DEEP-LEARNING-BASED HUMAN INTENTION PREDICTION WITH DATA AUGMENTATION

Shengchao Li[1], Lin Zhang[2], and Xiumin Diao[3]

[1]Arrow Electronics, Centennial, CO, USA
[2]Department of Physics & Astronomy, University of Central Arkansas, Conway, AR, USA
[3]School of Engineering Technology, Purdue University, West Lafayette, IN, USA

ABSTRACT

Data augmentation has been broadly applied in training deep-learning models to increase the diversity of data. This study ingestigates the effectiveness of different data augmentation methods for deep-learning-based human intention prediction when only limited training data is available. A human participant pitches a ball to nine potential targets in our experiment. We expect to predict which target the participant pitches the ball to. Firstly, the effectiveness of 10 data augmentation groups is evaluated on a single-participant data set using RGB images. Secondly, the best data augmentation method (i.e., random cropping) on the single-participant data set is further evaluated on a multi-participant data set to assess its generalization ability. Finally, the effectiveness of random cropping on fusion data of RGB images and optical flow is evaluated on both single- and multi-participant data sets. Experiment results show that: 1) Data augmentation methods that crop or deform images can improve the prediction performance; 2) Random cropping can be generalized to the multi-participant data set (prediction accuracy is improved from 50% to 57.4%); and 3) Random cropping with fusion data of RGB images and optical flow can further improve the prediction accuracy from 57.4% to 63.9% on the multi-participant data set.

KEYWORDS

Human Intention Prediction, Data Augmentation, Human-Robot Interaction, Deep Learning

1. INTRODUCTION

Humans can predict the intentions of others by observing their actions. We would also expect robots to be able to predict human intentions such that we can have safer and more efficient human-robot interactions [1][2][3], just like humans would do in collaboration with others. Besides human-robot interaction, human intention prediction is also the core technology for a variety of applications (e.g., rehabilitation devices to predict trainees' intention of slowing down [4], pedestrians' intention of crossing the road [5], driving assistance systems to predict drivers' intention of lane change [6], and surveillance and security to predict the intentions behind detected abnormal human activities [7]).

Human intention prediction is closely related to human action recognition but with a different purpose of classification. Action recognition [8][9] classifies different actions based on observed action sequences. However, intention prediction [10][11][12] predicts the intention of an action from the subtle motion patterns of the same action. Figure 1 illustrates the difference between action recognition and intention prediction. For action recognition, one needs to distinguish different action sequences, such as shooting an arrow and pitching a ball. The spatial and temporal patterns of these two action sequences are very different. For intention prediction, one predicts which target the participant pitches the ball to. These two pitching sequences have similar spatial and temporal patterns, which makes the intention prediction challenging.

DOI: 10.5121/ijaia.2022.13101

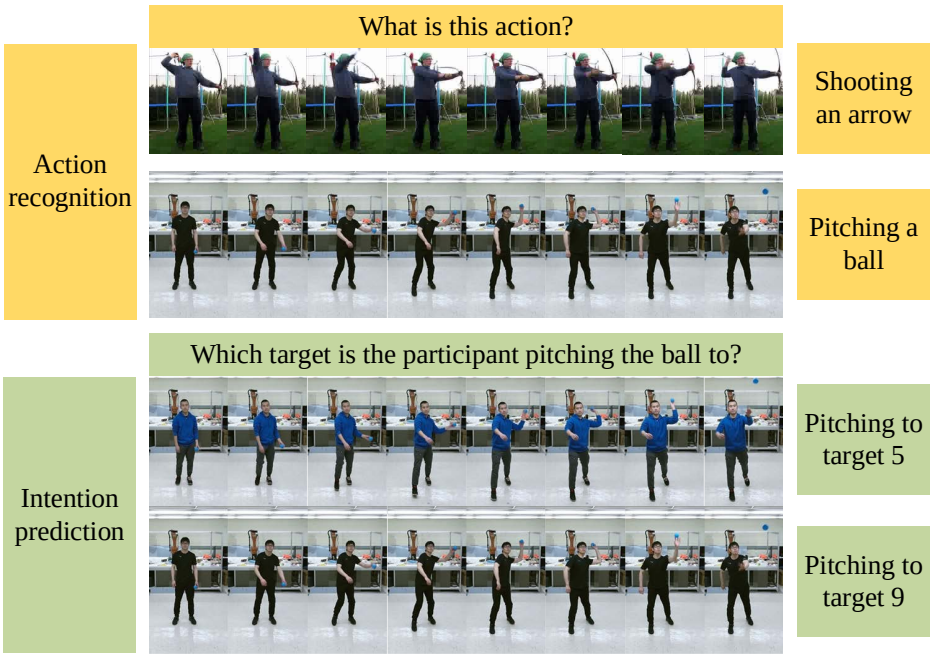

Figure 1. Difference between action recognition and intention prediction

Various approaches, such as Partially Observable Markov Decision Process (POMDP) [1], Markov Decision Process (MDP) [2], and neural networks [3], have been studied to predict human intentions in various scenarios. Furthermore, deep learning [13] has shown outstanding performance in various research areas (e.g., image recognition [14][15], object detection [16][17], and action recognition [18][19]). Deep neural networks are able to learn complicated representations in high-dimensional data for classification and prediction. In order to achieve satisfactory performance, massive data is required to train deep-learning models. However, it is usually complicated, and resource- and time-consuming to acquire sufficient data.

