Multi-round Dialogue Intention Recognition Method for a Chatbot Baed on Deep Learning

Junmei Li(✉)

School of Computer Engineering, Jingchu University of Technology, Jingmen 448000, China
chenweiliang7895@163.com

Abstract. With the continuous development of human-computer dialogue system, more and more dialogue robot products come into people's lives. However, when human beings use short sentences and omit words, and in the process of identification often face problems such as more text noise, sparse characteristics, polysemy, backward and backward dialogue information. In order to solve the above problem, a deep learning based chatbot multi-round dialogue intention recognition method, according to the fit of deep learning algorithm and chatbot multi-round dialogue intention recognition model, by transforming the problem into a mathematical model, and obtain the final dialogue intention through the calculation of the model. First, the chat dialogue text was preprocessed, and the BERT model was established based on the processing results, the BERT model fused the deep learning model in the B E R T model, established a joint model, and data vectorized the short text of the human-computer dialogue. Finally, the multi-round dialogue intention identification similarity is calculated through the robot, realizing the dialogue intention recognition, and experiments show that the highest accuracy of the recognition method can reach 0.9912, the highest recall rate can reach 0.9914, and the highest f price is 0.9914, which can prove the superiority of the design method.

Keywords: Deep learning · Chatbots · Multiple rounds of dialogue · Dialogue intent · Intent recognition

1 Introduction

With the advent of the AI era, more and more intelligent products have been widely used in everyday life, such as the emotional escort robot personal mobile assistant Siri, voice assistant Google Now, and Cortana. XiaoBing, an intelligent chatbot launched by Microsoft Asia Research Institute, and a Xiaodu robot launched by Baidu. These intelligent dialogue systems cannot only communicate with normal information with users, but also bring a lot of convenience to users' lives [1]. The dialogue system consists mainly of five parts: Automatic Speech Rec-ognition (ASR), Spoken Language Understanding (SLU), dialogue management (Dialog Management, DM), dialogue generation (Dialogue Generation, DG), and Textto Speech (TTS), as shown in Fig. 1. In order to let the machine better understand the expression of users, and then feedback the correct

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 561–572, 2022.
https://doi.org/10.1007/978-3-031-18123-8_44

information, spoken understanding plays an extremely important role. IntentDetection (ID), as a submodule of oral understanding, is also the key to the human-computer dialogue system. Traditional spoken understanding is mainly divided into two subtasks: intent recognition and semantic slot filling. Since the early research is limited by application scenarios, data and computing power, most oral comprehension is limited to some fields. However, with the innovation of technology and the emergence of multi domain dialogue system, today's oral understanding is often divided into three tasks: domain recognition, intention recognition and semantic slot filling [2].

In dialogue systems, intention recognition is crucial. The intention is the user's intention, namely what the user wants to do. Intent is sometimes referred to as "dialogue behavior" (Dialog Act) I', the behavior where the information status or context shared in the conversation changes and is constantly updated. The intention is generally named after the "verb + noun", such as weather inquiry, hotel booking, etc. intent recognition, also known as intent classification, is classified into a previously defined category of intent based on the areas and intent involved in the user's utterance.

With the widespread use of the human-computer dialogue system, users may have different intentions in different occasions, so they will involve multiple fields in the human-computer dialogue system, including the task-type vertical field and small chat [3]. The purpose of the task text is clear and easy to retrieve, such as flight tickets, weather, hotels, etc. The chat intention text generally has the characteristics of unclear theme, semantic width and short statements, paying attention to the communication with humans in the open domain. In the dialogue system only clear the user's topic field, to correctly analyze the specific needs of the user's intention, otherwise will cause later intention error identification when the user input a query, first need to clarify the user input text topic field is "train" "flight", because the intention category more granular than the topic field, so need to determine the user's specific semantic information is refund or query time, and semantic slot filling is also helpful to the user intention judgment. Therefore, in the intention recognition module of the human-computer dialogue system, it is first necessary to identify the user topic field, and then the specific intention needs of the users should be defined, and finally express the form of the semantic framework [4]. However, previous methods often face some problems in the process of recognition, such as text noise, sparse features, polysemy, backward dialogue information and so on. In order to solve the above problems, this paper proposes a multi round dialogue intention recognition method of chat robot based on deep learning. According to the deep learning algorithm and the multi round dialogue intention recognition model of chat robot, the problem is transformed into a mathematical model, and the final dialogue intention is obtained through the calculation of the model, in order to provide some help to improve the accuracy of multi round dialogue intention recognition of chat robot.

