·Review·

Data fusion methods in multimodal human computer dialog

Ming-Hao YANG[1,2], Jian-Hua TAO[1,2,3*]

1. *National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Sciences, Beijing* 100190, *China*

2. *The Center for Excellence in Brain Science and Intelligent Technology of Chinese Academy of Sciences, Beijing* 100190, *China*

3. *University of Chinese Academy of Sciences, Beijing* 100049, *China*

*** Corresponding author,** jhtao@nlpr.ia.ac.cn

Received: 15 October 2018 **Accepted:** 20 December 2018

Supported by the National Natural Science Foundation of China (61873269, 61425017, 61332017, 61831022); the National Key Research & Development Plan of China (2017YFB1002804).

Citation: Ming-Hao YANG, Jian-Hua TAO. Data fusion methods in multimodal human computer dialog. Virtual Reality & Intelligent Hardware, 2019, 1(1): 21—38
DOI: 10.3724/SP.J.2096-5796.2018.0010

Abstract In multimodal human computer dialog, non-verbal channels, such as facial expression, posture, gesture, etc, combined with spoken information, are also important in the procedure of dialogue. Nowadays, in spite of high performance of users' single channel behavior computing, it is still great challenge to understand users' intention accurately from their multimodal behaviors. One reason for this challenge is that we still need to improve multimodal information fusion in theories, methodologies and practical systems. This paper presents a review of data fusion methods in multimodal human computer dialog. We first introduce the cognitive assumption of single channel processing, and then discuss its implementation methods in human computer dialog; for the task of multi-modal information fusion, serval computing models are presented after we introduce the principle description of multiple data fusion. Finally, some practical examples of multimodal information fusion methods are introduced and the possible and important breakthroughs of the data fusion methods in future multimodal human-computer interaction applications are discussed.

Keywords Intention understanding; Multimodal human computer dialog

1 Introduction

Since the advent of computers, humans have dreamed of one day having a natural conversation with computers. In the beginning of the 21st century, the information service represented by call center represents the arrival of the era of highly intensive voice and text information. It has provided information support for the social service industry and economic benefits. It was predicate by Negroponte Nicholas in his well-known book "Being Digital"[1] that "in 1000 years, people and machines will talk more than people". With no more than 20 years, with the development of speech recognition, speech synthesis and natural languages understanding technologies, natural conversation between human and computer has made great progress. After decades of development, the man-machine dialogue system has developed from the early telephone voice system, such as language learning, ticket and hotel booking, etc,[2–4] to the current inaccurate questions and answers, such as speech assistant: Apple Siri and Google Duplex.

In traditional human computer dialog techniques, human-computer interaction focused on speech and language information processing. While in people' daily face-to-face communication, their information is often transmitted from multiple channels, including facial expressions, emotional voice, postures and gestures. For example, in the process of human interaction, when one's voice or tone is not enough to reflect the specific meaning of expression, human tend to express their intentions using not only speech but also facial expressions, body movements or gesture[5]. In these cases, a simple expression, such as gestures' fast and slow movement, amplitude changes in smile, contain rich interactive information in human computer dialog. The multi-modal human-computer dialogue is superior to the traditional single mode in the efficiency and integrity of information expression.

Compared with traditional single-channel interaction, multimodal human-computer interaction has a wider and potential application in mobile interaction and natural interaction[6,7], such as smart home[8], smart human-computer dialogue[9,10], somatosensory interaction[11–13], education[14], etc. In recent years, artificial intelligence technology contribute to largely improvement for single channel behavior perception, including speech recognition[15,16], face recognition[17,18], emotional understanding[19–23], gesture comprehension[24–27], posture analysis[12,28,29], handwriting comprehension[30–32], eye trace[33–35], touch[36–39], etc. Nowadays, the computer is able to understand the users well for one channel behavior.

Traditional single channel human computer interaction methods, such as mouse and keyboard, or graphic interface, obtain users' input signals accurately from capture device. However, under the condition of multimodal human dialog and natural interaction, the machine needs to judge users' intention from multiple channels, for example in the field of home service robots, when a user points to an apple on the table and says, "please give me the apple", the robot needs to accurately understand the user's target "apple" from the speech context and the "position" demonstrated by the gesture. Understanding of user intentions from users' multi-modal information fusion plays key role in the nature interaction in human-computer dialog. Therefore, Multimodal data fusion in multimodal human-computer dialog is a very dynamic and extensive research field. The goal of this paper is not to present a complete summary of this field, but to analyze the characteristics of multimodal information processing in multi-modal human computer dialog, and then to introduce how multi-modal information fusion can increase the interaction feeling of human computer dialog.

The remainder of this paper is organized as follows: we first introduce the concept of multi-modal fusion and the assumption of multimodal information fusion in cognitive science in section 2; in section 3, we will introduce some multi-modal information fusion methods in practical applications; discussion and conclusions are presented in section 4 and section 5.

2 Multi-modal information fusion

Various types of channels exist in interaction. Although the data acquisition and storage methods of these channels are quite different, they share some common characteristics in information processing from the perspective of cognitive science. In this section, we will introduce the calculation of single channel information, and then discuss the processing of multimodal information.

2.1 Single channel information processing

2.1.1 Information processing in cognitive assumption

Cognitive psychology supposes that human processing information follows three hypotheses: Channel-Filter (CF), Limited-Capacity (LC) and Active-Processing (AP). The first two hypotheses are mainly related to single-channel information processing, the latter is more related to multimodal information processing[40,41].

Single CF hypothesis refers to that human beings first process the information of each channel separately, and then form the cognitive characteristics of each channel[40,41]. According to the external situation of the channel, they select the representation of the learning object[42]. LC hypothesis refers to the limited capacity of data processed one time, which means that the channel information capacity that learners can maintain has certain limitations at a time[43,44]. Experiments found that human memory span is about seven units, such as subjects in the order of one word per second. After reading a string of numbers, the subjects were able to repeat about seven digits accurately[45]. Some scholars even believe that there is less capacity[46]. This shows that the information that humans can focus on is very limited without using cognitive functions such as association, reasoning, and memory[47].

Recently, more psychological experiments and cognitive discoveries have further confirmed the hypothesis of CF and LC[42,48]. Cognitive science regards that human memory structure is consisted of three levels: sensory memory, working memory and long-term memory[42]. The hypothesis of Limited Capacity mostly occurs in the sensory level, while information process last including these three levels, from sensory memory to working memory and until to long-term memory[49].

2.1.2 Single channel information computing

In early interactive systems, users' interactive intentions are converted to position, click and touch pressure in screen from different devices, such as keyboard and mouse. Similarly, in recent human computer dialog system, different kind of channel information is converted into behavior recognition, and then the system gives feedback according to the behavior recognition results. The conversion of these single channel numbers into intention understanding can be simplified by formula (1).

$$y^t = f(x^t, x^{t-1}, \cdots, x^{t-l}) \tag{1}$$

In formula (1), x^t is the input signal at time t for one channel x, such as the positions of strokes, gestures, emotions and speech. l is the length of information units to process, y^t means the recognition result at time t. In this way, the processing of single channel information becomes a problem of function fitting or pattern classification.

