Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 187 (2021) 140–145

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Identification, Information and Knowledge in the internet of Things, 2020

Research on gesture image recognition method based on transfer learning

Fei Wang[a], Ronglin Hu[a,*], Ying Jin[a]

[a]*Huaiyin Institute of Techology, Huai'an, China*

Abstract

To solve the problem of low gesture image recognition rate, we propose a transfer learning based image recognition method called Mobilenet-RF. We combine the two models of MobileNet convolutional network with Random Forest to further improve image recognition accuracy. This method firstly transfers the model architecture and weight files of MobileNet to gesture images, trains the model and extracts image features, and then classifies the features extracted by convolutional network through the Random Forest model, and finally obtains the classification results. The test results on the Sign Language Digital dataset, Sign Language Gesture Image dataset and Fingers dataset showed that the recognition rate was significantly improved compared with Random Forest, Logistic Regression, Nearest Neighbor, XGBoost, VGG, Inception and MobileNet.

Keywords: Gesture image identification, transfer learning, convolution neural network, random forest;

1. Main text

Gesture recognition can not only help deaf and dumb people communicate with the outside world, as an interactive means of wearable devices, automobile electronics and smartphones, the abundant information contained in it can be applied to a wider range of fields [1, 2].

* Ronglin Hu. Tel.: +86-0517-83591046.
E-mail address: huronglin@hyit.edu.cn.

10.1016/j.procs.2021.04.044

The goal of gesture recognition is to perceive specific human gestures, for which domestic and foreign research experts attempted to apply neural network to gesture recognition [3-6]. In order to progressively improve the accuracy of recognition, the number of layers of the neural network is continuously increasing. As the number of layers deepens, the storage problems of the model and the efficiency of the prediction come along. For this reason, this paper firstly transfers the lightweight MobileNet neural network model to the gesture image for feature extraction, and uses its more efficient network computing method to reduce network parameters and reduce the loss of network performance. Secondly, Random Forest is used as the classification model to effectively improve the recognition and accuracy of the algorithm. Experimental results show that this algorithm can achieve good recognition effect.

2. Related work

In this section, the related work for gesture image recognition technique is described. A brief introduction to the major transfer learning techniques will also be provided.

2.1. Gesture recognition

The initial gesture recognition mainly used wired machines and devices in direct contact with the hand, such as data gloves [7], accelerometers [8] and multi-touch screens [9], to collect data. These sensors can detect the orientation and angle of fingers, joints and arms, etc. This method has the advantages of fast reaction speed, high recognition accuracy and good stability. However, the equipment used in this method is costly in practice and seriously impacts the flexibility of the hands. Compared with contact sensor devices, visual gesture recognition can more directly convey hand movements, which has gradually become the mainstream way to study gesture recognition.

Gesture recognition in computer vision can be divided into two main tasks: static gesture recognition and dynamic gesture recognition. In the research of static gesture recognition, Zhang et al. [10] first used the parameters of the Hidden Markov Model (HMM) to train, and then recognized static gesture categories by layers. However, topological structure of HMM is general, which makes the calculation too large and the calculation speed too slow. The detailed description of the temporal and spatial changes of gesture signals is more suitable for dynamic gesture recognition. Josiane et al. [11] applied the Extreme Learning Machine (ELM) method to the feature classification of gesture images and obtained satisfactory results. ELM is a supervised learning algorithm proposed for the feedforward neural network of a single hidden layer. The shallow network layers cannot extract the image features better. In this paper, we use the deep learning model MobileNet network to extract more comprehensive and independent features of static gesture images and use them for classification. The research on static gesture recognition will promote the development of continuous gesture recognition.

2.2. Transfer learning

For the recognition of static gestures, neural network has been rapidly popularized in this field due to its characteristics of anti-interference, self-learning, easy control and high efficiency [12]. Although the accuracy of gesture recognition is constantly improved by neural network, the complexity of its structure is difficult for most beginners to build successfully. Therefore, in view of the limitations of neural network, it is necessary to break through the limitation of gesture recognition by building a network architecture and make full use of the advantages of neural network to extract features autonomously. This paper utilizes deep transfer learning [13] to deal with this challenge. ImageNet image dataset in a given source domain $D_s = \{X_s, F_s(X)\}$ and classification recognition task T_s, target domain gesture image dataset $D_t = \{X_t, F_t(X)\}$ and gesture recognition task T_t, by using classification T_s ImageNet image data sets and the source domain $Ds = \{X_s, F_s(X)\}$ acquired knowledge to help learn gestures in the target domain image dataset D_t prediction function $F_t(.)$ used to process gesture recognition task T_t, obviously $D_s \neq D_t$.