Fortunately, data augmentation offers an effective way to increase the quantity of training data. Data augmentation methods have been intensively studied for applications such as image classification [20], speech recognition [21], pose estimation [22], and action recognition [8][9] [23]. However, it is not clear whether data augmentation working for the above applications also work for intention prediction. Take Figure 2 and Figure 3 for example. Applying cropping on the original images changes the orientation and location of the human action (i.e., pitching a ball) in the images. For action recognition, one can still tell this action. However, for intention prediction, we still do not know how the changed orientation and location of the human action in the cropped images affect predicting the intention (e.g., the pitching direction) of the human action.

Figure 2. Images before cropping

Figure 3. Images after cropping

Our previous study [24] investigated the effectiveness of data augmentation on a single-participant data set. Since different people have different motion variations, it's still unknown whether data augmentation is effective for a multi-participant data set. Moreover, it's well known that optical flow is crucial for learning spatio-temporal patterns [9][19][23][25][26] which are important for intention prediction. However, to the best of our knowledge, how data augmentation is effective on the fusion data of RGB images and optical flow has not been explored for intention prediction. Motivated by the above three open issues, this study aims to answer the question - whether data augmentation is effective for intention prediction. Various data augmentation methods are evaluated using single-participant data set, multi-participant data set, and fusion data of RGB images and optical flow. The main contributions of this study are summarized as follows:

1) Intention prediction experiments show that not all data augmentation methods are effective for intention prediction. For the 10 data augmentation groups evaluated in this study, only the data augmentation groups that either crop or deform images can effectively improve the performance of intention prediction.

2) Random cropping can be generalized to the multi-participant data set. The accuracy of intention prediction is improved from 50% (without data augmentation) to 57.4% (with random cropping).

3) Random cropping also works on the fusion data of RGB images and optical flow. Random cropping attains the best prediction accuracy of 63.9% with fusion data of RGB images and optical flow on the multi-participant data set.

The organization of the paper is as follows. Section 2 presents the most relevant work in literature that dealt with intention prediction and data augmentation. The experiment setup is introduced in section 3. Seven data augmentation methods used for data augmentation experiments in this study are briefly introduced in section 4. The effectiveness of 10 data augmentation groups on a single-participant data set is investigated in section 5. Section 6 evaluates the generalization capability of random cropping on a multi-participant data set. Section 7 studies the effectiveness of random cropping on the fusion data of RGB images and optical flow. Finally, the paper is concluded in section 8.

2. RELATED WORK

This section briefly reports the most related work that tackled human intention prediction in various applications and data augmentation methods.

Human intention prediction has been widely investigated in various research areas (e.g., human-robot interaction, self-driving system, rehabilitation, and security). For human-robot interaction, various approaches have been proposed for human intention prediction to achieve safer and more efficient interactions. A neural network was used to predict human intentions in a table-carrying task [3]. A two-agent collaborative MDP approach [2] was proposed for anticipatory planning in human-robot teams. POMDP [1] was used to select anticipatory actions in human-robot table tennis playing. By analyzing the trajectories of human arms, the social intention prediction

problem was investigated in [27]. Intention prediction was also studied in [10] using motion information (i.e., kinematic data and video data) only. Besides human-robot interaction, human intention prediction has also been studied for applications such as driving systems [5][6], rehabilitation [4], and security [7].

Data augmentation has been used in many applications such as image recognition [20][28], speech recognition [21][29], SAR target recognition [30], environmental sound classification [31], singing voice detection [32], and pose estimation [22]. Data augmentation methods are usually categorized into two classes: pre-designed data augmentation methods [33][34] and learned data augmentation methods [35][36]. Pre-designed data augmentation methods use pre-designed signal processing functions to generate more data. For example, translation, color normalization, cropping, and deformation [37][38] are common pre-designed data augmentation methods for image data. Audio perturbation [21][29][39] (e.g., speech rate perturbation, speed perturbation, and reverberation) is usually used by pre-designed data augmentation methods for audio data. Pre-designed data augmentation methods have become a standard technique which can be implemented for most deep learning applications. Learned data augmentation methods usually learn a policy [40] or neural network [41][35] to generate new data. Compared with pre-designed data augmentation methods, learned data augmentation methods provide more variations of the generated data but are more complex to be implemented. Even though data augmentation has been studied for many applications [20][21][22][29][31][32], only our previous study [24] conducted preliminary research about the effectiveness of pre-designed data augmentation methods on intention prediction. Based on the previous study [24], this study further explores the effectiveness of pre-designed data augmentation methods on both multi-participant data set and fusion data of RGB images and optical flow for intention prediction using deep-learning models.

3. EXPERIMENT SETUP

3.1. Experiment Overview

The intention prediction experiment is illustrated in Figure 4. A human participant pitches a ball toward a robot (represented by the 9 targets in the figure). We expect the robot to be able to predict which of the 9 targets the participant pitches the ball to. Figure 4 shows the workflow of the experiment. Firstly, the motion of the participant is captured by a Kinect V2 camera (Microsoft Corporation, Redmond, WA, USA). The target hit by the ball is recorded. Secondly, a deep-learning-based prediction model is trained using the data collected in the previous step. Finally, new motion data is fed to the prediction model to predict the intention of pitching a ball.