The assumption of Channel Filter and Limited Capacity are related to the two variables x and l in formula (1). x corresponds to the signals or the feature presentation. Before the rise of deep learning, artificially designed features played key role in calculation of intention understanding. Now, the features obtained by deep network structures automatically do not perform less than the features by designed domain experts[50,51]. Generally speaking, it is an important task to find better feature representation in channel filter processing.

l in formula (1) presents the length of features in memory, which is related to the LC assumption. In the traditional machine learning process, l is viewed the time window of the input signal. The values of l partly affect the accuracy of recognition results. For example, in the task of facial emotion recognition, the emotion recognition results are relatively better when the values of l is set to 11 in cascaded support vector machines[52]. The similar situations are also discussed in the field of speech recognition[16,53] and gesture recognition[27,54]. The LC assumption is also validated in the long-short memory recurrent neural network (LSTM). When the pooling characteristic are added in the middle layer of LSTM, the model achieved higher performances in emotional recognition[21], gesture recognition[24,25] and speech recognition[16,53].

2.2 Multimodal information processing

2.2.1 Cognitive assumption of multi-modal information fusion

With the increasing computing power of computer, the performance achieved by computer for single channel signal is close to or even outperform the users without domain knowledge. Just like single channel processing, the multi-modal information processing is a process of re identification after sample learning.

It was regarded that cognition processes fall into two categories: memory and understanding. Memory refers to the ability of a person to recall, recognize and learn materials or information presented in the past. Understanding is the ability to construct a mental representation of the learned content and skills. Understanding occurs after memory and can be applied in a new situation. This ability enables learners to apply what they have learned in a new situation, and the effect can be measured by a transfer test[55]. An important character of understanding is that learners will face new situations, and they need to transfer the knowledge learned to solve the problem. multi-modal information fusion is conducive to understanding through memory, and in turn continue to promote memory[56,57].

Cognitive enhancement is based on third assumptions in cognitive science learning process: AP hypothesis. Active processing refers to the strategy of encoding, organizing and integrating new information with learned knowledge to participate in cognitive processing. The target of AP is to establish a consistent psychological representation of different channel knowledge and experience[49].

2.2.2 Framework of multi-modal information fusion

Multimodal information fusion runs through sensory memory, working memory and long-term memory modules, and it correspond to three hypotheses of human cognition after multi-modal signal acquisition[55]. Human perceptual channels acquire information, including images, voice, touch and other signals, complete the feature representation and coding, selectively enter the working memory area. Because of the LC principle, the selected features are retained in the corresponding brain regions for a short period of time, and multimodal features are correlated and fused. For specific tasks, AP include attention, association and reasoning mechanisms, triggering long-term memory of the learned historical knowledge, synthesizing the multimodal fusion information in working memory and obtain the final judgment[58].

The cognitive process of multimodal learning indicates that the fusion of multimodal information occurs in working memory, and the fusion process triggers the formation of long-term memory knowledge. Although there is no definite conclusion about the storage of formation and knowledge in the brain, a task for cognitive science is to explore the enhancement effect of intention understanding from multimodal information fusion over single-channel information processing. At the same time, computer engineers and researchers tried to verify whether multimodal information fusion would help information understanding based on the single-channel knowledge, and if so, how much more accurate the latter will be than the single-channel information.

Based on the presentation of formula (1), formula (2) gives the description form of multimodal information fusion.

$$y^t = f\,[\oplus_{k=1}^{K}(x_k^t), \oplus_{i=1}^{k}(x_k^{t-1}), \cdots, \oplus_{i=1}^{k}(x_k^{t-1})] \tag{2}$$

In formula (2), $k(2 \leqslant k \leqslant K)$ is the number of channels. x_k^t is the input signals at time t for the k-th channel. The symbol $\oplus_{k=1}^{K}(x_k^t)$ presents the fusion of multimodal signals from x_1^t to x_k^t. y^t is the fusion results of all channels at time t. Being similar to the single channel processing, Eq. (2) can be simplified as a function fitting problem. However, because of the distinct difference of different various channels, it is difficult to write $\oplus_{k=1}^{K}(\mathrm{x}_k^t)$ in an uniform format.

A quantitative analysis of multimodal information fusion is introduced in [59]. Supposing that x_k^t follows Gaussian distribution, denoted as $\mathrm{x}_k^t \sim \mathrm{N}(\mathrm{u}_k, \delta_k)$, and the degree of confidence for each channel could be described as $w_k = \frac{1/\delta_k^2}{\sum_k 1/\delta_k^2}$. Then, the multi-modal signal fusion process could be expressed as a formula

(3). Formula (3) actually is a maximum likelihood estimation for multiple Gauss signals. With the presentation of formula (3), the calculation of multimodal information fusion is expediently converted to a patterns classification problem.

$$\tilde{S} = \oplus_{k=1}^{K}(s_k) \sim N\left(\sum\nolimits_{k=1}^{K} w_k u_k, \sum\nolimits_{k=1}^{K} w_k \delta_k\right) \tag{3}$$

Ernst and Banks[59] further discussed the information fusion through a perception experiment for depth information fusion from visual and tactile dual-channels perception by virtue of virtual reality environment. Statistical results show that the results of scene depth estimation experiment obey the description of formula (3) in the depth information of visual and tactile dual-channel perception scene. The fusion of multimodal signals can be expressed as the maximum likelihood estimation of multiple signals when the single channel signal obeys Gaussian distribution. And at the same time, the channel with higher confidence plays a more dominant role in fusion.

3 Fusion method of multimodal information

Comprehension of users' behaviors happens at or after the time of multimodal information fusion. The multimodal information fusion styles influence the calculation model. According to the data where and when fused happened, fusion could be happened at data (feature) level, model level and decision level respectively. According to the calculation method, the fusion can be divided into rule-based and statistic (machine learning) based fusion[10,60]. Considering the relativity of different channels, some literatures concluded their relations into the three categories: complementary, mutual exclusion and redundancy[61]. Among these fusion strategies, the statistical and machine learning methods are widely applied in users' behavior comprehension. Therefore in the following subsections, we focus introduce statistics and machine learning fusion methods.

