Application of FMCW Radar for Dynamic Continuous Hand Gesture Recognition

Zhenyuan Zhang, Zengshan Tian, Mu Zhou, Yi Liu
zhangzhenyuangm@gmail.com, tianzs@cqupt.edu.cn, zhoumu@cqupt.edu.cn, liuyi21@cqupt.edu.cn
Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications
Chongqing 400065, China

Abstract

Recently, dynamic hand gesture recognition system is of great importance for human-computer interaction research. However, real-world dynamic hand gesture recognition system still exists some challenging problems, such as: 1) the system should be robustness to illumination conditions; 2) it is difficult to recognize diverse gestures performed by different people; 3) to avoid noticeable lag between its classification and performing a gesture, the system must detect and recognize continuous gestures utilizing unsegmented input streams. To solve these challenges, we present a novel system in this paper for recognizing dynamic continuous hand gestures based on Frequency Modulated Continuous Wave (FMCW) radar sensor. The radar system is not affected by noise, lighting or atmospheric conditions. We use a recurrent three-dimensional convolutional neural network to perform hand gesture classification. In addition, in order to improve recognition performance, Connectionist Temporal Classification (CTC) algorithm is utilized to predict class labels using unsegmented input streams. The experimental results demonstrate that the proposed system is able to achieve high recognition rate of 96%, which outperforms the state-of-the-art gesture recognition systems. What's more, the conclusion in this paper can be applied to real-time hand gesture recognition system design.

Index Terms

FMCW radar system, dynamic continuous hand gesture recognition, convolutional neural network, connectionist temporal classification

I. Introduction

Hand gesture recognition (HGR) is an interesting topic in the area of human-computer interaction[1-6], focusing on interpreting hand gestures based on various sensors and machine learning algorithms. Recently, HGR has been regarded as an effective interface for machines to understand human instructions. For example, with the popularisation of wearable devices, HGR system has been applied to micro portable electronic apparatus to replace small buttons and touch screens, which brings reliability and design flexibility improvements. In driver assistance

This work is supported in part by the National Natural Science Foundation of China (61771083, 61704015).

MOBIMEDIA 2018, June 21-22, Qingdao, People's Republic of China

DOI 10.4108/eai.21-6-2018.2276537

systems, HGR has realized non-contact interaction between drivers and vehicle navigation systems, improving driving security and operation convenience.

Compared to traditional camera-based HGR systems, radar-based systems have low computational resource cost and show promising performance in fine-gained gesture recognition under various illumination conditions. However, there is a number of open challenges existing in real-world radar-based systems for dynamic continuous HGR[7]. Firstly, to avoid noticeable lag between classification and performing a gesture, HGR systems must be capable of detecting and classifying gestures simultaneously using continuous streams of unprocessed radar data. Secondly, it is advantageous to address hand gesture segmentation and classification jointly, since they are highly interdependent. Thirdly, a dynamic hand gesture includes three temporally overlapping phases: preparation, nucleus, and retraction. In the preparation and retraction phases, there are similar actions cross different hand gestures, such as settling back a hand and stretching out. The nucleus mainly contains key hand movement features, such as different movement durations and trajectories. Compared to the other two phases, nucleus is the most discriminative. Therefore, for a dynamic HGR system, the key challenge is how to extract the nucleus phase from the above three phases. In addition, it is difficult for users to segment nucleus from the unsegmented input streams.

In this paper, we propose a Connectionist Temporal Classification (CTC) algorithm[8] based method for dynamic continuous HGR to address the above challenges. CTC algorithm makes gesture classification be performed based on the nucleus phase with no demand of explicit pre-segmentation. The system is based on a 24GHz FMCW radar platform. Diverse hand gestures are classified by the combination of three-dimension Convolutional Neural Network (3D-CNN), Long Short Term Memory (LSTM) network and CTC algorithm. The main contributions of this paper are summarized as follows.