2 Design a Multi-round Dialogue Intent Recognition Model Based on Deep Learning

Intent recognition has become a new research hotspot in academia and industry, and to correctly understand user intentions in human-computer dialogue systems, intention identification can be solved as a short text classification problem, where a category is

automatically determined for text in certain specific categories according to pre-defined topic categories. Intent to identify the user intention from the user short text of the dialogue process through the short text classification. The process of intention recognition is expressed as mathematically symbolic, as shown in Fig. 1, which can be viewed as a mapping relationship $f : U \rightarrow I$, where $I = 1, 2....$, Where the set U is the type of intent preset by the natural language statement i entered by the user, and I is the list of intentions resolved from the natural statement i.

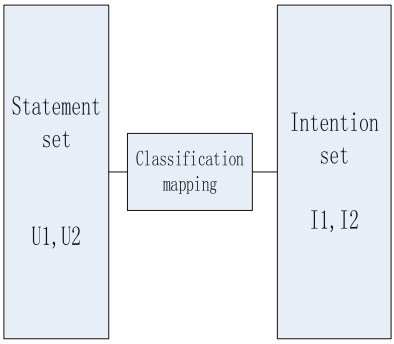

Fig. 1. Intent to identify the mathematical model

The process of intention identification can be roughly described as follows: models with parameters conduct parameter learning and model optimization on the training set with intention markers, and then use the trained optimal model to identify each data on the test set of hidden intention labels, calculate the identification results and compare them with the real label.

2.1 Chat Conversation Text Preprocessing

Text preprocessing is at the beginning of the entire intent recognition process, and many anomaly data or values in text data will directly or indirectly affect the results of downstream tasks, so preprocessing has great significance and necessity for conventional anomaly data or values [5–7], With important implications for the results of intent recognition. Because the short text format is not standard and the number of words is small, if it is not reasonably standardized processing, it will affect the results of intention recognition, so in the process of intention recognition, the preprocessing of short text text is a step that cannot be ignored. Data preprocessing can avoid many non logical exceptions before the algorithm starts, such as denoising data, processing outliers, missing values and so on. Text preprocessing mainly includes noise removal, parti, and removal of stop words.

Usually, the short text data crawled from the Internet platform contains not only the text and characters that identify the semantic, but often has a large number of additional structures with low connection to text labels and content, such as hyperlinks, emojis, description symbols, HTML markers, XML tags, pictures, etc., etc., which have no significance to build intention recognition, they usually have no clear semantic information,

but only express the text information more intuitively. However, in the intention recognition task, these data do not only contribute to the semantic expression of short text data, but will increase the processing time and operation scale of short text data, and will also adversely affect the operation process such as word segmentation and text vectorization, ultimately resulting in reducing the accuracy and reliability of text processing methods. Therefore, these abnormal information and noise data should be cleaned and sorted out before the short text data is formally processed.

In our daily life, English is the most widely used language, involving major countries in various regions, and Chinese is the most widely used language, but the two most representative languages differ greatly in terms of processing. It is relatively simple to divide English, only to divide according to the space characters between words. There is no clear separation between Chinese words, and the words in Chinese are separated by semantic and context. Therefore, compared with English segmentation segmentation, it is difficult to handle Chinese segmentation, and more rules and restrictions should be considered. Word segmentation is the basic step of Chinese text preprocessing and the premise of text representation. Under certain syntactic semantic rules, word segmentation is the process of dividing the original continuous expression into single words or words. In this process, many components with little correlation to the semantic expression of the original text will be ignored, and only the key and core words or words are retained. The effect of participle will directly affect the effect of words, semantic representation, we can choose the appropriate participle tool according to the different use scenarios and requirements.

There are three text segmentation methods in natural language processing: one is based on grammar and rules, one is based on dictionaries, and the other is based on statistics. This paper will adopt the method of stuttering segmentation, which is based on statistical segmentation methods. The rationale of a statistical-based partitioning method is to determine whether a string constitutes a word based on the statistical frequency of its occurrence in the corpus. Words are a combination of words, and the more times adjacent words appear simultaneously, the more likely it is to form a word [8]. Therefore, the frequency or probability of co-occurrence adjacent to words can better reflect their credibility to becoming words.