3.1 Bayesian decision model

The characteristic of Bayesian decision-making is that it can adopt subjective probability estimation to some unknown states from incomplete information. It is able to modify the probability of occurrence, and finally make the optimal decision with expected probability[62]. When the joint distribution probability of multimodal signals is known, Bayesian decision-making is able to retrieve some missing signals from historical experience, and obtain the global optimal evaluation of the whole multi-modal signal fusion. Supposing the probability of joint distribution of different channel signals is $p_s(S)$=$p_s(\mathrm{s}_1, \mathrm{s}_2, \cdots, \mathrm{s}_D)$, where D means the number of channels, then the probability of the edge distribution of a channel observation signal, denoted as $p_d(\mathrm{d}_{obs})$, could be written as $p_d(\mathrm{d}_{obs}|\mathrm{S}) \times p_s(S)$, where $p_d(\mathrm{d}_{obs})$ is the observation value of the d-th channel. According to Bayesian theory, given the prior knowledge and the joint distribution probability of a given channel signals could be written as formula (4) and formula (5).

$$p_d(\mathrm{S|d}_{obs}) = p_d(\mathrm{d}_{obs}|\mathrm{S}) \times p_s(S) / p_d(\mathrm{d}_{obs}) \tag{4}$$

$$p_d(\mathrm{d}_{obs}) = \int_S p_d(\mathrm{d}_{obs}|\mathrm{S}) \times p_s(S) \times \mathrm{d}V_s \tag{5}$$

In (4) and (5), $p_d(\mathrm{d}_{obs})$ is the observation of the edge distribution, which corresponds to the actual observation of a channel signal. According to the initial values of $p_d(\mathrm{d}_{obs})$,$p_d(\mathrm{d}_{obs}|\mathrm{S})$ and $p_s(S)$, we could obtain the accurate values of $p_s(S)$ and $p_d(\mathrm{d}_{obs})$ iteratively using (4) and (5).

Because of the advantages in multi-channel information integration, Bayesian decision model can retrieve the optimal decision under partial observation conditions from incomplete information. It potentially improves the performance of users' behaviors comprehension from incomplete observation signals. This make Bayesian decision model well adopted in the fields of face tracking[63], user behavior perception[64], robot pose estimation and obstacle avoidance[65], emotional understanding[66] and multi sensor information alignment and observation data analysis[67].

3.2 Neural network model for multimodal information fusion

The traditional neural network model has good performance in nonlinear function fitting, and the neural networks with deep structure are widely used in speech recognition, man-machine dialogue, machine translation, semantic understanding, object recognition, gesture detection and tracking, human body detection and tracking. In the field of emotional recognition, the best result obtained by computer using depth-long-short-term memory neural network model (Long Short-Term Memory Neural Networks, LSTM) is about 10% different from that obtained by professionals[21,23]. In the field of speech recognition, for dialect accent speech recognition, depth recurrent neural networks (RNN) can achieve a word recognition accuracy of 95%, which is very close to human level[68]. In the field of image target recognition, the Very Large Scale Convolution Neural Network (CNN) has exceeded the level of ordinary human identification[69,70]. In the single channel depth neural network model technology, many researchers synthesize the above LSTM, CNN, RNN structures to build large-scale depth neural network model for multi-channel information fusion, trying to process multi-channel information indistinguishably in the fusion stage. Generally speaking, multi-modal information fusion is applied to deep neural networks. Like ordinary multi-channel information fusion, it takes place in three levels in structure, namely early data level fusion, medium-term model level fusion and late rule level fusion[60]. Figure 1(a), 1(b) and 1(c) correspond to the above three kinds of multi-channel information fusion abstract representation. We can see that data or feature-level fusion is performed by using the extracted features from each modality and concatenating these features into one large vector. While those of model-level and decision level fusion for different modality are performed at middle layers and top layers of whole fusion structures.

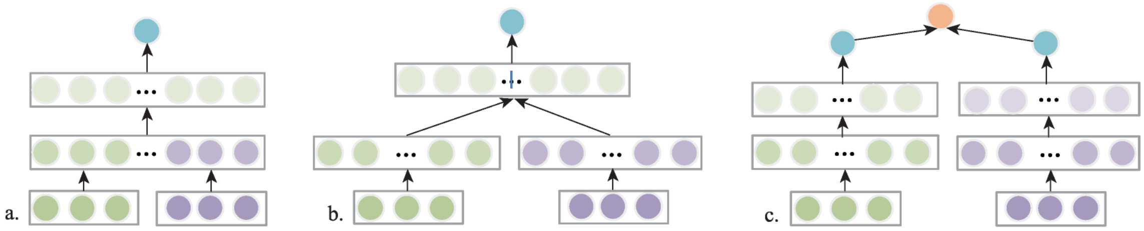

Figure 1 Structure of classical neural networks for information fusion. (a) Fusion at data or feature level; (b) Fusion at model level (c) Fusion at decision level.

Based on the above structure, a variety of multi-channel information fusion can further combine the above structure, build more complex large-scale structure, and realize multi-task learning[71 - 73]. Cross modal learning[74]. At the same time, in the case of joint training based on multi-modal data, this kind of structure can achieve good results even if one modal information is missing. It has achieved good results in multi-channel emotion recognition, semantic understanding, target learning and other fields. Nevertheless, such networks are relatively targeted to certain tasks. If the tasks are changed, users need to modify the network structure including network structures and parameters, which make the design of deep neural network structure a time-consuming and labor-intensive work. Therefore, researchers hope that a hybrid neural network structure can perform multiple tasks simultaneously to reduce its workload in structural design and training. In view of this, researchers began to focus on building a multi-channel joint feature sharing layer by using large data joint training, and then a deep multi-modal fusion structure which can process multi-tasks simultaneously in the recognition stage. For example, Google scholars try to suggest a unified in-depth learning model that adaptively addresses multiple different types of tasks in different domains and data modes without significant performance loss on specific tasks[75]. A detailed description of the three parts is given in Figure 1[75].

3.3 Graph model based information fusion

Graph model combines probability calculation with graph theory to provide a better tool for calculating uncertainties. Different from neural network models, the nodes and the connections between nodes make it advantageous to calculate the relationship between variables and adjacent variables. Graph models can be divided into undirected graph model and directed graph model according to the direction of the connection between nodes. With the help of multi-scale analysis, undirected graph is widely used in scene segmentation, video content analysis, text semantic understanding and so on. Segmentation, detection and tracking of human motion in video based on Markov field model[76], Undirected graph model joint multi document summarization extraction[77], Undirected graph estimation based on groupings is used to retrieve missing features in multi-channel information[78], sentiment classification of text, video and audio information based on undirected graph model[66], detection of brain pleasure activity distribution in multimodal brain regions based on Gauss graph model[79].

Compared with undirected graph model, the connection between nodes of directed graph model not only remembers the data flow direction, but also records the state jump probability in the learning process. Directed graph model can be used not only in uncertainty calculation, but also in decision-making reasoning for time series problems. For example, dynamic Bayesian model imitates human writing process[80]. Gesture and gesture understanding based on Markov decision process understanding[81], Optimal decision of multi-user behavior conflict based on directed graph model[61,82], Multimodal man-machine dialogue based on weighted finite state automata for decision-making strategy of modal conflict dialogue[61].

In addition to the above multi-channel information fusion calculation model, there are many other models also used for multi-channel information fusion, such as multi-level support vector machine, decision regression tree, random forest and other methods, because the length of this article, we will not describe them totally here.