- We present a radar-based dynamic HGR system. An end-to-end trained fusion network is proposed, which includes 3D-CNN and LSTM neural networks, and extracts motion features from not only short clip of input radar frames, but also long-term temporal information existing in hand movement sequences.
- CTC algorithm is utilized to recognize dynamic continuous hand gestures with no demand of hand gesture pre-segmentation.

II. System Description

A. Radar system description

To satisfy the the range and velocity resolutions demand for HGR, the chirp frequency bandwidth is $B = 4GHz$ and pulse duration is $T = 1ms$, achieving a range resolution $\Delta R = \frac{c}{2B} = 3.75\mathrm{cm}$. The transmitted signal of a FMCW radar can be modeled as:

$$S_T(t) = \exp\left(j2\pi\left(f_c t + 0.5Kt^2\right)\right) \tag{1}$$

where f_c denotes the carrier frequency, $K = B/T$. By assuming that a reflected signal with the distance R and moving radial velocity v_r, the received signal can be expressed as:

$$S_R(t) = \exp\left(j2\pi\left(f_c(t-\tau) + 0.5K(t-\tau)^2\right)\right) \tag{2}$$

where $\tau = 2\left(R + v_r t\right)/c$ is the round trip time-delay and c is the speed of light. In FMCW radar systems, the intermediate frequency signal $S_{IF}\left(t\right)$ of the low-pass filter output is then obtained:

$$S_{IF}\left(t\right) = \exp\left(j2\pi\left(f_c\tau + Kt\tau - 0.5K\tau^2\right)\right) \tag{3}$$

At last, 2D Fourier transform algorithm is applied to detect moving targets and estimate range information. From the the recorded spectrograms, as shown in Figure 1, it is easy to observe that different hand gestures have their own special trajectory features. From the spectrograms of sliding a hand from right to left, sliding a hand from left to right, pushing, and pulling, we can find the inverse hand movements have inverse trajectories.

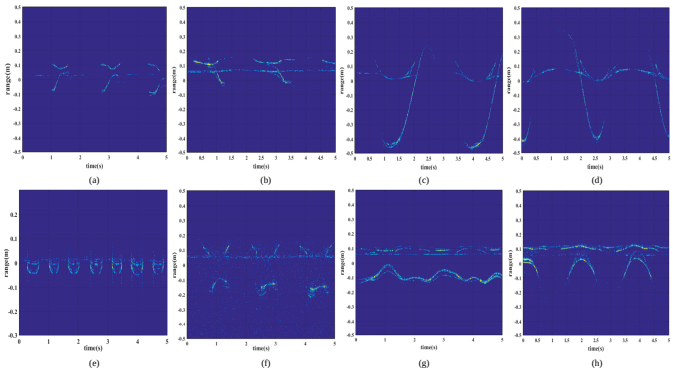

Fig. 1. Spectrograms of eight hand gestures: (a) sliding a hand from right to left, (b) sliding a hand from left to right, (c) pulling, (d) pushing, (e) knocking, (f) moving a hand up and down, (g) waving a hand, (h) patting.

(1)3D-CNN

In this paper, 3D-CNN is utilized to extract spatial-temporal feature from short consecutive radar spectrograms. The difference between 2D-CNN and 3D-CNN is that 3D-CNN uses a convolution kernel cube, rather than traditional 2D convolutional kernel, to perform the 3D-convolution operation. Formally, the value at position (x, y, z) on each feature map can be given as

$$\begin{aligned} F\left(x,y,z\right) &= \mathrm{ReLU}\left(b + X\left(x,y,z\right) \otimes H\left(x,y,z\right)\right) \\ &= \mathrm{ReLU}\left(b + \sum_{i=0}^{H_1-1}\sum_{j=0}^{H_2-1}\sum_{k=0}^{H_3-1} X\left(x+i, y+j, z+k\right) H\left(i,j,k\right)\right) \end{aligned} \tag{4}$$

where $X\left(x,y,z\right)$, $H\left(x,y,z\right)$, and $F\left(x,y,z\right)$ stand for the values at position (x,y,z) in previous radar spectrograms, convolution kernel cube, and feature map. b is the bias for this feature map. This system uses Restricted Linear Units ($\mathrm{Relu}\left(\right)$) function as the activation function. H_1, H_2 and H_3 is the length, width, and height of the convolution kernel cube. Similar to 2D-CNN, 3D-CNN relies on 3D pooling operation to down sample the feature map.