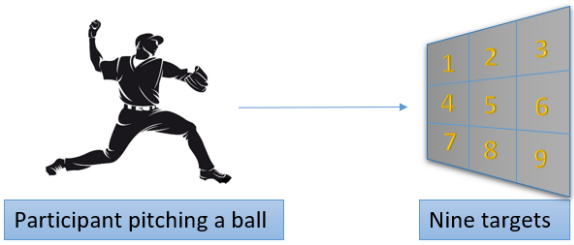

Figure 4. Experiment scenario

The workflow of the experiment is shown in Figure 5. Firstly, we employ a Kinect V2 camera (Microsoft Corporation, Redmond, WA, USA) to capture the motion of the participant. At the same time, we also record the target that is hit by the ball for each pitch. Secondly, the collected training data is used to train a deep-learning-based prediction model. Finally, we use the prediction model to predict human intention.

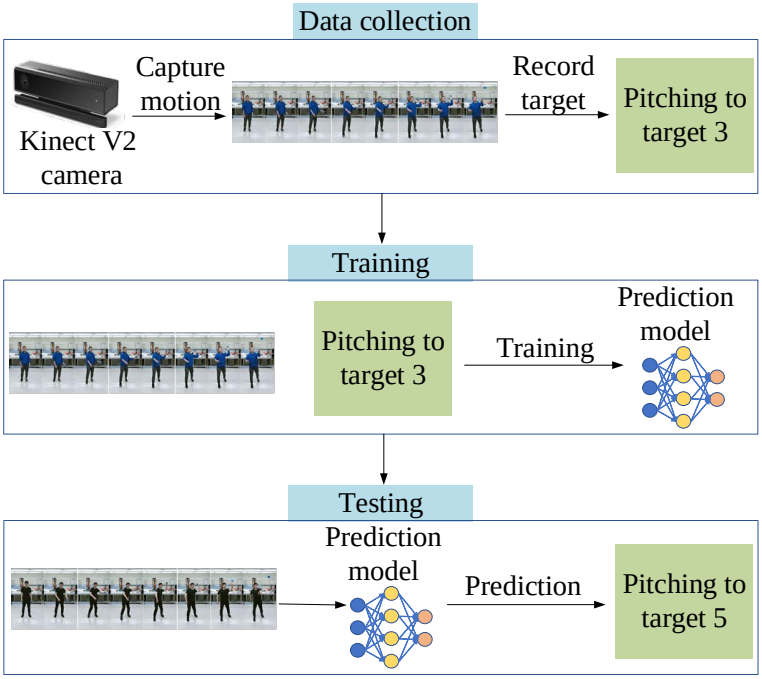

Figure 5. Experiment workflow

3.2. Data Collection Rules

For both single- and multi-participant data sets, we follow the same rules to collect data:

- It is a valid pitching trial if the ball hits the target the participant intends to hit. We save the data (i.e., the motion of the participant and the target hit) of this experiment to train or test the prediction model for human intention prediction.

- It is an invalid pitching trial if the ball hits the target the participant does not intend to hit. We save the data for future research.

- It is an invalid pitching trial if the ball hits none of the 9 targets (i.e., the ball hits outside the target area). We save the data for future research.

- It is an invalid pitching trial if the ball hits the border line between two targets. We discard the data.

4. DATA AUGMENTATION METHODS

This section introduces 7 data augmentation methods used in data augmentation experiments. Data sets are augmented by one or more data augmentation methods to assess the effectiveness of different data augmentation methods.

4.1. Contrast normalization

Applying contrast normalization to an image is to scale the pixel values of the image. With contrast normalization, one can create more new images from an original image. The pixel values after contrast normalization are represented by the Gamma Contrast function [42]

$$X_{i,j,k(cn)} = 255 \times \left(\left(\frac{X_{i,j,k}}{255}\right)^{\gamma}\right) \quad (1)$$

where i, j, and k are the row index, column index, and channel index of an image, respectively. $X_{i,j,k}$ and $X_{i,j,k(cn)}$ are pixel values of the original image and the new image after contrast normalization, respectively. $\gamma$is a parameter for contrast normalization. A comparison of the original image and a new image created with contrast normalization is shown in Figure 6.

Figure 6. Original image (left) and new image with contrast normalizaztion (right)

4.2. Gaussian noise

By adding Gaussian noise, a new image can also be created from the original image. A comparison of the original image and a new image created by adding Gaussian noise is shown in Figure 7. The pixel values of the new image are calculated as:

$$X_{i,j,k(gn)} = X_{i,j,k} + z \tag{2}$$

where $X_{i,j,k}$ and $X_{i,j,k(gn)}$ are pixel values of the original image and the new image with Gaussian noise, respectively. $X_{i,j,k}$ ranges from 0 to 255. $X_{i,j,k(gn)}$ is set to 255 if it is larger than 255 and 0 if it is smaller than 0. z is a variable representing random noise. The probability density function of z can be found in [43].

Figure 7. Original image (left) and new image with Gaussian noise (right).