Stop words refer to be frequent in text, but from the perspective of semantic understanding and expression, it has little influence on tasks such as text representation or intention recognition. Stop words mainly include public stop words and professional stop words. Public stop words usually have commonly used prepositions, crowns, aids, pronouns, conjunctions, etc. Based on the empirical summary of the numerous research work, Stop words are roughly divided into two categories: one refers to some words that are very widely used, Words like “I”, “just” that appear in almost every text, However, its association with the intent labels is very low, Not only did there have any positive impact on the identification task, Instead, because of the excessive number of appearances, consumption of time, Also reduces the efficiency of identification; The other category refers to the high frequency of both appearing and being used in the text, But these words are not substantive or decisive in semantic expression, Its role in the text is only to ensure the standardization and integrity of the text in the grammatical structure, This type of words usually contain tone aids, adverbs, prepositions, conjunctions, etc., They have no

practical meaning of themselves, Nor decisive determine the intention and emotional tendency of the text, Only put into the complete sentence can show a certain auxiliary effect, Such as the common “of”, “in”, “and”, “then”, etc.

2.2 Establish the BERT Model

This chapter presents a joint model BERT _ word2vec based on BERT and word2vec to quantify the short text of human-computer dialogue, whose model structure diagram is shown in Fig. 3-1. The vectorization representation method first trains the word vector in the word2vec model, and calculates each word vector to the sentence hierarchy vector pre-trained by BERT, then transforms the resulting similarity value into weights assigned to the corresponding word vector, and finally combines the weighted word vectors into the sentence vector and the sentence vector of BERT [9] (Fig. 2):

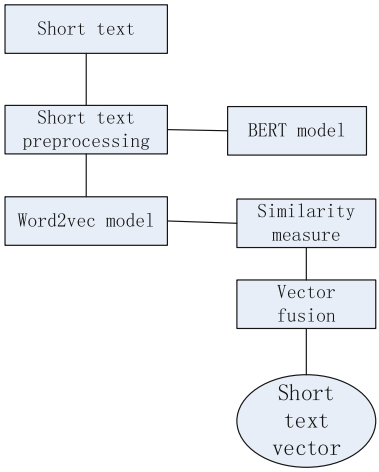

Fig. 2. Structural diagram of the BERT_word2vec model

The BERT (Bidirectional Encoder Representations from Transformers) model, is essentially implemented on the basis of a two-way Transformer encoder, where E, E, …, Ex is the input vector of the model, and bidirectional Transformer coding yields a vectorized representation of the text T. Transformer is a self-attention (Self-attention)-based seq2seq model, a Encoder-Deocder-structured neural network whose input and output are a sequence [10]. In the model, Encoder transforms the input sequence of variable length into a fixed-length vector expression, and then decodes this fixed-length vector into a variable-length target signal sequence by Decoder, and Figs. 3-3 are the structural diagram of the traditional model, where C is the state vector between Encoder and Deocder. In fact, the basic Encoder-Decoder structure is implemented based on RNN, and its core module is composed of RNN units, but with the increase of sequence length, RNN itself has some unavoidable problems, such as unable to parallel, slow operation, etc. At the same time, because the state vector size of the connecting Encoder

and Decoder is fixed, the Decoder is unable to directly follow more details of the input information [11]. To improve on the above deficiencies, Transformer uses self-attention to replace the RNN. Since the Encoder part of the Transformer is mainly used in the BERT model, the Encoder structure in the Transformer model is highlighted below, as shown in Fig. 3:

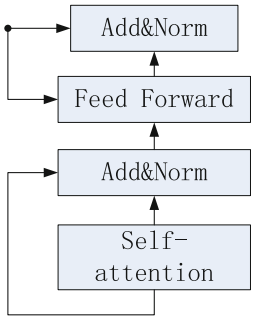

Fig. 3. The Encoder structure

As can be seen from Fig. 3, the input of Encoder consists of a vectorial representation of the short text and the location information of each word in the text, and then through the Self-attention layer gives Encoder the information to view the word before and after the word when encoding each word. Another Add & Norm layer, Add adds the input and output of the self-attention layer, Norm normalizes the added output, so that the output of the self-attention layer has a fixed mean 0 and standard deviation 1, and the normalized vector representation will be processed through a fully connected feedforward neural network (Feed Forward). Similarly, the Feed Forward layer also contains the Add & Norm layer [12]. As mentioned above, a new list of word vectors will be output. The core module in part encoder is Self-attention, whose main idea is to calculate the mutual relationship between each word in the short text and all the words in the short text, and then adjust the weight of each word to obtain a new expression of each word [13]. This new expression not only contains the semantic meaning of the word itself, but also contains the relationship between other words and the word, so it is a more global expression compared with the traditional word vector. The Self-attention procedure is calculated as follows:

