A Review of Multimodal Interaction

Rongjian Liang
Navigation and Control Research Center
Tsinghua University
Beijing, China
liangrj14@mails.tsinghua.edu.cn

Bin Liang, Xueqian Wang *
National Laboratory for Information Science and Technology, Graduate School at Shenzhen, Department of Automation
Tsinghua University
Shenzhen, China
bliang@robotsat.com, wang.xq@sz.tsinghua.edu.cn
* Corresponding Author

Tao Zhang
Navigation and Control Research Center
Tsinghua University
Beijing, China
taozhang@mail.tsinghua.edu.cn

Gang Li, Kang Wang
National Laboratory for Information Science and Technology, Graduate School at Shenzhen, Department of Automation
Tsinghua University
Shenzhen, China
wychug@163.com,wangkang14@mails.tsinghua.edu.cn

Abstract—Due to the contribution to a more natural and efficient human-computer interaction style, multimodal interactive systems are becoming more and more common-place and cover a wide range of applicants. Therefore, there seems to be an urgent need for an introduction to related concepts of multimodal interaction and an analytical review of the-state-of-the-art. In this paper, we try to present a brief review of some of the significant aspects in multimodal interaction. At first, the definition and advantages of multimodal systems are discussed, and then a brief development history of multimodal interaction is introduced. Next, we touch on the recent development in the breadth of work on input/output modalities and fusion engine. Finally, we provide a personal outlook for the research of multimodal interaction. It is our hope for this paper to provide a picture of the contemporary state of multimodal interaction research and a meaningful outlook for this research field.

Keywords-multimodal; human-computer interaction; input/out modalities; fusion engine; human centered interaction

I. INTRODUCTION

Thanks to the great advancement in the manufacture of sensors and the breakthrough in the recognition based interactive technologies, more and more novel communication channels has been employed by human-computer interactive systems. Unlike traditional WIMP interface, the newly interactive systems enable us to exchange information with computers by a wide range of modalities, such as speech, body movement, haptic interaction, bio-electricity signals and so on.

In recent decades, our comprehension of human-computer interaction is shifting from computer centered interaction to human centered one [1]. Traditionally, human is supposed to adapt themselves to the way in which machines transmit and receive information. Users often have to sit before computers and type our commands with keyboard or click the buttons on the screen with mouse, which was unnatural to us and inefficient. To eliminate or ease such problems, human centered interaction style was introduced. Machines are supposed to adapt to the communication pattern of human beings. Inherently, human interacts with others in a multimodal way. We employ different kinds of modalities or communication channels to express opinions and perceive information. Human centered interactive systems should be multimodal and support natural and intuitive communication with human beings.

Multimodal interactive system has been booming in the past 30 years. It owes to two main reasons. The first one is the widespread of mobile computing devices with rapidly growing computing power. It results in more abundant, even ubiquitous computing resources, and provides a broad application platform for multimodal interaction. Secondly, with the great advance in sensor manufacturing and recognition based interactive technology, more and more communication channels can be applied to interactive systems. We can use many different kinds of communication methods to interact with computer systems. Multimodal interaction has been attracting many attentions and covers a wide range of application domains.

II. DEFINITION AND ADVANTAGES OF MULTIMODAL INTERACTION

With respect to the definition of multimodal interaction, the scientific community hasn't yet reach a consensus. In some literatures, multimodal interaction is defined as “interfaces that use either multiple modalities or multiple channels” [2]. In other words, multimodal interaction systems leverage different modalities (sight, hearing, touch, smell, taste and etc.) and different pathway to exchange information with human. According to this definition, multimedia systems also belongs to multimodal systems. However, multimedia systems do not necessary to extract meaning from the information streams they carry. Some scientists argue to distinguish multimodal systems with multimedia ones [3]. They points out that “both multimedia and multimodal systems use multiple communication channels,” however, a multimodal system

strives for extracting meaning from the information it carries while a multimedia system doesn't.

Whatever the definition adopted, a multimedia system isn't a multimodal system in the true sense. Information is transmit through different communication channels in multimedia systems, but the signals in different channels are simply and independently processed, and the overall meaning is not be strived by the system. The shift from single modality to multimodalities does not reflect only in an increasement of the number of modalities, in fact, it means a leap from machine centered interaction to human centered interaction. Multimodal interaction systems leverage many different modalities and strive for the overall meaning of the information streams, so to better understand users' intention. They aim to provide natural, intuitive and efficient interaction methods for human, towards the full use of human interaction capabilities.