3. Methods

The Mobilenet-RF algorithm transfers the MobileNet network, that is, it extracts gesture image features by using the MobileNet network, while the classification part utilizes the Random Forest (RF) model. This section will briefly introduce its theoretical basis.

3.1. The architecture of Mobilenet-RF

The overall architecture of Mobilenet-RF is shown in figure 1. The overall structure is composed of two main parts. The feature extraction part adopts the model of MobileNet and the classification part adopts the model of random forest. Feature extraction and training of gesture image dataset $G_1 = \{img_1, img_2, ..., img_d, ... img_{len(G1)}\}$ by MobileNet, where img_d represents the single gesture image information set after d-th preprocessing, $len(G_1)$ denotes the total length of gesture dataset, and d is restricted to $1 \leq d \leq len(G_1)$. In the course of MobileNet network training, gesture images with original dimension (224×224×3) is firstly processed by a standard convolution, and then feature extraction is carried out through stacking depthwise convolution and pointwise convolution. Part of the depthwise convolution adopts the stride of 2 for down sampling. The Batch Normalization (BN) and ReLU activation functions are added for each depthwise convolution and pointwise convolution. BN is used to deal with the slow convergence speed of neural networks or untrainable situations such as gradient explosions. Compared with other activation functions, ReLU has great computing advantages, which can make the network design more in-depth. When depthwise convolution and pointwise convolution are calculated separately, the entire MobileNet network has 28 layers.

The algorithm of Mobilenet-RF is modified on the infrastructure of MobileNet. Only the first 28 layers of the network are used to extract gesture image features, and these features are directly input into the Random Forest model for classification.

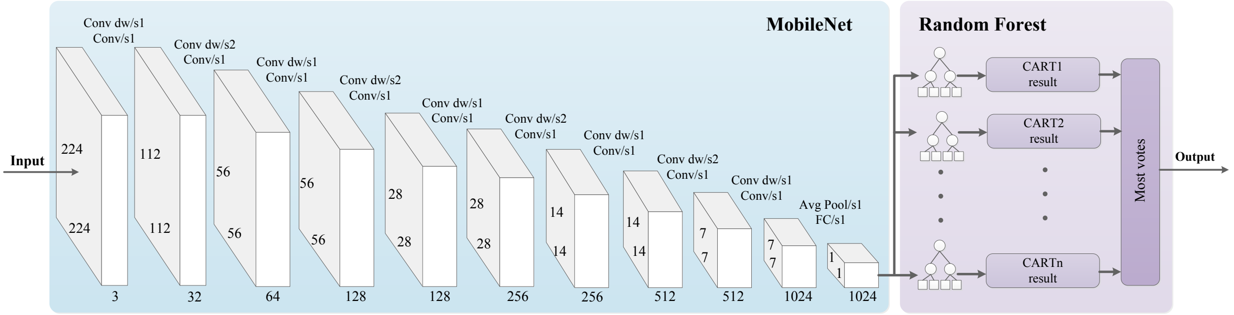

Fig.1. The body architecture of MobileNet-RF.

3.2. Gesture feature extraction

MobileNet is used to extract gesture features. Because MobileNet uses depthwise separable convolution filters to construct a lightweight neural network [14]. As illustrated in figure 2, MobileNet decomposes a standard convolution into a depthwise convolution and a pointwise convolution (1×1 convolution kernel). The formula of the convolution process is as follows:

$$F(m_i) = f_{relu}^{1\times1}(Conv(f_{relu}^{3\times3}(Dwise(m_i)))) \tag{1}$$

Where, m_i represents the feature map of the gesture image after the i depthwise separable convolution operation. The function *Dwise*(.) represents depthwise convolution, and the function *Conv*(.) represents pointwise convolution. Both use the *relu* activation function. The size of convolution kernel for pointwise convolution is 1 and that for depthwise convolution is 3. The feature map of gesture image $F(m_i)$ is obtained by depthwise separable convolution operation.

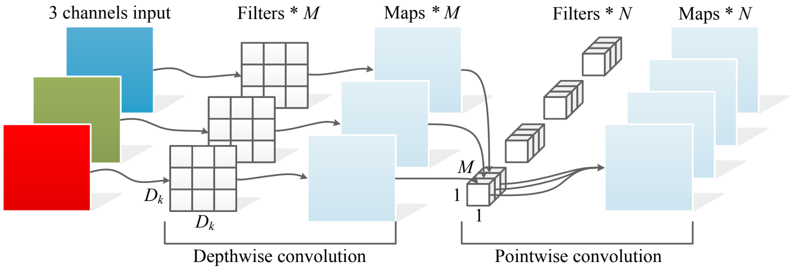

Fig.2. Depthwise separable convolution.