4 Multimodal information fusion in human computer dialog

4.1 The general framework of multimodal human computer dialog

In human's daily life, various interaction channels, such as speech, gesture, body movement, facial and emotional expression, gaze, touch and so on, contribute to human's nature social communication. In spite of various interaction channels, speech and visual information are the still two main channels that influence and determine the process and quality of communication in dialogs. Like human dialog, human-computer dialog relies on the reception and analysis of multi-modal users' behaviors, as well as the understanding of the context and scene context of interaction. Therefore correspondingly, in the framework of multimodal human computer dialog, there contains information acquisition, information processing, information output modules. The information acquisition part usually refers to the computer through camera, microphone and other sensors to obtain human multi-modal behaviors, and further transmitted to the information processing part. According to the multi-modal information input by the system, the information processing module analyzes human interaction information, and produces dialogue content and feedback to users. Figure 2 presents a general framework for multimodal human computer dialog.

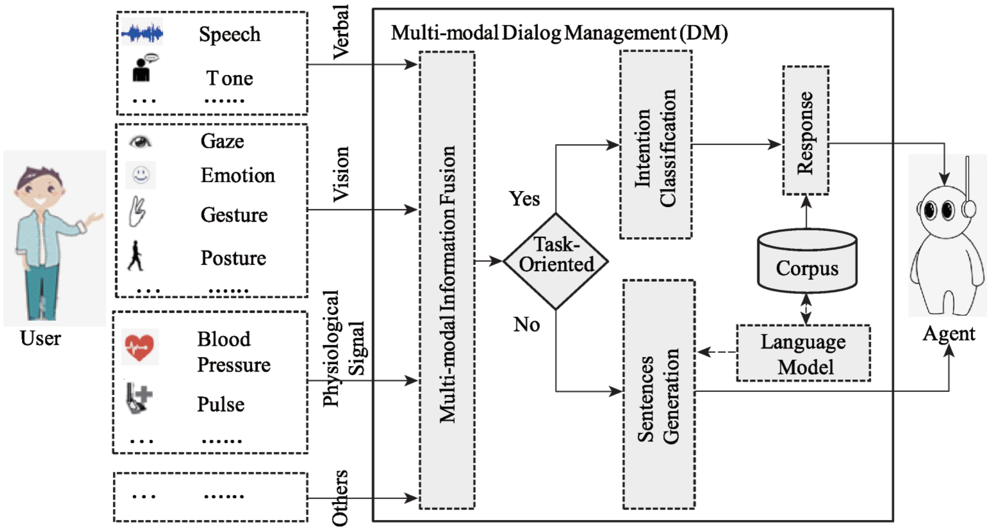

Figure 2 The general framework of multimodal information fusion in human computer dialog.

In Figure 2, the users' behavior information, such as verbal information from speech and tones, visual information from gaze, facial expression, gesture and posture, physiological signal from blood pressure and pulse, are captured and inputted to multi-modal dialog management (multi-modal DM) module. In multi-modal DM, multi-modal information fusion is used to combine all users' behavior information. The methods introduced in section 4 could be used here for multi-modal information fusion. In most cases, users' multimodal information could be presented by a composite feature vector[21] or multi-layer cascaded feature vectors[5]. Generally, we need to consider two types of multi-modal human computer dialog tasks: Task-Oriented and Non-Task-Oriented dialog. For Task-Oriented dialog task, the users' behaviors are needed to be classified into different intentions from feature vectors or cascaded feature vectors. According to different intentions, DM's target is to find appropriate answers to users. While for Non-Task-Oriented dialog, DM's target is to generate suitable sentences to users. We will introduce the feedback generation methods for these two types of tasks in the following subsections.

4.2 Task-Oriented multi-modal human computer dialog

Task-Oriented dialog systems are generally designed to accomplish specific tasks, such as website customer service, vehicle assistants, etc. In most cases, computer Task-Oriented dialog systems are designed to provide information and support users for tasks. Given certain tasks, the target of Task-Oriented dialog systems are designed to complete tasks in possible small amount of interaction turns. Therefore, workflows behind the dialog interaction procedures are important factor for an practical Task-Oriented dialog system. As for Task-Oriented dialog managements, the computer needed to first understand users' intentions accurately, and then generate suitable answers to users.

Depended on the fusion method of multimodal information introduced in section 4, users' multi-modal behavior information could be presented as combined features. Then users' intention calculation is transferred to a classification problem. Intention calculation is essentially targeted to generate suitable answers to users. At this step, DM usually considers answer generation as a problem of sentences retrieve or sentence generation from given patterns according to the dialog states. Sentences retrieve often select appropriate sentences from database for the each state. In this way, dialog state tracking is crucial for Task-Oriented dialog management.

The dialog state tracking challenge (DSTC) endorsed by SIGdial and supported by Microsoft Research, Institute for Infocomm Research and COLIPS, expect the challenge on human dialogs will contribute to progress in developing much more human-like systems[83]. It defines the representation of dialog states and updates them at each moment on a given on-going conversation, and proposes a series of pilot tracks for the core components in developing end-to-end dialog systems based on the same dataset. However, DSTC are originally designed for nature language processing. While in multi-modal dialog system, we need combine multimodal channel information together to obtain users' dialog intention or state. In early researches on this point, in order to reduce the uncertainty and ambiguity of each analysis module on state analysis, multimodal DM usually consider the speech information as the main interactive channel, and used, slot-filling, finite state machine (FSM), graph model, partially observable Markov decision processes (POMDPs), etc, to manage the dialog states. For example, the paper presents a POMDP optimization technique composite summary point-based value iteration, which enables optimization to be performed on slot-filling POMDP-based dialog managers of a realistic size[84]; SmartKom separately parses each channel and gradually extends it to the fusion of each channel through adaptive confidence measurement, forming user intention grid for subsequent processing[85]. FSM is used to integrate and understand the channel information from gestures and voices[86]. According to the influence of different channels on speech interaction, the literature divides their processing methods into three modes: information complementary mode, information fusion mode and information independent mode[10]. Then, semantic understanding and dialogue management are carried out according to the results of speech recognition, expression, posture and emotion. Michaelis and Mutlu constructed the interactive system for children reading companion robot, the interaction process is divided into eight states, according to the different interaction needs and situations of children, the system jumps between eight states to ensure the interaction of robustness and coherence[87]. Directed graph Model is one of the most important tools for man-machine interaction task management[61,88,89]. It can be seen that the purpose of interactive management is to plan the computation of multi-channel information in time and space, so that user interaction is more robust and natural. Given a message, the system retrieves related responses from the repository and returns the most reasonable response. That is to say, we would not generate a new response, but select the most suitable response (originally made to other messages) as reply to the current message[90]. However, these sentence retrieve methods depended on the sentence previously set in database, somehow it is still inflexibility for spoken dialog, where some key words for slot or dialog states were possibly omitted in conversation in a long term dialog. This is a great challenging in dialog tracking traditional methods.