The advantages of multimodal systems differ from particular applicants and the particular modalities and structure of the systems, therefore it is difficult to draw general conclusions. However, many studies has shown the great advantages of multimodal systems over unimodal ones [4, 5]. Multimodal systems may have the following advantages.

- They may extend the information transmission bandwidth and widens the capture range of information, so to improve the information transmission efficiency;
- Different transmission channels may provide redundant or complementary information. Redundant information can attribute to the higher reliability of the understanding of users' intention. Taking advantage of the complementary information, the systems can perceive much more information which cannot be obtained through a single channel alone;
- They may promote the natural and intuitive interaction between human and machines;
- They may have better adaption and usability to different users and application scenarios than unimodal systems.

III. HISTORY

The "Put that there" system [6] developed by MIT Bolt's team in 1980 is considered to the first multimodal system in history. "Put that there" was a spatial data management system that users can use voice and gesture to interact with. As shown in Fig 1. "Put that there" system demonstrated the power of multimodal system to eliminate the ambiguity of user's command and provided a relatively natural and intuitive interaction style.

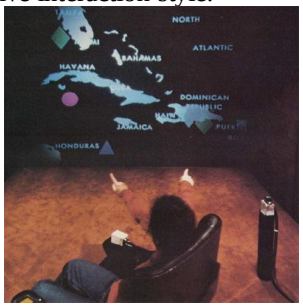

Figure 1. The "Put that there" System.

Following the "Put that there" system, various human-computer interactive systems featuring novel communication channels had been developed and applied in different application domains. Koons's team [7] developed a map-based interactive system integrating voice, gestures and eye gaze channels. Cohen's team developed the Quickset system [8] to which voice and pen-based interactive methods were introduced. To facilitate interaction with the visually impaired person, Nikolaos Kaklanis's team [9] introduced sound and tactile feedback in a map service application. Mohamad Eid's team [10] developed a shooting game integrating tactile, voice, and visual feedback, which greatly improved the sense of immersion to users. F. Galán's team [11] developed a smart wheelchair that can be controlled by electroencephalogram signal. Multimodal interactive systems now cover a wide range of application domains, including education, entertainment, medical, military area, etc.

With the development of multimodal interaction, some related conferences and journals with internationally influence arise, such as ICMI conference (the international conference on multimodal interfaces), ISIF conference (International Society of Information Fusion), the Transactions on Interactive Intelligent Systems, and etc.

IV. INPUT AND OUTPUT MODALITIES

Up to now, the modalities having been used in multimodal systems can be roughly divided into four main directions: vision based [12], sound based, haptic based [13] and bioelectricity based. Each direction include some input and output channels, as the Table.1 shows. Surely, some communication channels utilized in interactive systems do not belong to anyone of the four directions, such as the olfaction based [14] channels and the magnetism based channels. Due to the space limitation we do not discuss them in this paper.

The vision based input channels used in interactive systems include facial expression [15, 16], gestures [17-21], head movement [22], the whole body action [23-26], eye gaze [27], and etc. Skin color [28, 29], shape of the head, facial features [30, 31] are often used to recognize and locate the faces. Then the machine learning technics can be employed to recognize the facial expression. Vision based motion detection has a lot of advantages. It has small interference to human body, and the measurement condition is easy to be satisfied. However, the processing algorithm for vision based motion detection faces some enormous challenges, such as the self-occlusion problem and ambiguity problem [32]. The vision based detection of head movement is relatively easier than the detection of gestures and the whole body action. Because the shape of head is rather simple and unchanging, while the geometry of hands and the whole body are much more complex and changeable [33], and the self-occlusion problem is unavoidable in most applications. Image display is one of the most common used output channels and can provide a great amount of information to users. Recently, 3D display is becoming more and more popular and has been applied in many interaction systems. Compared with traditional 2D screen display, 3D technology can provide the users with vivid stereoscopic images which are no longer confined to

the screen plane. It improves users' immersive experience greatly.