3.3. Classification of gesture features

RF was used to classify gesture features. Given a gesture image feature training dataset $D= \{(x_1, y_1), (x_2, y_2) ..., (x_n, y_n)\}$, and the number of sample subsets is T. In order to obtain the final gesture image strong classifier $f(x_t)$, m gesture image feature sample points are randomly extracted from the original sample dataset D and put back, and a sampling set d_t is obtained. The limit of t is $1\leq t \leq T$. Then the t CART decision tree is trained by using the sampling set d_t. In the training process, the segmentation rule for each node is to randomly select k features from all gesture image features, and then select the optimal segmentation point s from these k features to divide the left and right subtrees. T decision trees, namely T weak learners, were obtained after repeated T operations. In the classification algorithm, the final category predicted was the category in which the sample point cast the most votes in the T weak learners. For gesture image classification task, the whole classification process can be expressed by formula:

$$f(x_t) = Max_v\{H_i(x_t)\}_{i=1}^{T} \tag{2}$$

Where, $H_i(x_t)$ represents the output of the decision tree for the sample x_t to be tested. The function Max_v represents majority voting.

4. Experiment and result analysis

In this section, we present the execution details and experimental results for the proposed model.

4.1. Gesture datasets

In order to prove that the proposed method can effectively improve the accuracy of image classification, we selected the Sign Language Digital Dataset, Sign Language Gesture Image Dataset and Fingers Dataset as the image dataset of the experiment.

Sign Language Digital Dataset: The dataset contains 10 digital sign language images, provided by students at Ayranc Anadolu High School in Ankara, Turkey. A total of 218 students participated in the dataset production, and each student provided 10 samples, with a range of digital gestures ranging from 0 to 9.

Sign Language Image Dataset: The dataset is made up of 37 different gestures, including the a-z letter gesture, the 0-9 number gesture, and a space representation gesture, which means how deaf people represent the space between two letters or two words when communicating. Each gesture has 1,500 images, a total of 55,500. The data set consists of two parts. The first part is composed of colour images of t hands with different gestures. The second part of the image is colour gesture image, after the threshold binary conversion image for training and testing.

Fingers Dataset: the dataset is composed of left hand and right hand images. The left hand images are generated by flipping the right hand images. The total number of images is 21600, and the background is processed by noise. The original purpose was to distinguish the left hand from the right hand. In order to achieve a better experimental effect, it was reclassified according to the number of fingers and divided into 6 categories, each containing 3,600 gesture images.

4.2. Image preprocessing

In this experiment, apparent features of the hand (skin colour, contour and texture, etc.) were used to improve the accuracy of gesture recognition. The gesture images were preprocessed by gaussian filtering, binarization processing, morphological operation and image adding operation. The gesture images in the gesture dataset are processed in batches to complete the gesture segmentation of the whole gesture dataset. Figure 3 is the whole gesture image pre-processing process and processing effect diagram.

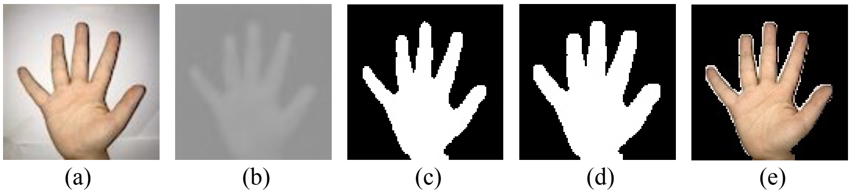

Fig.3. Gesture image preprocessing:(a)original gesture image;(b)gaussian filtering;(c)binarization processing;(d)morphological operation;(e)hand image extraction.

4.3. Implementation details

The MobileNet model is used to extract the features of gesture images. The migrated MobileNet network structure does not retain the top layer's fully connected layer network and pooling layer network, and the output of the last convolutional layer is 4D tensor. According to the size of the image in the dataset, the shape of the image in the input training is indicated. For example, (64,64,3) means that the input is a three-channel image with a length and a width of 64 pixels. The default number of filters is used to control the width of the network, and the depth multiplier of the deep convolution is 1. The weight of pre-training was loaded, and the parameters of training on ImageNet were transferred to the gesture image dataset. Dropout is 0.001.

It is proved by many experiments that the best result is obtained when the number of decision trees is 60 and the maximum depth of decision trees is 30. The performance of split mass was measured by the standard Gini index. The minimum number of samples needed to segment the internal nodes is 2. The minimum number of samples on leaf nodes is 1. When building a decision tree, sampling with a fallback is used, and out-of-pocket samples are not used to estimate generalization accuracy.

4.4. Experimental results

The experiments are carried out on Sign Language Digital Dataset, Sign Language Gestures Image Dataset and Fingers Dataset to test the accuracy of four common machine learning algorithms, RF, LR, KNN, and XGBoost, and three convolutional networks, VGG, Inception, and MobileNet, which are superior in classification performance for gesture recognition. The comparison of experimental results is shown in the table below.