4.3. Gaussian blur

Gaussian blur [44] is a method of blurring an image by convolving the image with a Gaussian kernel. The Gaussian kernel coefficients can be calculated as [44]

$$G_i = \alpha e^{\frac{-(i-(ksize-1)/2)^2}{2\sigma_{(gb)}^2}} \tag{3}$$

where $i = 0, 1, \ldots ksize - 1$, α is a factor guaranteeing that $\sum_i G_i = 1$, and $\sigma_{(gb)}$ is the Gaussian standard deviation. This function computes the $ksize \times 1$ matrix of Gaussian kernel coefficients. After the Gaussian kernel coefficients are obtained, each pixel after Gaussian blur can be calculated in each image channel [44]. Figure 8 shows the original image and the new image created with Gaussian blur.

Figure 8. Original image (left) and new image with Gaussian blur (right).

4.4. Random cropping

Applying random cropping to an image is to randomly crop a rectangular area of the image and then resize the cropped rectangular area to a certain size. Suppose the original image has a size of (I, J) and it is cropped with a parameter c. Then a random value a can be chosen from [0, c]. The size of the cropped rectangular area is calculated as $(\mathrm{I} \times (1-a), \mathrm{J} \times (1-a))$. Finally, the cropped rectangular area is resized to (224, 224). A comparison of the original image and a new image created with random cropping is shown in Figure 9.

Figure 9. Original image (left) and new image with random cropping (right).

4.5. Translation

One can also translate an image along either its row or column direction with a parameter $t \in$[-0.5, 0.5]. Randomly choose a row translation value t_1 and a column translation value t_2 from [$-t, t$]. t_1 and t_2 are the percentages of the height and the width of the original image, respectively. For example, t_2 = 0.5 represents 50% of the width of the image. After translation, only the pixels of the original image within the original image are kept. The pixels of the blank area are set to 0. A comparison of the original image and the new image with a translation is shown in Figure 10.

Figure 10. Original image (left) and new image with a translation (right).

4.6. Piecewise affine transformation

Piecewise affine transformation creates a new image by locally distorting the original image. More specifically, piecewise affine transformation places a regular grid of points on the image and randomly moves the neighborhood of these points via a normal distribution. Detailed implementation of the piecewise affine transformation was discussed in [45]. A comparison of the original image and the new image created with piecewise affine is shown in Figure 11.

Figure 11. Original image (left) and new image with piecewise transformation (right).

4.7. Perspective transformation

A new image can be created by applying a random four-point perspective transformation on an original image. Each of the four points is placed on the original image with a random distance from the respective corner of the original image. The random distance is sampled from a normal distribution. Detailed implementation of the perspective transformation was discussed in [42]. A comparison of the original image and the new image created with perspective transformation is shown in Figure 12.

Figure 12. Original image (left) and new image with perspective transformation (right).

5. Evaluation of Data Augmentation Methods on a Single-Participant Data Set

The effectiveness of data augmentation for intention prediction is evaluated with a single-participant data set in this section. One or more data augmentation methods are used as a data augmentation group. Totally, 10 groups of data augmentation methods are evaluated. Details of the single-participant data set, data augmentation methods, prediction model, and evaluation results are introduced in the following subsections.

5.1. Single-Participant Data Set

In the single-participant data set, one human participant conducts 292 valid pitching trials. 256 pitching trials are used for training the model of human intention prediction while the rest 36 pitching trials for testing the model of human intention prediction. Each pitching trial is recorded with 55 frames of RGB images. The single-participant data set is available at: https://github.com/deePurrobotics/Intention_Prediction.

5.2. Data Augmentation

We classify the data augmentation methods discussed in section 4 into two categories. Gaussian blur, Gaussian noise, and contrast normalization belongs to the first category while perspective transformation, piecewise affine transformation, translation, and cropping belongs to the second category. New images created using data augmentation methods in the first category are blurred but neither deformed nor cropped. New images created using data augmentation methods in the second category are either cropped or deformed, but their clearance is not changed. One or more

Table 1. Data augmentation methods and training data sets overview

Data Augmentation Groups	**Data Augmentation Parameters**	**Size of Data**	**Method Category**
Original RGB images	N/A	14,080 frames 938 MB	N/A
1. Random contrast normalization	Randomly choose γ from [0.75,1.5]	28,160 frames 1.88 GB	First
2. Random contrast normalization and Gaussian noise	Randomly choose γ from [0.75,1.5], adding Gaussian noise with probability of 50%, $\sigma^2_{(gn)} = 0.05 \times 255$	42,240 frames 3.62 GB	First
3. Random contrast normalization, Gaussian noise, and Gaussian blur	Randomly choose γ from [0.75,1.5], adding Gaussian noise with probability of 50%, $\sigma^2_{(gn)} = 0.05 \times 255$, $\sigma^2_{(gb)} = 0.5^2$	56,320 frames 4.93 GB	First
4. Random contrast normalization, Gaussian noise, Gaussian blur and Cropping (0, 10%)	Randomly choose γ from [0.75,1.5], adding Gaussian noise with probability of 50%, $\sigma^2_{(gn)} = 0.05 \times 255$, $\sigma^2_{(gb)} = 0.5^2$, $c = 0.1$	70,400 frames 6 GB	First and second
5. Cropping (0, 10%)	c = 0.1	70,400 frames 4.8 GB	Second
6. Cropping (0, 20%)	c = 0.2	70,400 frames 4.8 GB	Second
7. Translation (-10%, 10%)	t = 0.1	70,400 frames 4.8 GB	Second
8. Translation (-20%, 20%)	t = 0.2	70,400 frames 4.4 GB	Second
9. Translation (-30%, 30%)	t = 0.3	70,400 frames 4.4 GB	Second
10. Piecewise affine transformation and perspective transformation	pa_1 = 0.015, pa_2 = 0.045 p_1 = 0.025, p_2 = 0.075	70,400 frames 4.9 GB	Second

data augmentation methods in section 4 are used as a data augmentation group to augment the single-participant data set. As listed in Table 1, 10 data augmentation groups are investigated to evaluate their effectiveness on the single-participant data set. Take data augmentation group 1 for example. The original data set has 14,080 images. After random contrast normalization, we get 14,080 new images. Therefore, by combining the original images with the new images, we can obtain a training data set of 28,160 images.