TABLE I. INPUT/OUTPUT CHANNELS UTILIZED IN INTERACTIVE SYSTEMS

	Input channels	Output channels
Vision based	Facial expression Gesture Head movement Whole body action Eye gaze	2D display 3D display
Sound based	Speech Non-speech audio	Speech Non-speech audio
Haptic based	Touch based location and selection Touch based gesture Force and force moment input	Force and force moment output Tactile output
Bioelectricity based	EEG (electroencephalogram) EMG (electromyogram) EOG (electrooculogram)	/

Sound based interaction is natural, intuitive and easy to learn, and has been applied to many interactive systems. Sound based interaction can be divided into speech interaction [34-36] and non-speech audio interaction [37]. Speech interaction has always been the research hotspot in the field of human-computer interaction. In the last 3-5 years, as the Deep Learning Theory [38, 39] being introduced to the research of speech recognition, the recognition rate of speech increases by big percentages. Deep Neural Network has replaced the traditional Hidden Markov-Gaussian Mixture Model [40, 41] and become the mainstream in the research of speech recognition. Major Internet giants such as Google, Baidu, etc., have invested a lot of money and research resource to develop their speech recognition systems, and bring out more and more mature commercial products.

Haptic is one of the five main senses (sight, hearing, haptic, smell, and taste) of human and is the only bidirectional transmission modality among. It means that human can transmit and perceive information simultaneously through haptic modality. Haptic based communication channels allow us to push or pull an object, to rotate it, and to feel its inertia, damp, shape and surface texture. Haptic based input channels, including touch based location and selection, touch based gesture, force input, and etc. Haptic based output channels can be roughly divided into force output [42], and tactile output [13, 43]. The first two input channels have been widely used in smart phone, tablet PC, and other devices with a screen. Force input as well as force and tactile feedback is of great importance to interactions in which human has to manipulate objects in virtual environment or remote site. They allow users to operate intuitively and feel what is happening in the environment.

In recent years, bioelectricity technology has made great progress and became accessible to interactive systems. Bioelectricity-based input channels including EEG [44](electroencephalogram), EMG(Electromyogram), EOG (electrooculogram), and etc. EEG signal is of special importance to severely paralyzed person or patients with neuromuscular disease [45]. Such patients cannot send control signals to their limbs so they cannot move their legs or arms, while they can generate EMG signals normally. By processing their EEG signals, their intended motions can be identified and be used to control the prosthesis or some smart instruments such as wheel chair, as shown in Fig. 2. EMG can also be used to estimate users' action. There is a small time interval between the moment EMG signals can be detected and the moment muscle is activated. Taking advantage of this time interval, human's action can be estimated before it really happens. Comparing with actions estimation with EEG signals, estimating actions with EMG signals is more precise. Recently, EOG signal has attracted some attentions. EOG can be utilized to detect outer retinal disease, and serve as control signal to wheelchair or prosthetics.

Figure 2. Brain-controlled Wheelchair.

V. FUSION ENGINE

Fusion engine – also referred to as multimodal integration – is the critical technic for multimodal interaction. Fusion engine aims to combine and interpret information from various input modalities. As mentioned in previous papers [46], multimodal fusion engine is facing a number of challenges:

- Heterogeneous data – Multimodal systems collect the qualitatively different data such as visual, auditory, haptic and bioelectric measurements. Both cases have to be handled properly by fusion engine.
- Imperfect or inconsistent data [47] – Data provided by input devices is unavoidable to be affected by some level of uncertainty and impreciseness. Data from different sources may be highly conflicting.
- The temporal issue [48] – Different modalities may be employed sequentially or in parallel. Fusion engine has to handle the problem of representing temporal knowledge and temporal reasoning.

…

The existing multimodal interactive systems have provided experience worthy of learning in dealing with above issues. To address the problem of heterogeneous data fusion, Cohen's team employed the Feature Structure [49] and integrated information from pen-based input and voice through unification in Quickset system. Nigay's team employed the Melting pot [50] and integrated data from different modalities in a tabular form. Multimodal parse tree [51] was utilized by Milota's team. The Finite State machine [52] was utilized by Bourguet's team.