Suppose the input short text is expressed as:

$$X = (x_1, x_2, ..., x_n)^T \tag{1}$$

where, x_i is the i word in the short text, and now it is expressed as a_i by one hot vector. There are n words in total, and the vector matrix is obtained:

$$A = (a_1, a_2, ...a_n)^T \tag{2}$$

Then multiply the vector matrix by three different weight matrices W_g, W_k and W_v to obtain query, key and value matrices Q, K and V. The importance s of each word is

calculated as follows:

$$s = Q \bullet K \tag{3}$$

The smoothed result of s is multiplied by the softmax function to obtain the value of attention. Each line represents the attention vector of the corresponding word in the input short text. The vector has been integrated with the information of other position words, which is a new vector representation. The specific calculation formula is shown in 4:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where: d_k is the penalty factor to ensure that the inner product of Q and K is not too large. As can be seen from the calculation formula above, the whole computational process is a series of matrix multiplication that can be parallelized and superior to RNN. In practice, Transformer uses Multi-headSelf-attention, or multiple Self-attention in parallel, to enhance the attention power of the model.

2.3 Robot Multiple Rounds of Dialogue Intent Recognition Similarity Calculation

The similarity metric adopted is cosine similarity, which measures the cosine value of the two vector clip angles in vector space as the size of the difference between two individuals, which pays more attention to the difference between the two vectors in direction than distance or length, so the method is more suitable for the vectized data in this chapter [14]. Assuming that the output vector of the Bert model is expressed as A, the word vector trained by word2vec is B_i, $i = 1, 2, ..., n$, which represents the word vector of the i-th word in the short text, and n is the total number of words in the short text, the calculation formula of the similarity S between the i-th word in the short text and the short text is shown in 5, and the weight w corresponding to the th word can be obtained from the similarity S, as shown in formula 6:

$$S_i = \frac{A \bullet Bi}{\|A\|\|B\|} \tag{5}$$

$$w_i = \frac{S_i}{\sum_{i=1}^{n} S_i} \tag{6}$$

Then multiply each word vector by its corresponding weight, splice it into a vector, and then add it to the short text preprocessed by Bert to obtain the final vectorized representation of the short text. The vector contains both semantic features at the sentence level and highlights words closer to the sentence meaning, which not only compensates for the disadvantage that word2vec cannot reflect the word polysemy, but the presence of the sentence vector generated by BERT also complements the semantic information lost during word vector splicing. Finally, the scaling point product attention results of h times were spliced from left to right, and the attention matrix X obtained from a second linear transformation was used as the result of multi-head attention. The specific calculation formula is as follows:

$$head_i = Attention(Qw_i, Kw_i, Vw_i) \tag{7}$$

$$MultiHead(Q, K, V) = (head_1 + head_2 + head_3 + ... + head_i) \tag{8}$$

In Eqs. (7) and (8): Q, K and V represent query matrix, key matrix and value matrix respectively, with equal values, which are vectorized output E; $\sqrt{d_k}$ refers to the square root of the bond vector dimension, which plays a regulatory role and controls that the inner product of Q and K will not be too large; W is the parameter of linear transformation, and W is different every time Q, K and V are linearly transformed; h represents the number of heads and i represents the attention head.

3 Test Experiments

In order to verify the effectiveness of the deep learning-based multi-round dialogue intent recognition method designed in this paper, compare the method designed in this paper with the traditional dialogue intent identification method (literature [11] method), highlighting the advantages of the method designed in this paper.