We start by formalizing the operation performed by the system. Firstly, we use a volume $\mathbf{C}^{(t)} \in \mathrm{R}^{k\times l\times m}$ $((m \geq 1))$ to represent a radar spectrogram clip having m sequential spectrograms with $k \times l$ spatial size at time t. Using a 3D-CNN $\mathrm{F}_{\mathrm{3D-CNN}}$, $\mathrm{f}^{(t)} = \mathrm{F}_{\mathrm{3D-CNN}}\left(\mathbf{C}^{(t)}\right)$, each spectrogram clip is transformed into a feature map $\mathbf{f}^{(t)}$.

(2) LSTM_CTC

The traditional LSTM cannot classify a sequence of diverse hand gestures in real-time, because it only outputs a classification result after several timesteps. However, LSTM_CTC, the combination of LSTM and CTC algorithm, can recognize a sequence of gestures in real-time. LSTM has several advantages over traditional Recurrent Neural Network (RNN). For example, LSTM is not limited to fixed length input, which is widely used in sequential data modeling. In addition, by introducing learned gating functions, LSTM units allow state to be propagated without modification, be updated, and be reseted. The core of LSTM units is memory cells. A memory cell includes input gate, forget gate, cell, and output gate. As stated above, 3D-CNN extracts the feature vector at time t, so an input feature sequence can be described as $\mathbf{f} = \left(\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \cdots \mathbf{f}^{(T)}\right)$. The input gate is a sigmoid unit, which takes activation from the current data input $\mathbf{f}^{(t)}$ and the hidden layer at the previous timestep. The forget gate is responsible for learning to flush the memory of the internal state. The cell in LSTM memory cell has a self-connected recurrent edge with fixed unit weight. In addition, in LSTM network, error can flow across several timesteps without vanishing or exploding, which is because that the edge spans adjacent timesteps with constant weight. At last, by multiplying the internal state by the output gate, the output of LSTM memory cell is obtained.

In this system, CTC algorithm is used as a cost function for sequence modeling based on unsegmented spectrogram streams. CTC is utilized to detect and recognize the nucleus of hand gestures, while assigning the *no gesture* class to represent the remaining sequence input clips and addressing the alignment of class labels to particular clips in the radar data.

As shown in Figure 1, let $\mathrm{L} = \{L_1, L_2, \cdots L_6\}$ denote the label dictionary of existing hand gestures. Firstly, the dictionary is extended with a *no gesture* class: $\mathbf{L}' = \mathbf{L} \cup \{no\ gesture\}$. Therefore, the output of softmax layer contains a class-conditional probability for this additional *no gesture* class. In this paper, an input radar sequence is described as $\gamma = \{\gamma_0, \gamma_1, \cdots \gamma_{p-1}\}$, where γ_i represents a training sample mini-batch in the form of unsegmented radar spectrogram data. Each radar spectrogram data includes T clips, making γ is composed of $N = T \times P$ clips. The network aims to compute the probability of observing a particular gesture (or *no gesture*) k at time t in an input sequence γ, $p\left(k, t|\gamma\right) = s_t^k\ \forall t \in [0, N)$, rather than averaging predictions across clips in a pre-segmented hand gesture.

