 pixels. This video camera has 120 fps recording ability.

We converted each frame image to a gray scale image with 100×100 pixels and created static images for input data. Identical mapping of these static images is performed on a multilayered neural network. Then specific features of swing motions shown in a center layer of the network were extracted and analyzed the information appearing in the feature space of the center layer.

We provided a 7-layered neural network shown in Figure 1 and operated identical mapping of each flame image extracted from the video image of swing motions. The network has only 3 neurons in the 4th layer to form 3-dimensional feature space. It is expected that each flame will be represented by one single point in this 3-dimensional feature space. If each flame is input in time order, points corresponding to the sequence of flames should form an orbit in the 3-dimensional feature space.

We adopted MSGD (Momentum Stochastic Gradient Decent) as an optimizer for training the network and set the learning constant $\eta=0.5$, the momentum coefficient $\alpha=0.8$.

Sigmoid function and ReLU function are tried as an activation function and we examined the feature of each orbit appearing in its feature space. It is clear from these orbits shown in Figures 2 that Sigmoid function is more suitable for our purpose. Learning efficiency of mini batch learning is also examined with changing the size of a batch from 1 to 120. Consequently, we decided to adopt one batch process learning for the purpose of this paper.

(b)

Fig.2. An example of orbits in the feature space of the neural network using Sigmoid function (a) and ReLU function (b) as the activation function

Fig. 3-(b). Learning curve of the auto encoder 7-layered NN for 10,000 epochs.

Fig. 3-(b). Samples of orbits created in the feature of the auto encoder 7-layered NN after 10,000 learning epochs.

3. Classification of orbits corresponding to swing motions

Fig. 4. Orbits corresponding to 100 training datasets of swing motions created in the feature space

We provided 120 static images clipped out from video image data for each swing.

Totally image data of 290 swings including 210 forehand swings and 80 backhand swings is provided as the input data of the network.

The structure of a multilayered neural network that we used for this experiment consists of 7 layers with 100×100 neurons in input and output layers and 3 neurons in the

center layer which is supposed to create a feature space. After 10,000 epochs of the training for 100 swings with 81 forehand and 19 backhand swings, we input all of 290 swings into the network and analyzed characteristics of orbits appeared in the feature space (Fig. 4). A clustering method often helps to categorize a set of data. The k-means method is one of strong clustering methods. The optimal number of k for the orbital data in the feature space was determined by use of Elbow method. The result of clustering using k-means method is shown at k=4 in Figure 5. Analyzing details of each cluster indicated the explicit difference between forehand and backhand swings represented by difference of clusters but also nonobvious differences in forehand swings are mainly classified into 3 clusters automatically. Careful examination of original forehand swing video images corresponding to each cluster suggested that difference of body position and angle of the player also affected to orbits in the feature space. We observed in the cluster 1 in Figure 5 a similar behavior that the player was rather staying at the same position during the rally.

Fig. 5. The result of X-means method applied to data points in the feature space for 290 swing datasets. Slant line bars are for forehand swings and a meshed bar for backhand swings.

4. Relation between ball placements and clusters of orbits in the feature space

In order to find out the meanings depending on each cluster that is hard to notice visually, analysis of the relation between ball placements and clusters was made with increasing the number of clusters of k-means method. Focusing on a ball position extracted from video images

may help us to find out the meaning of encoded results. We divided a table of a player's own side court into 4 areas and 9 areas for an opponent side court as shown in Figure 6. The player in this experiment is righthanded and we counted the number of balls which was returned from an area among 4 divisions on the player's own side and landed in an area among 9 areas on the opponent side for every cluster. Orbits in the feature space of the auto encoder NN corresponding to 290 forehand swing motions are divided into k clusters by k-means method and we examine ball placements for each cluster and estimate the probability of each course for a return ball. Some of results are shown in Figure 7. The average probability over clusters shown in the figure indicates the tendency that most of returns landed in the area E or F. However, some of clusters show characteristic features as shown in Figure 7 (b). The figure 7 (b) indicates that the cluster 1 yields quite high return probability at the area E and contrarily the cluster 12 has a high peak at the area I.

Fig. 6. Definition of divided table areas

These tendencies suggest us that prediction of ball returns might be possible from the analysis of orbits combining with clustering methods.

Expectation value of area prediction for ball placements is defined as the function of cluster number k by the following formula.

$$E_k = \sum_{x=I}^{IV} \sum_{c=1}^k P(x, c) \max_i T_{xc}(i)$$

Here, $P(x, c)$ is the joint probability that a ball corresponding to a cluster c lands on an area x among the four different areas from I to IV. $T_{xc}(i)$ is the conditional probability that the return in the cluster c is hit from an area x and lands at an area i among 9 divided areas in an opponent side court.

Expectation values estimated for the different number of clusters are shown in Figure 8. Naturally better prediction rate is realized as the cluster number k increases. However, it is recognized the large improvement of expectation at relatively small k. Possibly there might be the optimal number of k around 25 and some particular clusters of orbits in the feature space may represent certain movements corresponding to particular returns.

Fig. 7. Probabilities of ball placement returned from the area I for average over clusters and for two major clusters at k=4 (a) and k=15(b).

Fig. 8. Expectation value of area prediction for ball placements plotted as the function of the number of clusters

5. Conclusion

Proposition of a method using an auto encoder multilayered neural network that could classify and extract small scale difference in human behavior which would be hard to detect visually was made in this paper.

We tried to classify the video images of ping-pong swing motions recorded by the high-speed camera and analyzed differences in player's movement depending on each different cluster obtained by k-means method applied to target data in the feature space of the auto encoder multilayered neural network. Optimization and adjustment of parameters in its leaning processes lead the neural network converges to its minimum error state.

It is known that identical mapping for time sequential image data draws an orbit in the feature space of a multilayered neural network. We confirmed that input data of time-sequence swing images also formed various orbits in the feature space of the neural network. Applying k-means method to a set of orbits split the orbits into clusters of k. The clustering showed that the possibility of identifying swing movements is feasible from classifying such orbits in the feature space of a neural network.

Furthermore, analysis of such orbits focusing on placements of a pin-pong ball was considered. Expectation value to predict return placements was estimated as changing the number of clusters. It was shown that better prediction would be possible if optimal number of clusters is provided.

We confirmed that the orbits created in the feature

space of the auto encoder NNs represent some kind of characteristics in swing motions and showed the possibility of automatic detection of human behavior by using machine learnings based on neural networks.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Numbers 17K00356 and 16K01300.

References

- (1) T. Kamishima et. al. (Edit. JSAI) "Deep Learning" Kindaikagaku Co. (2016)
- (2) T. Ohnuma, Y. Hayakawa "Extracting continuous feature expression of handwritten characters", Proceeding of NOLTA Society, 367, (2016)
- (3) Yoshihiro Hayakawa, Takanori Oonuma, Hideyuki Kobayashi, Akiko Takahashi, Shinji Chiba, Nahomi M Fujiki, "Feature Extraction of Video Using Artificial Neural Network," International Journal of Cognitive Informatics and Natural Intelligence, Vol 11(2), pp.25-40, 2017.
- (4) Shiro Usui and Shigeki Uchinaka, "Theory of calculating color vision -- Color vision model in neural networks," Science, Vol. 65, No. 7, pp.469- 476, 1995-07. (Japanese)
- (5) M. Chida, K. Sato, N. M. Fujiki and Y. Hayakawa "Behavior classification method based on the feature space in a neural network" ICIEC Bulletin Technical Report, vol. 118, no. 498, NLP2018-128, pp. 23-26, 2019