To this end, researchers now combine deep learning techniques in state tracking, and achieve more accurate state tracking in long conversation environments. Neural Belief Tracking (NBT) framework use representation learning to compose them into distributed representations of user utterances and dialogue context from pre-trained word vectors[91]. NPCEditor provides a user-friendly editor for creating a natural language processing component for virtual humans capable of engaging users' behaviors in spoken dialog on a limited domain. It uses statistical language classification technology for mapping from a user's text input to system responses[92]. Recurrent neural network (RNN) encoder-decoder model is used Lowe et al. provide baselines in two different environments[93]: one where models are trained to select the correct next response from a list of candidate responses, and one where models are trained to maximize the log likelihood of a generated utterance conditioned on the context of the conversation. And these two schemes are both evaluated on a recall task, and using vector-based metrics that capture the topicality of the responses. The experiments demonstrate that Dual Encoder model using both RNNs and LSTMs outperform traditional term frequency-inverse document frequency (TF-IDF) methods. It could be seen that in Task-Oriented multi-modal human computer dialog, deep structures outperform traditional statistical model in response generation on the situations that slot key words and dialog states possibly omitted.

4.3 Non-Task-Oriented Multi-modal human computer dialog

Non-Task-Oriented multi-modal human computer dialog system is also called open domain dialogue system or chat system. The challenge of Non-Task-Oriented multi-modal human computer dialog is to generate suitable responses in non-standard chat environment. In recent years, with the rapid growth of social data on the internet, data-driven open domain dialogue system has gradually become a hot topic in academia and industry. The role of human-computer dialogue system has gradually changed from the role of service to the role of chat partner.

As for Non-Task-Oriented multi-modal human computer dialog, a successful sentence-matching algorithm therefore needs to capture not only the internal structures of sentences but also the rich patterns in their interactions, since natural language sentences have complicated structures, both sequential and hierarchical, that are essential for understanding them. Most researchers adopt deep neural network to generate the sentences from large scale of social data. RNN and CNN based end-to-end encoder-decoder model have been widely used in Non-Task-Oriented multi-modal human computer dialog[94,95]. In this task, a typical example is XiaoBin, which presents a response retrieval approach to find responses based on unstructured documents. For each user utterance and inputted images, instead of looking for the best Q-R pair or generating a word sequence based on language generation techniques, XiaoBin selects a sentence by ranking all possible sentences based on language and visual features designed at different levels of granularity[96]. The proposed method, encoder-labeler LSTM, first encodes the whole input sequence into a fixed length vector with the encoder LSTM, and then uses this encoded vector as the initial state of another LSTM for sequence labeling. With this method, it can predict the label sequence while taking the whole input sequence information into consideration[97]. These deep structures have the following advantages compared to traditional language models: (1) the hierarchical sentence modeling through layer-by-layer composition help to the capturing of the rich matching patterns at different levels of abstraction; (2) helps to generate the continuous latent variable representing the high-level semantic content of the response and the response word by word conditioned from the context[98,99].

Other methods, such as attention model[100], reinforcement learning[101], incremental learning[102] and adversarial Learning[103] have been introduced in sentence generation. These machine learning methods help non-goal-driven systems to carry out natural language understanding, reasoning, decision making and natural language generation in order to replicate or emulate the behavior of the agents in the training corpus.

4.4 Conversational schema that interleave Task and non-Task content

Some researchers tried to find some conversational schemes that interleave Task and Non-Task content in a practical system. A type of conversational humanoid robot, which can engage in both task-oriented dialogues and non-task-oriented dialog for accurately understanding human requests or non-task-oriented dialogues to allow humans to enjoy conversations[104]. Reinforcement learning was used to combine these two types of conversation systems smoothly by training a response selection policy[105]. Similarly, different from these pure task or pure non-task systems, Q-learning method and reinforcement learning algorithms were used to train policies that choose among task and non-task candidate responses to optimize towards a coherent, consistent and informative conversation with respect to different users[106,107]. We can see from the literatures that selection policy is the key component for conversational schema that interleaves task and non-task content.

4.5 Challenges

Multichannel information provides robustness interface, error correction mechanism and interactive options to different situations and environments. However, in essence, multichannel user behavior does not belong to interface operations, so simply converting multichannel signals into interface operations or events may be invalid or even useless[6]. Researchers have suggested that multi-channel interaction can lead to a multi-channel problem. The reason is that the multi-channel human-computer interaction gives users a greater degree of freedom of expression, such as voice, posture, emotional expression in the interaction has the characteristics of uncertainty and casualness. The richness and fuzziness of this expression cannot be accurately mapped to the traditional human-computer interaction interface operation, which makes the system feedback inaccurate. The system needs to manage the integration of multi-channel interaction information[108].

Under the premise that each channel can be obtained and their features could be unified representation synchronously, the current multimodal information fusion relies on the rationality of the design of the interactive system, in addition to the accuracy of the single channel information identification. At present, the existing artificial intelligence methods have partially explored the first three points above and achieved some results on some data sets. However, in the multi-modal human-computer dialog applications, there is a lack of a multi-channel human-computer interaction model which satisfies the four characteristics of the appeal. Constructing a multi-channel information fusion and understanding model with intelligent growth, which enables computer systems to learn, understand and integrate new knowledge into existing knowledge in interaction with users, will be an important breakthrough direction of human-computer dialog in future. To adapt to the characteristics of user's freedom of behavior and changing interactive environment in human-computer dialog, multi-channel information fusion technology needs at least the ability to learn and grow with human-computer dialog in a new environment through simply presetting.

5 Conclusion

This paper briefly reviews the hypothesis and experimental verification of multi-channel information fusion in cognitive science and computer science, and then introduces the common multi-channel information fusion models, as well as some examples of practical multimodal human computer dialog managements. We can see that in spite of high performance of users' single channel behavior computing, it is still great challenge to understand users' intention accurately from their multimodal behaviors. One reason for this challenge is that there is still no applicable method for multimodal information fusion. This paper presents a review of data fusion methods in multimodal human computer dialog. We first introduce the cognitive assumption of single channel processing, and then discuss its implementation methods in human computer dialog; then for the task fusion of multi-modal information, we present serval computing models after we introduce the principle description of multiple data fusion.

After we discussed the existing defaults of current multimodal information methods, for example: in the case of minor errors in speech recognition and video analysis, the existing multi-channel fusion models and interaction systems lack the ability of self-correction. The future human-computer dialog applications, such as home service robots, intelligent education, etc., are hoped to own the abilities of learning from users' interactions. Finally, some practical examples of multimodal information fusion methods are presented and the possible and important breakthroughs of the data fusion methods in future multimodal human-computer interaction applications are discussed.