Data augmentation methods in data augmentation groups 1~3 belong to the first category of data augmentation methods and those in data augmentation groups 5~10 belongs to the second category of data augmentation methods. Data augmentation group 4 belongs to both the first and

the second categories of data augmentation methods. The original images and the corresponding new images created using example data augmentation methods are shown in Figures 13∼16.

Figure 13. Original images.

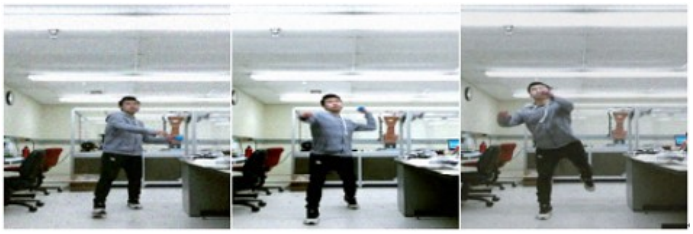

Figure 14. New images after applying random contrast normalization, Gaussian noise, Gaussian blur, and Cropping (0, 10%).

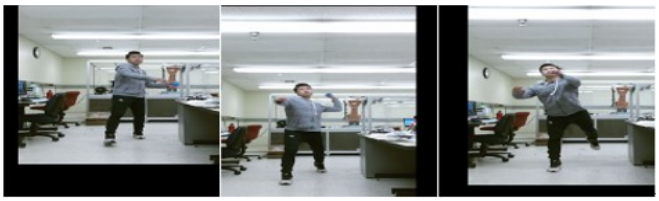

Figure 15. New images after applying Translation (-30%, 30%).

Figure 16. New images after applying piecewise affine transformation

5.3. Prediction Model

AlexNet [14] is modified and used as the baseline model for human intention prediction in this study. The Softmax classifier of AlexNet is changed to 9 classes. For each image in a pitching trial, a prediction is produced by the prediction model. By voting using the prediction from each image, one obtains the final intention prediction for the whole pitching trial.

1) Architecture of AlexNet for intention prediction: Figure 17 shows the architecture of AlexNet used for intention prediction in this study. This network has 5 convolutional layers and 3 fully connected layers in total. The neural network accepts an input tensor size of 224×224×3. Using short notation in [46], *Conv*(*d*, *f*, *s*) is a convolutional layer that has *d* filters with a stride of *s* and a spatial size of *f*×*f*. *N* represents a normalization layer. *P* represents a pooling layer. *FC*(*n*) represents a fully connected layer with *n* nodes.

2) Voting method: The voting result of a pitching trial can be calculated using n predictions from *n* frames of images:

$$P = \text{argmax}\left(\sum_{i=1}^{n}\left(\gamma^{n-i} \times sign(p_i)\right)\right) \quad (4)$$

where p_i is the prediction of the ith frame in a pitching trial, γ is the discount factor that weighs more for the later frames in the pitching trial, P is the final prediction result of the pitching trial. n = 45 and γ = 0.9 are used in experiments.

The prediction model is trained with the single-participant data set using stochastic gradient descent. The weights of the neural network are initialized from the Gaussian distribution $N(0, 0.001)$. The learning rate is initialized to 0.01. Each pooling unit has a stride of 2 and a size of 3×3. Dropout is applied with a probability of 0.5.

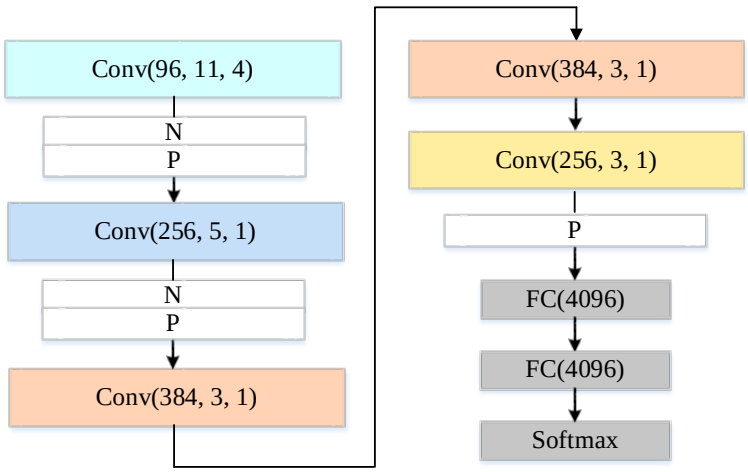

Figure17. Architecture of AlexNet used for intention prediction

5.4. Evaluation Results

For the single-participant data set, evaluation results of different groups of data augmentation methods are listed in Table 2. One can notice that, without data augmentation (i.e., the prediction model is trained by the original images only), the prediction model can achieve a prediction accuracy of 50% only. By analyzing the evaluation results, we can conclude that:

- The prediction model trained by augmented data obtained by Cropping (0, 10%) achieves the best prediction accuracy of 75% (with an improvement of 25% compared with original RGB images without data augmentation).