Table 1. Experimental results of algorithm (1).

Method	Recognition rate /%								
	Sign Language Digital Dataset			Sign Language Gestures Image Dataset			Fingers Dataset		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
RF	45.77	48.67	46.75	29.08	30.33	29.10	91.97	91.89	91.89
LR	50.59	44.55	44.56	20.49	23.37	19.77	77.32	77.33	77.13
KNN	35.53	35.83	34.07	27.34	27.98	27.30	87.26	87.03	87.05
XGBoost	49.26	50.12	49.31	31.15	31.78	30.03	94.53	94.44	94.43
MobileNet-RF	**80.97**	**81.13**	**80.10**	**98.12**	**98.11**	**97.89**	**99.72**	**99.72**	**99.69**

Table 2. Experimental results of algorithm (2).

Method	Recognition rate /%		
	Sign Language Digital Dataset	Sign Language Gestures Image Dataset	Fingers Dataset
VGG	78.86	97.59	99.58
Inception	78.14	96.67	98.73
MobileNet	74.25	94.12	97.02
MobileNet-RF	**80.97**	**98.12**	**99.72**

As shown in table 1, the recognition accuracy of the convolutional neural network is superior to the classical and common machine learning algorithms (RF, LR, KNN and XGBoost) on the whole, and it is more suitable for gesture image recognition. Table 2 shows that our method achieves higher recognition accuracy than VGG, Inception, and MobileNet.

5. Conclusion

In this paper, we use deep transfer learning to transfer the structure of convolutional neural network from source domain to gesture image for feature extraction, avoid rebuilding the neural network architecture, and use the unique calculation method of MobileNet model to reduce the computational complexity of the model and achieve the effect of compression model. We use random forest model to classify the extracted features to overcome the problem of long time of network training. By comparing with Random Forest, Logistic Regression, K-Nearest Neighbor, XGBoost, VGG, Inception and MobileNet, it is determined that higher accuracy is achieved in gesture dataset.

References

[1] Li-Juan Sun, Li-Cai Zhang, Cai-Long Guo. (2008) "Technologies of hand gesture recognition based on vision." *Computer Technology and Development* **18**(**10**): 214-216.

[2] Jing-Guo Yi, Jiang-Hua Cheng, Xi-Shu Ku. (2016) "Review of gestures recognition based on vision." *Computer Science* **43**(**6A**): 103-108.

[3] Oyedotun O K, Khashman A. (2017) "Deep learning in vision-based static hand gesture recognition." *Neural Computing and Applications*, **28**: 3941-3951.

[4] Zhu G, Zhang L, Shen P, et al. (2017) "Multimodal gesture recognition using 3-D convolution and convolutional LSTM." *IEEE Access* **5**:4517-4524.

[5] Mohanty A, Rambhatla S S, Sahay R R. (2017) "Deep gesture: static hand gesture recognition using CNN." *Proceedings of International Conference on Computer Vision and Image Processing*, Springer, Singapore

[6] Cheng W, Sun Y, Li G. et al. (2019) "Jointly network: a network based on CNN and RBM for gesture recognition." *Neural Computing and Application* (**31**): 309.

[7] Lv Lei, Zhang Jinling, Zhu Yingjie, et al. (2015) "A static gesture recognition method based on data glove." *Journal of Computer Aided Design and Graphics* **27** (**12**): 2410-2418.

[8] Bourke A, O'Brien J, Lyons G. (2007) "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm." *Gait & Posture* **26**(**2**):194−199.

[9] Xue Jiao, Sun Peng, Deng Feng, et al. (2014) "Gesture remote control system based on touch screen." *Computer Engineering* **40** (**06**): 285-290.

[10] Zhang Lizhi, Zhang Yingrui, Niu Lianding, et al. (2019) "HMM static hand gesture recognition based on combination of shape features and wavelet texture features." *International Conference on Wireless and Satellite Systems(WiSATS)*, Springer, 187-197.

[11] Yu Haibin, Liu Jinguo, Liu Lianqing, et al. (2019) "Static hand gesture recognition for human robot interaction." *Intelligent Robotics and Applications*, ICIRA, Springer, 417-430.

[12] Wu Xia, Zhang Qi, Xu Yanxu. (2013) "A review of the development of gesture recognition." *Electronic Science and Technology* **26** (**6**): 171-174.

[13] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). "A survey on deep transfer learning." *In International conference on artificial neural networks*, 270-279. Springer, Cham.

[14] Howard A G, Zhu M, Chen B, et al. (2017) "MobileNets: efficient convolutional neural networks for mobile vision applications." *Computer Vision and Pattern Recognition*