References

1 Olyanitch A V. Information technologies in economics: semiolinguistic aspect. In: Perspectives on the Use of New Information and Communication Technology (ICT) in the Modern Economy. Cham: Springer International Publishing, 2019, 630‒638
DOI:10.1007/978-3-319-90835-9_73

2 Brustoloni J C. Autonomous agents: characterization and requirements. School of Computer Science, Carnegie Mellon University, 1991

3 Engwall O, Bälter O. Pronunciation feedback from real and virtual language teachers. Computer Assisted Language Learning, 2007, 20(3): 235‒262
DOI:10.1080/09588220701489507

4 Wik P, Hjalmarsson A. Embodied conversational agents in computer assisted language learning. Speech Communication, 2009, 51(10): 1024‒1037
DOI:10.1016/j.specom.2009.05.006

5 Yang M, Tao J, Chao L, Li H, Zhang D, Che H, Gao T, Liu B. User behavior fusion in dialog management with multi-modal history cues. Multimedia Tools and Applications, 2015, 74(22): 10025‒10051
DOI:10.1007/s11042-014-2161-5

6 Cohen P R, McGee D R. Tangible multimodal interfaces for safety-critical applications. ACM, 2004, 47(1): 41‒46
DOI:10.1145/962081.962103

7 Jaimes A, Sebe N. Multimodal human‒computer interaction: A survey. Computer Vision and Image Understanding, 2007, 108(1): 116‒134
DOI:10.1016/j.cviu.2006.10.019

8 Meyer S, Rakotonirainy A. A survey of research on context-aware homes. In: Proceedings of the Australasian information security workshop conference on ACSW frontiers 2003. Adelaide, Australia: Australian Computer Society, Inc. , 2003, 21: 159‒168

9 Yang M, Tao J, Gao T, Zhang D, Sun M, Li H, Chao L. The error analysis of intention classification and speech recognition in speech man-machine conversation. In: The 11th Joint Conference on Harmonious Human Machine Environment. Huludao, Liaoning, China: 2015

10 Yang M, Tao J, Li H, Chao L. A nature multimodal human-computer-interaction dialog system. In: the 9th Joint Conference on Harmonious Human Machine Environment. Nanchang, Jiangxi, China: 2013

11 Duric Z, Gray W D, Heishman R, Fayin L, Rosenfeld A, Schoelles M J, Schunn C, Wechsler H. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. Proceedings of the IEEE, 2002, 90(7): 1272‒1289
DOI:10.1109/JPROC.2002.801449

12 Wang L, Hu W, Tan T. Recent developments in human motion analysis. Pattern Recognition, 2003, 36(3): 585‒601
DOI:10.1016/S0031-3203(02)00100-0

13 Seely R D, Goffredo M, Carter J N, Nixon M S. View invariant gait recognition. In: Handbook of Remote Biometrics: for Surveillance and Security. Tistarelli M, Li S Z, Chellappa R, eds. London: Springer London; 2009: 61‒81
DOI:10.1007/978-1-84882-385-3_3

14 Chin K, Hong Z, Chen Y. Impact of using an educational robot-based learning system on students' motivation in elementary education. IEEE Transactions on Learning Technologies, 2014, 7(4): 333‒345
DOI:10.1109/TLT.2014.2346756

15 Pierre-Yves O. The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies, 2003, 59(1): 157‒183
DOI:10.1016/S1071-5819(02)00141-6

16 Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. Neural

Information Processing Systems (NIPS), 2015

17 Ming-Hsuan Y, Kriegman D J, Ahuja N. Detecting faces in images: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(1): 34‒58
DOI:10.1109/34.982883

18 Zhao W, Chellappa R, Phillips P J, Rosenfeld A. Face recognition: A literature survey. ACM Computing Surveys, 2003, 35(4): 399‒458
DOI:10.1145/954339.954342

19 Pantic M, Rothkrantz L J M. Automatic analysis of facial expressions: the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1424‒1445
DOI:10.1109/34.895976

20 Tao J, Tan T. Affective Computing: A Review. In: Affective Computing and Intelligent Interaction. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005: 981‒995
DOI:10.1007/11573548_125

21 Chao L, Tao J, Yang M, Li Y, Wen Z. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. Brisbane, Australia: ACM, 2015: 65‒72
DOI:10.1145/2808196.2811634

22 Wang S, Yan W, Li X, Zhao G, Zhou C, Fu X, Yang M, Tao J. Micro-expression recognition using color spaces. IEEE Transactions on Image Processing, 2015, 24(12): 6034‒6047
DOI:10.1109/TIP.2015.2496314

23 He L, Jiang D, Yang L, Pei E, Wu P, Sahli H. Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. Brisbane, Australia: ACM; 2015: 73‒80
DOI:10.1145/2808196.2811641

24 Ge L, Liang H, Yuan J, Thalmann D. Robust 3D hand pose estimation from single depth images using multi-view CNNs. IEEE Transactions on Image Processing, 2018, 27(9): 4422‒4436
DOI:10.1109/TIP.2018.2834824

25 Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. In: International Conference on Computer Vision. Venice, Italy: 2017, 1, 3
DOI:10.1109/ICCV.2017.525

26 Ruffieux S, Lalanne D, Mugellini E, Abou Khaled O. A Survey of datasets for human gesture recognition. In: Human-Computer Interaction Advanced Interaction Modalities and Techniques. Cham: Springer International Publishing, 2014: 337‒348
DOI:10.1007/978-3-319-07230-2_33

27 Hasan H S, Kareem S A. Human computer interaction for vision based hand gesture recognition: a survey. In: 2012 International Conference on Advanced Computer Science Applications and Technologies. Kuala Lumpur, Malaysia: 2012, 55‒60
DOI:10.1109/ACSAT.2012.37

28 Weiming H, Tieniu T, Liang W, Maybank S. A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2004, 34(3): 334‒352
DOI:10.1109/TSMCC.2004.829274

29 Fagiani C, Betke M, Gips J. Evaluation of tracking methods for human-computer interaction. In: Sixth IEEE Workshop on Applications of Computer Vision, 2002 (WACV 2002) Proceedings. Orlando, FL, USA: IEEE, 2002, 121‒126
DOI:10.1109/ACV.2002.1182168

30 Oviatt S, Cohen P, Wu L, Duncan L, Suhm B, Bers J, Holzman T, Winograd T, Landay J, Larson J, Ferro D. Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. Human‒Computer Interaction, 2000, 15(4): 263‒322
DOI:10.1207/S15327051HCI1504_1

31 Tian F, Xu L, Wang H, Zhang X, Liu Y, Setlur V, Dai G. Tilt menu: using the 3D orientation information of pen devices

to extend the selection capability of pen-based user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Florence, Italy: ACM, 2008: 1371−1380
DOI:10.1145/1357054.1357269

32 Tian F, Lu F, Jiang Y, Zhang X, Cao X, Dai G, Wang H. An exploration of pen tail gestures for interactions. International Journal of Human-Computer Studies, 2013, 71(5): 551−569
DOI:10.1016/j.ijhcs.2012.12.004