- The first category methods do not improve the intention prediction performance. The prediction model trained by augmented data obtained by using the first category methods even has a worse performance than that trained by the original images only (i.e., without augmentation).

- By choosing proper augmentation parameters, the second category methods can effectively improve the intention prediction performance.

Table 2. Prediction accuracy on the single-participant data set

Data Augmentation Groups	**Prediction Accuracy**	**Method Category**
Original RGB images	50%	N/A
1. Random contrast normalization	47.2%	First
2. Random contrast normalization and Gaussian noise	41.7%	First
3. Random contrast normalization, Gaussian noise and Gaussian blur	38.9%	First
4. Random contrast normalization, Gaussian noise, Gaussian blur and Cropping (0, 10%)	55.6%	First and second
5. Cropping (0, 10%)	**75%**	**Second**
6. Cropping (0, 20%)	61.1%	Second
7. Translation (-10%, 10%)	66.7%	Second
8. Translation (-20%, 20%)	61.1%	Second
9. Translation (-30%, 30%)	50%	Second
10. Piecewise affine transformation and perspective transform	63.9%	Second

6. Evaluation of Random Cropping on a Multi-Participant Data Set

Experiments with the single-participant data set show that random cropping achieves the best performance in human intention prediction. Therefore, we are interested in the effectiveness and the generalization capability of random cropping for a multi-participant data set. The multi-participant data set contains more pitching trials from multiple participants with more motion variations.

6.1. Multi-Participant Data Set

There are 6 participants in the multi-participant data set. In total, 540 valid pitching trials from all 6 participants (90 valid pitching trials from each participant) are collected. Each valid pitching trial contains 90 frames of RGB images. 540 valid pitching trials of data are divided into two sets: 432 pitching trials for training the model of human intention prediction and 108 pitching trails for testing the model of human intention prediction. Two example valid pitching trials in the multi-participant data set are shown in Figure 18.

Figure 18. Two example valid pitching trials

6.2. Data Augmentation

Different from the original RGB images in the single-participant data set, the original RGB images of the multi-participant data set have a size of 640×360. Since the central area of the RGB images contains the most informative motion field of participants needed for intention prediction, we first crop the central area of the original images and then apply random cropping data augmentation on the cropped central area. More specifically, the central area of the original images is first cropped with a size of 240×240. Then random cropping with a size of 235×235 is applied to the cropped central area. Finally, the randomly cropped images with a size of 235×235 are resized to 224×224.

6.3. Prediction Model

ResNet18 [15] pre-trained with ImageNet [47] is used as the prediction model for the multi-participant data set. The architecture of ResNet18 is shown in Figure 19. The architecture of ResNet18 is shown in Figure 19. The channel number of the first convolutional layer is modified according to the numbers of the input channels. The last layer is changed to a Softmax classifier with 9 classes. Use the same notation as in section 5, For simplicity, convolutional with notation Block×2 mean that there are two identical blocks in the network architecture. The MaxPooling layer has a kernel size of 3 and a stride of 2. The AveragePooling layer has a kernel size of 7 and a stride of 7. To transfer the pre-trained weights of ImageNet, the cross-modality pre-training in [23] is used to transform the shape of the weights. Lastly the model is fine-tuned based on our own data set.

The prediction model is trained with the multi-participant data set using stochastic gradient descent. The learning rate is initialized to 2.5×10^{-4} and is divided by 10 when the accuracy stops improving.

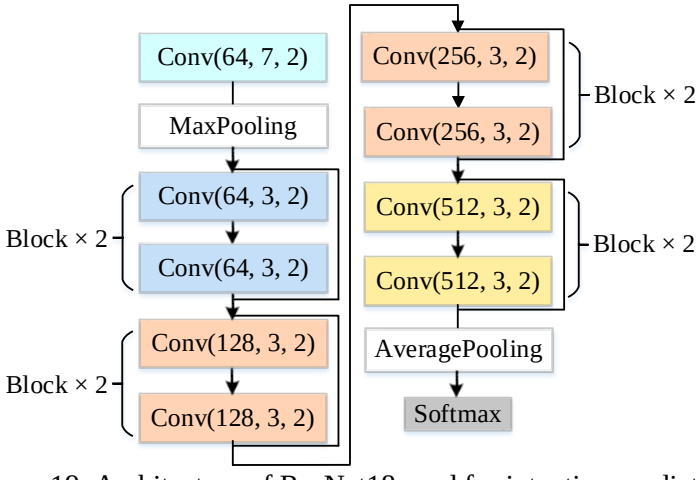

Figure 19. Architecture of ResNet18 used for intention prediction

6.4. Testing results on single-participant dataset Evaluation Results

The evaluation results of random cropping on the multi-participant data set are shown in Table 3. A prediction accuracy of 50% is achieved by using the original RGB images only. Random cropping improves the prediction accuracy to 57.4%. Such an improvement on prediction accuracy shows that the random cropping can be generalized to the multi-participant data set. Compared with the best prediction accuracy (75%) on the single-participant data set, random

cropping on the multi-participant data set achieves a lower prediction accuracy (57.4%). This could be caused by the variations of the spatial and temporal patterns from different participants in the multi-participant data set.