3.1 Dataset

The experimental data are selected from the corpus of the booked restaurant, but because the corpus only contains the data of a single round of dialogue, considering the specific requirements of the experiment in this chapter, the dialogue of the data set is required for this chapter. Multiple rounds of dialogue in the dataset are shown in Table 1:

Table 1. Multiple-round dialogue data

Conversation object	Dialog box text	Intention
p1	Is there any restaurant near here?	find_restaurant
b1	Please provide your current location	The dialogue robot identified the key word “nearby”
p2	I am on XX Road, XX No	ask_ location
b2	Recommend XX restaurants to you	The dialogue robot identified the questioner

A list of intent labels involved in the conversation and data examples are shown in Table 2:

The multi-round dialogue dataset annotated in this paper contains 10,432 dialogue data and 100 sets of dialogue, ranging from 1 to 8 rounds, with 11 intention labels appearing. It can be seen from the data distribution that the intention categories of data are unbalanced, especially the number of data of greet, thanks and deny, which has a great impact on the accuracy and is easy to bring large errors to the experiment. Therefore, so

Table 2. Schematic labels and data examples

Order number	Intent name	The intent label	Example data
1	Query the restaurant	find restaurant	Is there any restaurant near here?
2	Number of information	info people	About 5 people had their meals
3	Location information	ask_location	I am on XX Road, XX No
4	Telephone information	phone_number	My cell phone is 99999999999999
5	Price information	info _price	How is your meal price here
6	Confirm information	confirm	Yes, that is what I am looking for
7	Denial information	deny	This restaurant does not meet my requirements
8	send one's respects to	greet	shalom
9	thanks	thanks	Thank you very much
10	good-bye	goodbye	a form of greeting by women
11	other	others	Recommend XX restaurants to you

these three intentions are not included when calculating the comprehensive performance of each model in the subsequent experiments. During the experiment, the datasets were divided into training, test, and validation set in a ratio of 6:2:2. To ensure the contrast of the experiment, the datasets in the experiment used the same division method, and the data used the same short text vector representation method.

3.2 Experimental Result

Due to the data imbalance, the amount of data in the intention categories greet, thanks and d e n y does not affect the evaluation of the overall performance of the model. The identification results of these three intentions are not included in the calculation of the model indicators: Precision (accuracy) and RecallK (recall) and F-measure are used to evaluate the classifier performance, as shown in (9), (10), (11):

$$\Pr ecision = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{10}$$

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{11}$$

where: if sample A belongs to a category B and is recognized as such by the model, it is recorded as TP; If sample A does not belong to category B but is identified as category B

by the model, it is recorded as *FN*; If sample *A* belongs to a category *B* but is identified by the model as a category other than category *B*, it is recorded as *TN*; If sample *A* does not belong to category *B* and is identified by the model as a category other than category *B*, it is recorded as *TN*.

For the multi-classification problem of this task, the evaluation index of each category is first calculated, and then the macro average value (Macro-average) is used as the final evaluation index.

In the algorithm text extension method, the minimum support is set to 0.01 and the minimum confidence is 0.5; In the LDA subject extension method, the number of topics is set to 10. After obtaining the subject distribution of the chat text, the word with the maximum probability under the subject is added to the text. During the feature selection process, this paper starts from 100 dimensions and calculates the performance effect of the recognition model with 100 as a step length, finally retaining the most relevant 3,000 words as the final feature. Different recognition models were also tested for different parameters.

To reduce the contingency of the experimental results, this chapter uses a 5-fold cross-validation method to obtain the final identification results. Where "original" refers to the word feature + feature selection method; "+ apriori" means the content extension on the original base; "+ lda" means the theme extension on the original base; "+ apriori, 1da" means merging original, apriori and 1da3 methods.

The final results are shown in Table 3:

Table 3. Comparison results of different methods

Numerical classification	Naive Bayesian multi-round dialogue intent recognition methods			Random Forest multi-round dialogue intent identification method			The multi-round dialogue intent identification method designed in this paper		
	Precision	Recall	F price	Precision	Recall	F price	Precision	Recall	F price
Original	0.7445	0.7412	0.7541	0.7412	0.7544	0.7445	0.9135	0.9912	0.9178
+apriori	0.7845	0.7415	0.7153	0.7415		0.7845	0.9912	0.9145	0.9914
+lda	0.7746	0.7745	0.7544	0.7745		0.7746	0.9541	0.9914	0.9145
+apriori, lda	0.7256	0.7523	0.7826	0.7523		0.7256	0.9416	0.9514	0.9914