Then, it is assumed that the output probabilities at each timestep is independent to those at other timesteps (or rather, conditionally independent given γ). π denotes a path which possibly maps of the input sequence γ into a sequence of labels $\mathbf{y}$. $p\left(\pi|\gamma\right) = \prod_t s_t^{\pi_t}$ is the probability of observing path π ,where π_t is the class label predicted at time t in path π.

Considering the fact that different paths could lead to the same label sequence, we define a many-to-one function ψ as $\mathbf{y} = \psi\left(\pi\right)$ to remove *no gesture* labels and condense repeated labels. For instance, $\psi\left(1 * *1 * * * *2 * *\right) = \psi\left(* * *11 * *12\right) = 12$, where $1, 2$ denote actual gesture labels and $*$ is no hand gesture label. By operator ψ, different paths lead to the same gesture sequence $\mathbf{y}$. Given an input sequence γ, the probability of observing a particular sequence $\mathbf{y}$ is the sum of the conditional probabilities of all paths π mapping to that sequence, $p\left(\mathbf{y}|\gamma\right) = \sum_{\pi \in \psi^{-1}(\mathbf{y})} p\left(\pi|\gamma\right)$, where $\psi^{-1}\left(\mathbf{y}\right) = \{\pi : \psi\left(\pi\right) = \mathbf{y}\}$. We can compute $p\left(\mathbf{y}|\psi\right)$ simply by dynamic programing. Then a new vector $\hat{\mathbf{y}}$ is defined by inserting a *no gesture* label before and after each gesture clip in $\mathbf{y}$. By assuming that $\mathbf{y}$ contains P labels, the length of $\hat{\mathbf{y}}$ is $\hat{P} = 2P + 1$.

For sequence $\mathbf{q}$ with the length r, $\mathbf{q}_{1:p}$ and $\mathbf{q}_{r-p:r}$ represent its first and last symbols respectively. Then for a label sequence $\mathbf{y}$, the forward variable $g_t(s)$ is defined to be the total probability of $\mathbf{y}_{1:s}$ at time $t, g_t(s) = \sum_{\substack{\pi \in N^T \\ \psi(\pi_{1:t})=\mathbf{y}_{1:s}}} \prod_{t'=1}^{t} s_{t'}^{\pi_{t'}}$.

It is allowed for hand gesture sequence to start with either *no gesture* or the first hand gesture in $\mathbf{y}(\mathbf{y}_1)$. Therefore, we initialize the forward variable by the probability of a path beginning with *no gesture* or the probability of a path beginning with the first actual hand gesture. In addition, a valid path is not allowed to begin with a later hand gesture. The rules for initialization is given as follows:

$$\begin{cases} b_1(1) = s_1^{no\ gesture} \\ b_1(2) = s_1^{\mathbf{y}_1} \\ b_1(s) = 0,\ \forall s > 2 \end{cases} \tag{5}$$

and the transition function is

$$b_t(s) = \begin{cases} \left(b_{t-1}(s) + b_{t-1}(s-1)\right) s_t^{\hat{\mathbf{y}}_s} \\ \quad if\ \hat{\mathbf{y}}_s = \ no\ gesture\ or\ \hat{\mathbf{y}}_{s-2} = \hat{\mathbf{y}}_s \\ \left(b_{t-1}(s) + b_{t-1}(s-1) + b_{t-1}(s-2)\right) s_t^{\hat{\mathbf{y}}_s} \\ \quad \text{otherwise} \end{cases} \tag{6}$$

At last, any valid path π must end with the last gesture $\hat{\mathbf{y}}_{P'-1}$ or with *no gesture* at time $N-1$. Therefore, $p(\mathbf{y}|\gamma)$ can be represented by $p(\mathbf{y}|\gamma) = b_{N-1}\left(\hat{P}-1\right) + b_{N-1}\left(\hat{P}\right)$. Therefore, we can defined the CTC loss to be $L_{CTC} = -\ln\left(p(\mathbf{y}|\gamma)\right)$.