33 Pelz J B. Portable eyetracking in natural behavior. Journal of Vision, 2004, 4(11): 14−14
DOI:10.1167/4.11.14

34 Santella A, Decarlo D. Robust clustering of eye movement recordings. Eye Tracking Research and Applications (ETRA), 2003, 27−34
DOI:10.1145/968363.968368

35 Cheng S, Sun Z, Sun L, Yee K, Dey A K. Gaze-based annotations for reading comprehension. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul, Republic of Korea: ACM, 2015: 1569−1572
DOI:10.1145/2702123.2702271

36 Yu C, Sun K, Zhong M, Li X, Zhao P, Shi Y. One-Dimensional handwriting: inputting letters and words on smart glasses. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, California, USA: ACM, 2016, 71−82
DOI:10.1145/2858036.2858542

37 Yu C, Wen H, Xiong W, Bi X, Shi Y. Investigating effects of post-selection feedback for acquiring ultra-small targets on touchscreen. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, California, USA: ACM, 2016, 4699−4710
DOI:10.1145/2858036.2858593

38 Wang D, Zhao X, Shi Y, Zhang Y, Hou J, Xiao J. Six Degree-of-Freedom haptic simulation of probing dental caries within a narrow oral cavity. IEEE Transactions on Haptics, 2016, 9(2): 279−291
DOI:10.1109/TOH.2016.2531660

39 Yang W, Jiang Z, Huang X, Wu X, Zhu Z. Tactile perception of digital images. In: Haptic Interaction. Singapore: Springer Singapore, 2018, 445−447
DOI:10.1007/978-981-10-4157-0_74

40 Paivio A. Mental representation: A dual coding approach. New York: Oxford University Press, 1990
DOI:10.1093/acprof:oso/9780195066661.001.0001

41 Baddeley A D. Working memory. Oxford: Clarendon Press, 1986

42 Cowan N. What are the differences between long-term, short-term, and working memory? In: Sossin W S, LacailleJ-C, Castellucci V F, Belleville S, eds. Progress in Brain Research. Elsevier, 2008, 323−338
DOI:10.1016/S0079-6123(07)00020-9

43 Baddeley A. Working memory: looking back and looking forward. Nature Reviews Neuroscience, 2003, 4: 829
DOI:10.1038/nrn1201

44 Service E. The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. The Quarterly Journal of Experimental Psychology Section A, 1998, 51(2): 283−304
DOI:10.1080/713755759

45 Just M A, Carpenter P A. A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 1992, 99(1), 122−149
DOI:10.1037/0033-295X.99.1.122

46 Nelson C. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 2001, 24, 87−185
DOI:10.1017/S0140525X01003922

47 ChooiW-T, Thompson L A. Working memory training does not improve intelligence in healthy young adults. Intelligence, 2012, 40(6): 531−542
DOI:10.1016/j.intell.2012.07.004

48 Barrouillet P, Bernardin S, Camos V. Time constraints and resource sharing in adults′ working memory spans. Journal of Experimental Psychology: General, 2004, 133(1): 83
DOI:10.1037/0096-3445.133.1.83

49 Maehara Y, Saito S. The relationship between processing and storage in working memory span: Not two sides of the same coin. Journal of Memory and Language, 2007, 56(2): 212−228
DOI:10.1016/j.jml.2006.07.009

50 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. 2006, 313(5786): 504−507
DOI:10.1126/science.1127647

51 Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzago P A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 2010, 3371−3408

52 Chao L, Tao J, Yang M, Li Y. Bayesian fusion based temporal modeling for naturalistic audio affective expression classification. In: The 5th International Conference on Affective Computing and Intelligent Interaction. Geneva, Switzerland: IEEE, 2013, 173−178
DOI:10.1109/ACII.2013.35

53 Miao Y, Gowayyed M, Metze F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Scottsdale, USA: IEEE, 2015, 167−174
DOI:10.1109/ASRU.2015.7404790

54 Caramiaux B, Montecchio N, Tanaka A, Bevilacqua R. Adaptive gesture recognition with variation estimation for interactive systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 2015, 4(4): 18
DOI:10.1145/2643204

55 Mayer R E. Multimedia learning. In: Psychology of Learning and Motivation. Academic Press, 2002, 41: 85−139
DOI:10.1016/S0079-7421(02)80005-6

56 Revlin R. Cognition: theory and practice: Worth Publishers, 2012.

57 Fournet N, RoulinJ-L, Vallet F, Beaudoin M, Agrigoroaei S, Paignon A, Dantzer C, Desrichard O. Evaluating short-term and working memory in older adults: French normative data. Aging & Mental Health, 2012, 16(7): 922−930
DOI:10.1080/13607863.2012.674487

58 Maehara Y, Saito S. The relationship between processing and storage in working memory span: Not two sides of the same coin. Journal of Memory and Language, 2007, 56(2): 212−228
DOI:10.1016/j.jml. 2006.07.009

59 Ernst M O, Banks M S. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 2002, 415: 429
DOI:10.1038/415429a

60 Gunes H, Piccardi M. Affect recognition from face and body: early fusion vs. late fusion. In: 2005 IEEE International Conference on Systems, Man and Cybernetics. Waikoloa, USA: IEEE, 2005, 3437−3443
DOI:10.1109/ICSMC.2005.1571679

61 Yang M, Tao J, Chao L, Li H, Zhang D, Che H, Gao T, Liu B. User behavior fusion in dialog management with multi-modal history cues. Multimedia Tools and Applications, 2015, 74(22): 10025−10051
DOI:10.1007/s11042-014-2161-5

62 Li X, Gao F, Wang J, Strahler A. A priori knowledge accumulation and its application to linear BRDF model inversion. Journal of Geophysical Research: Atmospheres, 2001, 106(D11), 11925−11935
DOI:10.1029/2000JD900639

63 Fang L, Xueyin L, Li S Z, Yuanchun S. Multi-modal face tracking using Bayesian network. In: 2003 IEEE International SOI Conference Proceedings. Nice, France: IEEE, 2003, 135−142
DOI:10.1109/AMFG.2003.1240835

64 Town C. Multi-sensory and multi-modal fusion for sentient computing. International Journal of Computer Vision, 2007, 71(2): 235−253
DOI:10.1007/s11263-006-7834-8

65 Pradalier C. Colas F. Bessiere P. Expressing bayesian fusion as a product of distributions: applications in robotics. In: International Conference on Intelligent Robots and Systems IEEE. IEEE, 2015

66 Savran A, Cao H, Nenkova A, Verma R. Temporal Bayesian Fusion for Affect Sensing: Combining Video, Audio, and Lexical Modalities. IEEE Transactions on Cybernetics, 2015, 45(9): 1927–1941
DOI:10.1109/TCYB.2014.2362101

67 Li W, Lin G. An adaptive importance sampling algorithm for Bayesian inversion with multimodal distributions. Journal of Computational Physics, 2015, 294: 173–190
DOI:10.1016/j.jcp.2015.03.047