Table 3. Prediction accuracy on the multi-participant data set

Data Augmentation Methods	Prediction Accuracy
Original RGB images only	50%
Original RGB images + random cropping	57.4%

7. Evaluation of Random Cropping on Fusion Data of RGB Images and Optical Flow

Sections 5 and 6 evaluate data augmentation methods on RGB-image-based data. It is known that optical flow is crucial for learning spatio-temporal patterns [9][19][23][25][26] which is important for intention prediction tasks. Therefore, we are interested in intention prediction using fusion data of optical flow and RGB images. This section further explores the effectiveness of random cropping on fusion data of RGB images and optical flow.

7.1. Optical Flow Estimation

Optical flow describes the motion information of objects between two consecutive frames which is of great significance for video classification. FlowNet [48] was proposed to estimate optical flow using a supervised learning method. Specifically, a convolutional neural network was trained to estimate the optical flow from RGB images. What's more, FlowNet 2.0 [49] reduced the estimation error of FlowNet by more than 50%. In this study, FlowNet 2.0 is used for optical flow estimation. Figure 20 shows the optical flow of an example pitching trial.

Figure 20. Optical flow of an example pitching trial

7.2. Data Concatenation Fusion

For both single- and multi-participant data sets, the following two data concatenation fusion methods are considered to investigate the effectiveness of random cropping on the fusion data of RGB images and optical flow, as depicted in Figure 21.

- Fusion method 1: Concatenate three channels of RGB images with two channels of optical flow. Fusion method 1 aims to investigate the effect of random cropping on fusion data of RGB images and optical flow.

- Fusion method 2: Concatenate one channel of RGB images with two channels of optical flow. Fusion method 2 aims to explore the effect of random cropping on fusion data of RGB images and optical flow with increased temporal information (i.e., the proportion of optical flow in the fusion data is increased).

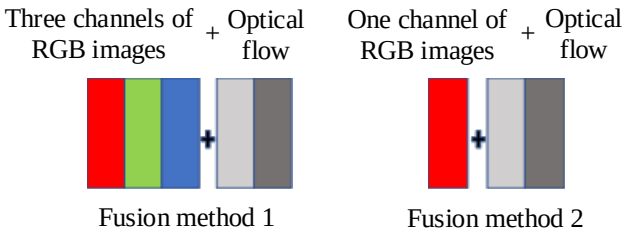

Figure 21. Data concatenation fusion methods

7.3. Data Augmentation

Random cropping is applied on the fusion data of RGB images and optical flow for both single- and multi-participant data sets. Specifically, for the single-participant data set, we apply Cropping (0, 10%). For the multi-participant data set, the same random cropping data augmentation method used in section 6 is applied. The original images are first cropped in the central area with a size of 240×240. Then random cropping with a size of 235×235 is applied to the cropped central area. Finally, the randomly cropped images with a size of 235×235 are resized to 224×224.

7.4. Prediction Model

ResNet18 [15] pre-trained with ImageNet [47] is used as the prediction model in this section. The cross-modality method in [23] is used to transform the pre-trained weights shape. The pre-trained weights are fine-tuned based on our own data sets.

The prediction model is trained on the fusion data of RGB images and optical flow using stochastic gradient descent. The learning rates for fusion methods 1 and 2 are initially set to 5×10^{-4} and 1×10^{-3}, respectively. The learning rates are divided by 10 when the accuracy stops improving.

7.5. Evaluation Results

After training the prediction models with the fusion data of RGB images and optical flow on both single- and multi-participant data sets, the prediction accuracies with fusion methods 1 and 2 are listed in Table 4. It can be noted that:

- For all comparison in Table 4, the random cropping data augmentation method significantly improves the prediction accuracy. It indicates that the random cropping data augmentation method is also effective on the fusion data of RGB images and optical flow. Using random cropping, fusion method 2 achieves the best prediction accuracy of 63.9% on the multi-participant data set.

- Besides the effectiveness of random cropping on fusion data of RGB images and optical flow, it is also observed that the prediction performance benefits from the increased proportion of optical flow in the fusion data. Without random cropping, fusion method 2 outperforms fusion method 1 by 2.8% on the single-participant data set. With random cropping, both fusion methods 1 and 2 achieve the same prediction accuracy of 75%. For the multi-participant data set, prediction accuracy of fusion method 2 is slightly lower than that of fusion method 1 when no random cropping is applied. However, with random cropping, fusion method 2 outperforms fusion method 1 by about 6.5%.

Table 4. Prediction accuracy on fusion data

Data Set	Methods	Prediction Accuracy
Single-participant data set	Fusion method 1	63.9%
	Fusion method 1 + random cropping	75%
	Fusion method 2	66.7%
	Fusion method 2 + random cropping	75%
Multi-participant data set	Fusion method 1	53%
	Fusion method 1 + random cropping	57.4%
	Fusion method 2	51.8%
	Fusion method 2 + random cropping	**63.9%**

8. CONCLUSION

This study investigates the effectiveness of various data augmentation methods in human intention prediction when limited training data is available. Evaluations of data augmentation methods are conducted on both single- and multi-participant data sets. Experiment results show that: 1) Data augmentation methods that either crop or deform images can improve the prediction performance; 2) The random cropping data augmentation method can be generalized to the multi-participant data set (improved prediction accuracy from 50% to 57.4%); and 3) Random cropping data augmentation method with fusion data of RGB images and optical flow can further improve the prediction accuracy from 57.4% to 63.9% on the multi-participant data set. The experiment results demonstrate that the random cropping data augmentation method is effective for the single-participant data set, the multi-participant data set, and the fusion data of RGB images and optical flow and can help improve the performance of deep-learning-based human intention prediction when only limited training data is available. In the future, we will investigate how to take advantage of generative adversarial networks [50] for data augmentation in human intention prediction.