Though CTC is used as a training loss function only, by inserting the extra *no gesture* label, it has the influence on the architecture of the network.

III. Experimental Results

A. Continuous Hand Gesture Recognition

To test dynamic continuous hand gesture recognition, radar data sequence is collected for 80 seconds. The comparison of the recognition performance of "LSTM" and "LSTM_CTC" is shown in Figure 2. This figure also shows the network predictions and ground truth labels during continuous operation on the sequence. Different hand gestures is represented by various colors and line types. The nucleus phase of each gesture is described by the ground truth in the top row. The experimental result shows that "LSTM_CTC" achieves higher recognition performace in HGR: 96.25% and 90.3% in terms of "LSTM_CTC" and "LSTM", respectively. In addition, it is obvious to find that the two networks behave differently when the same hand gesture is performed sequentially by observing that instances of the same gesture conducted at 22-35s and 54-70s. The LSTM_CTC network generates an individual peak for each repetition, whereas LSTM merges them into a single activation.

Conclusion

In this paper, we propose a novel radar-based HGR system, which is capable of recognizing diverse dynamic hand gestures in real-time. We employ the combination of 3D-CNN and LSTM networks to extract spatial-temporal

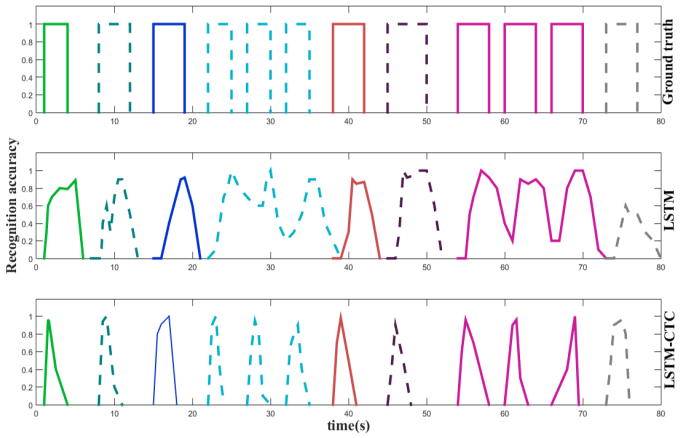

Fig. 2. Comparison of the recognition performance of LSTM and LSTM_CTC.

features from radar sequence data. In addition, CTC algorithm is responsible for classifying gestures with zero or negative lag using unsegmented input data in our system. The experimental results demonstrate that the proposed system is able to achieve higher recognition rate compared to the state-of-art systems. In the future, we will continue to study how to extract the spatial-temporal features of dynamic gestures using single network to reduce the computational resource.

REFERENCES

[1] Ge L, Liang H, Yuan J, et al. "3d convolutional neural networks for efficient and robust hand pose estimation from single depth images", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1991–2000, 2017.

[2] Wan, Chengde, et al. "Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[3] Sun, Yuliang, et al. "Gesture Classification with Handcrafted Micro-Doppler Features using a FMCW Radar", 2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM). IEEE, pp. 1–4, 2018.

[4] Smith, Karly A., et al. "Gesture Recognition Using mm-Wave Sensor for Human-Car Interface", IEEE Sensors Letters, vol 2, pp.1-4, 2018.

[5] Li, Gang, et al. "Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition." IEEE Transactions on Aerospace and Electronic Systems, vol. 54, pp. 655–665, 2018.

[6] Takamine, Asamichi, Y. Iwashita, and R. Kurazume. "First-person activity recognition with C3D features from optical flow images." *IEEE/SICE International Symposium on System Integration IEEE*, pp. 619-622, 2015

[7] A. Kendon. "Current issues in the study of gesture", in The biological foundations of gestures: motor and semiotic aspects, Lawrence Erlbaum Associates, pp. 23–47, 1986.

[8] Graves, Alex, and F. Gomez. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", in ACM International Conference on Machine Learning, pp. 369–376, 2006