68 Yu D, Li D, He X, Acero A. Large-Margin minimum classification error training for large-scale speech recognition tasks. In: International Conference on Acoustics, Speech and Signal Processing. Honolulu, HI, USA: IEEE, 2007, IV-1137

69 He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. IEEE, 2015, 1026–1034
DOI:10.1109/ICCV.2015.123

70 Yang F, Choi W, Lin Y. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2016, 2129–2137
DOI:10.1109/CVPR.2016.234

71 Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng A Y. Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). 2011, 689–696

72 Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. Helsinki, Finland: ACM, 2008: 160–167
DOI:10.1145/1390156.1390177

73 Seltzer M L, Droppo J. Multi-task learning in deep neural networks for improved phoneme recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013, 6965–6969
DOI:10.1109/ICASSP.2013.6639012

74 Tzeng E, Hoffman J, Darrell T, Saenko K. Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2015, 4068–4076
DOI:10.1109/ICCV.2015.463

75 Kaiser L, Gomez A N, Shazeer N, Vaswani A, Parmar N, Jones L, Uszkoreit J. One model to learn them all. arXiv preprint, 2017, 1706. 05137

76 Wang C, de La Gorce M, Paragios N. Segmentation, ordering and multi-object tracking using graphical models. ICCV, 2009, 12: 747–754

77 Wei F, Li W, Lu Q, He Y. A document-sensitive graph model for multi-document summarization. Knowledge and Information Systems, 2010, 22(2): 245–259
DOI:10.1007/s10115-009-0194-2

78 Myunghwan K, Jure L. Latent multi-group membership graph mode. Computer Science, 2012, 80

79 Honorio J, Samaras D. Multi-task learning of gaussian graphical models. ICML, 2010, 447–454

80 Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. 2015, 350(6266): 1332–1338
DOI:10.1126/science.aab3050

81 Wu J, Cheng J, Zhao C, Lu H. Fusing multi-modal features for gesture recognition. In: Proceedings of the 15th ACM on International conference on multimodal interaction. Sydney, Australia: ACM, 2013: 453–460
DOI:10.1145/2522848.2532589

82 Hamouda L, Kilgour D M, Hipel K W. Strength of preference in graph models for multiple-decision-maker conflicts. Applied Mathematics and Computation, 2006, 179(1): 314–327
DOI:10.1016/j.amc.2005.11.109

83 Kim S, D'Haro L F, Banchs R E, Williams J D, Henderson M. The fourth dialog state tracking challenge. In: Dialogues

with Social Robots: Enablements, Analyses, and Evaluation. Jokinen K, Wilcock G, eds,. Singapore: Springer Singapore; 2017: 435−449
DOI:10.1007/978-981-10-2585-3_36

84 Williams J D, Young S. Scaling POMDPs for spoken dialog management. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(7): 2116−2129
DOI:10.1109/TASL.2007.902050

85 Wahlster W. Smartkom: Symmetric multimodality in an adaptive and reusable dialogue shell. In: Proceedings of the human computer interaction status conference. Berlin, Germany: DLR, 2003, 3: 47−62

86 McGuire P, Fritsch J, Steil J J, Rothling F, Fink G A, Wachsmuth S, Sagerer G, Ritter H. Multi-modal human-machine communication for instructing robot grasping tasks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. Lausanne, Switzerland: IEEE, 2002, 1082−1088
DOI:10.1109/IRDS.2002.1043875

87 Michaelis J E, Mutlu B. Someone to read with: design of and experiences with an in-home learning companion robot for reading. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Denver, Colorado, USA: ACM, 2017: 301−312
DOI:10.1145/3025453.3025499

88 Cheng A, Yang L, Andersen E. Teaching language and culture with a virtual reality game. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Denver, Colorado, USA: ACM, 2017: 541−549
DOI:10.1145/3025453. 3025857

89 Sun M, Zhao Z, Ma X. Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Denver, Colorado, USA: ACM, 2017: 556−567
DOI:10.1145/3025453.3025469

90 Ji Z, Lu Z, Li H. An information retrieval approach to short text conversation. Computer Science, 2014

91 Mrksic N, ó. Séaghdha D, Wen T-H, Thomson B, Young S. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In: The 55th Annual Meeting of the Association for Computational Linguistics, 2017

92 Leuski A, Traum D R, Leuski A. Creating virtual human dialogue using information retrieval techniques. Ai Magazine, 2011, 32
DOI:10.1609/aimag.v32i2.2347

93 Lowe R T, Pow N, Serban I V, Charlin L, Liu C W, Pineau J. Training end-to-end dialogue systems with the ubuntu dialogue corpus. Dialogue & Discourse, 2017, 8(1), 31−65

94 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735−1780
DOI:10.1162/neco.1997.9.8.1735

95 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, 2014, 3104−3112

96 Yan Z, Duan N, Bao J, Chen P, Zhou M, Li Z, Zhou J. An information retrieval approach for chatbot engines using unstructured documents. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, 516−525
DOI:10.18653/v1/P16-1049

97 Kurata G, Xiang B, Zhou B. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016
DOI:10.18653/v1/D16-1223

98 Hu B, Lu Z, Li H. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In: International Conference on Neural Information Processing Systems, 2014

99 Serban I V, Sordoni A, Lowe R, Charlin L, Pineau J, Courville A C, Bengio Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In: AAAI, 2017, 3295−3301

100 Yao K, Zweig G, Peng B. Attention with intention for a neural network conversation model. Computer Science, 2015

101 Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D. Deep reinforcement learning for dialogue generation, 2016, 1192 −1202

102 Rieser V, Lemon O, Keizer S. Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vancouver, BC, Canada: IEEE, 2014, 22(5): 979‒994
DOI:10.1109/TASL.2014.2315271

103 Li J, Monroe W, Shi T, Ritter A, Jurafsky D. Adversarial learning for neural dialogue generation. Empirical methods in natural language processing, 2017, 2157‒2169

104 Nakano M, Hoshino A, Takeuchi J, Hasegawa Y, Torii T, Nakadai K, Kato K, Tsujino H. A robot that can engage in both task-oriented and non-task-oriented dialogues. In: 2006 6th IEEE-RAS International Conference on Humanoid Robots. Genova, Italy: 2006, 404‒411
DOI:10.1109/ICHR.2006.321304

105 Williams J D, Young S. Partially observable Markov decision processes for spoken dialog systems. Computer Speech & Language, 2007, 21(2): 393‒422
DOI:10.1016/j.csl.2006.06.008

106 Yu Z, Xu Z, Black A W, Rudnicky A. Strategy and policy learning for non-task-oriented conversational systems. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, 404‒412
DOI:10.18653/v1/W16-3649

107 Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D. Deep reinforcement learning for dialogue generation. In: arXiv preprint arXiv, 2016, 1606, 01541
DOI:10.18653/v1/D16-1127

108 Oviatt S. Mutual disambiguation of recognition errors in a multimodel architecture. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. Pittsburgh, Pennsylvania, USA: ACM, 1999: 576‒583
DOI:10.1145/302979.303163