From the experimental results, it can be seen that in three different intention recognition models, the content and theme are expanded through Apriori and LDA, and the effect is improved. However, compared with the other two methods, it can be seen that the multi round dialogue intention recognition method of the design method in this paper has better effect, with the highest accuracy of 0.9912, the highest recall rate of 0.9914 and the highest f price of 0.9914, which can prove the superiority of the design method. The reason for this result is that this design method first calculates the correlation on the corpus, then finds out the co-occurrence relationship of each word in the chat text, and completes the limited relationship with the text. Because there are some connections between common words, after completing the words, it can enrich the current text content and supplement the synonyms not mentioned in the original text. After the description information is added to the supplementary text, the recognition effect is

improved compared with the original text. Moreover, because deep learning needs to set word support, which is the minimum frequency, a priori mining common associations in the corpus and display content supplement is the premise, that is, the user's expression of high-frequency words also limits the user's need for relative expression norms, and try not to appear random rather than standard low-frequency words. However, in the actual process, users may have more irregular expressions, and the expression patterns are very diverse, resulting in a large number of repetitions of diversified spoken words, which can not be found according to the association mining algorithm. Based on this situation, this chapter uses LDA to mine the topic information of the text, which can bypass the word level supplement. The overall semantics of the text is very helpful to express non-standard text.

4 Conclusion

In order to improve the accuracy of short text intention recognition in man-machine dialogue system, this paper proposes a short text vectorization method and two intention recognition methods. Firstly, the text of chat dialogue is preprocessed, and the Bert model is established according to the processing results. The Bert model integrates the deep learning model in the Bert model, establishes a joint model, and quantifies the short text of man-machine dialogue. Finally, the similarity of multi round dialogue intention recognition is calculated by the robot, and the dialogue intention recognition is realized. The experimental results show that the recognition accuracy of the design method can reach 0.9912, the recall rate can reach 0.9914, and the f price is 0.9914, which can prove the superiority of the design method.

References

1. Al-Mayyahi, A., Aldair, D., Chatwin, C.R.: Control of a 3-RRR planar parallel robot using fractional order PID controller. Int. J. Autom. Comput. **17**(6), 822–836 (2020)
2. Rutschi, C., Dibbern, J.: Towards a framework of implementing software robots: transforming human-executed routines into machines. Data Base Adv. Inf. Syst. **51**(1), 104–128 (2020)
3. Travagnin, S.: From online Buddha halls to robot-monks: new developments in the long-term interaction between Buddhism, media, and technology in contemporary China. Rev. Relig. Chin. Soc. **7**(1), 120–148 (2020)
4. Perugia, G., Paetzel-Prüsmann, M., Alanenp, M., Castellano, G.: I can see it in your eyes: gaze as an implicit cue of uncanniness and task performance in repeated interactions with robots. Front. Robot. AI **8**, 1–18 (2021)
5. Zhang, L., Yang, Y., Zhou, J., Chen, C.C., He, L.: Retrieval-polished response generation for chatbot. IEEE Access **8**, 123882–123890 (2020)
6. Ren, F., Xue, S.: Intention detection based on Siamese neural network with triplet loss. IEEE Access **8**, 82242–82254 (2020)
7. Saha, T., Gupta, D., Saha, S., Bhattacharyya, P.: Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. Cogn. Comput. **13**(3), 277–289 (2020)
8. Gupta, D., Bansal, P., Kavita: Emotion recognition: differences between spontaneous dialogue and active dialogue. J. Shanghai Jiaotong Univ. (Sci.) **16**(9), 633–644 (2021)

9. Li, J., Guo, H., Chen, S., Yang, D., Zhao, L.: A novel semantic inference model with a hierarchical act labels embedded for dialogue act recognition. IEEE Access **7**, 167401–167408 (2019)
10. Yang, W., Wan, B., Qu, X.: A forward collision warning system using driving intention recognition of the front vehicle and V2V communication. IEEE Access **8**, 11268–11278 (2020)
11. Chen, Y., Li, C.: Simulation of target tactical intention recognition based on knowledge map. Comput. Simul. **36**(8), 5 (2019)
12. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)
13. Liu, S., et al.: Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multimedia **23**, 2188–2198 (2021)
14. Liu, S., Wang, S., Liu, X., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)