REFERENCES

[1] Z. Wang, A. Boularias, K. Mülling, B. Schölkopf, and J. Peters, "Anticipatory action selection for human–robot table tennis," *Artif. Intell.*, vol. 247, pp. 399–414, 2017.

[2] H. S. Koppula, A. Jain, and A. Saxena, "Anticipatory planning for human-robot teams," in *Experimental Robotics*, 2016, pp. 453–470.

[3] E. C. Townsend, E. A. Mielke, D. Wingate, and M. D. Killpack, "Estimating Human Intent for Physical Human-Robot Co-Manipulation," *arXiv Prepr. arXiv1705.10851*, 2017.

[4] E. A. Kirchner, M. Tabie, and A. Seeland, "Multimodal movement prediction-towards an individual assistance of patients," *PLoS One*, vol. 9, no. 1, p. e85060, 2014.

[5] J.-Y. Kwak, B. C. Ko, and J.-Y. Nam, "Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime," *Infrared Phys. Technol.*, vol. 81, pp. 41–51, 2017.

[6] I.-H. Kim, J.-H. Bong, J. Park, and S. Park, "Prediction of driver's intention of lane change by augmenting sensor information using machine learning techniques," *Sensors*, vol. 17, no. 6, p. 1350, 2017.

[7] S. S. Phule and S. D. Sawant, "Abnormal activities detection for security purpose unattainded bag and crowding detection by using image processing," in *Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on*, 2017, pp. 1069–1073.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015, pp. 4489–4497.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[10] A. Zunino, J. Cavazza, A. Koul, A. Cavallo, C. Becchio, and V. Murino, "Intention from Motion," *arXiv Prepr. arXiv1605.09526*, 2016.

[11] S. Li, L. Zhang, and X. Diao, "Deep-Learning-Based Human Intention Prediction Using RGB Images and Optical Flow," *J. Intell. Robot. Syst.*, pp. 1–13, 2019.

[12] L. Zhang, S. Li, H. Xiong, X. Diao, and O. Ma, "An application of convolutional neural networks on human intention prediction," *Int. J. Artif. Intell. Appl.*, vol. 10, no. 5, pp. 1–11, 2019.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitionision (CVPR)*, 2016, pp. 770–778.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[18] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017, pp. 4724–4733.

[19] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1933–1941.

[20] J. Wang and L. Perez, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning."

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] G. Rogez and C. Schmid, "Mocap-guided data augmentation for 3d pose estimation in the wild," in *Advances in neural information processing systems*, 2016, pp. 3108–3116.

[23] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9912 LNCS, pp. 20–36.

[24] S. Li, L. Zhang, and X. Diao, "Improving Human Intention Prediction Using Data Augmentation," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 559–564.

[25] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[26] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A key volume mining deep framework for action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016, pp. 1991–1999.

[27] M. Daoudi, Y. Coello, P. Desrosiers, and L. Ott, "A New Computational Approach to Identify Human Social intention in Action," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018.

[28] Y. Xu, Y. Zhang, H. Wang, and X. Liu, "Underwater image classification using deep convolutional neural networks and data augmentation," in *2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2017, pp. 1–5.

[29] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[30] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, 2016.
[31] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017.
[32] J. Schlüter and T. Grill, "Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks.," in *ISMIR*, 2015, pp. 121–126.
[33] L. Taylor and G. Nitschke, "Improving deep learning using generic data augmentation," *arXiv Prepr. arXiv1708.06020*, 2017.
[34] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv Prepr. arXiv1708.04896*, 2017.
[35] S. Hauberg, O. Freifeld, A. B. L. Larsen, J. Fisher, and L. Hansen, "Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation," in *Artificial Intelligence and Statistics*, 2016, pp. 342–350.
[36] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.
[37] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Adaptive data augmentation for image classification," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3688–3692.
[38] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?," in *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, 2016, pp. 1–6.
[39] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 9, pp. 1469–1477, 2015.
[40] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V Le, "Autoaugment: Learning augmentation policies from data," *arXiv Prepr. arXiv1805.09501*, 2018.
[41] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv Prepr. arXiv1711.04340*, 2017.
[42] A. Jung, "imgaug." [Online]. Available: https://github.com/aleju/imgaug.
[43] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*. Cengage learning, 2011.
[44] "OpenCV,Image Filtering." [Online]. Available: https://docs.opencv.org/3.1.0/d4/d86/group__imgproc__filter.html#gaabe8c836e97159a9193fb0b11ac52cf1.
[45] S. Van der Walt *et al.*, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 2014.
[46] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732.
[47] A. Berg, J. Deng, and L. Fei-Fei, "Large scale visual recognition challenge 2010." 2010.
[48] A. Dosovitskiy *et al.*, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.
[49] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, vol. 2, p. 6.
[50] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.