As reviewed by Bahador Khaleghi [47] et al, a number of mathematical theories have been introduced to the fusion engine research to tackle the issue of imperfect or inconsistent data. The probability theory [53], possibility theory, Dempster-Shafer evidence theory [54], fuzzy set

theory, rough set theory and random finite set theory [55, 56] are all useful tools to deal with the problem of imperfect or inconsistent data. However, all of these theories have their limitations and are only capable of addressing specific aspects of this issue. A more powerful theory is still in urgent need. When the information received is inconsistent, some useful policies have also been used, such as defining priorities amongst the input channels in advance [8], or iterative testing the "N-best" fusion results.

Representing temporal knowledge is of great significance in fusion engine. As summarized by Lalanne [3] et al., the time representation can be divided into two levels: quantitative and qualitative. Quantitative time means to attach a given amount of time or a precise moment in time to the event, while qualitative time pay attention to the order of the events, such as precedence, simultaneity and succession.

VI. Outlook

Multimodal interaction is still in a rapidly developing phase and it is difficult to forecast what will happen in multimodal interaction research community. Here, we just try to give personal opinions about the meaningful and promising issues in the field of multimodal interaction.

Integrating with user's implicit information as well as the context of use [57] - according to the Relevance Theory, the speaker only conveys as much information as needed in any given context, so that the audience has to recover their intended meaning taking into account what was said/written as well as the context. In order to realize human-centered interaction, computer systems are supposed to capture not only what the user explicitly conveys but also the user's implicit information (preference, emotion, fatigue, and etc.) as well as the context of use. Fusion engine could take advantage of the overall information to get a more reasonable and consistent interpretation.

Taking advantage of the domain knowledge – as mentioned in previous studies [58], it is difficult or infeasible to collect a sufficiently large dataset for classical machine learning algorithms due to the high variability of complex patterns. Therefore, it is obligatory to make use of additional domain knowledge to better integrate the input streams.

Arranging the outputs – while most of attention has been paid to input modalities and their fusion, the arrangement of the output channels attracts less attention [59]. In fact, the arrangement of output channels is of great importance to realize a seamless and efficient human-computer interaction. Multimodal systems are supposed to take advantage of their multiple output modalities to convey information in a more natural and efficient way.

Paying attention to the role of human in interactive systems – in some recent research, the concept of hard/soft sensors [60] has been introduced. The hard sensors represent the traditional electronic devices, which soft sensors represent human. Human is regarded as a special kind of sensor and plays an irreplaceable role in some scenarios (such as a judgment of the relationship). Human computer interactive system is a man-in-loop system. In an interactive system, human may act as a sensor, actuator, decision maker, audience, and etc. Considering human's role in a different angle may open up a novel research direction in the field of multimodal interaction.

VII. Conclusions

The present paper gives a brief survey of multimodal interaction, including the definition, advantages, and history of multimodal interaction, discussion about the key technical issues of input/output modalities and fusion engine, as well as a personal outlook. It is our hope for this paper to provide a picture of the contemporary state of multimodal interaction research and a meaningful outlook of this research field.

References

[1] J. Kong, W.Y. Zhang, N. Yu, X.J. Xia. Design of human-centric adaptive multimodal interfaces[J]. International Journal of Human-Computer Studies, 2011, 69(12): 854-869.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] M. Turk Multimodal interaction: A review[J]. Pattern Recognition Letters, 2014, 36: 189-195.K. Elissa, "Title of paper if known," unpublished.

[3] L. Nigay, J. Coutaz. A design space for multimodal systems: concurrent processing and data fusion[C]. Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems. ACM, 1993: 172-178.

[4] S. Oviatt, R. Coulston, R. Lunsford. When do we interact multimodally?: cognitive load and multimodal communication patterns[C]. Proceedings of the 6th international conference on Multimodal interfaces. ACM, 2004: 129-136.

[5] S. Oviatt, R. Lunsford, R. Coulston. Individual differences in multimodal integration patterns: What are they and why do they exist?[C]. Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2005: 241-249.

[6] R. A. Bolt. "Put-that-there": Voice and gesture at the graphics interface[M]. Morgan Kaufmann, San Francisco, CA, 1998.

[7] D. B. Koons, C. J. Sparrell, K. R. Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures[J]. MIT Press: Menlo Park, CA, 1993: 257-276.

[8] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, et al. QuickSet: Multimodal interaction for distributed applications[C]. Proceedings of the fifth ACM international conference on Multimedia. ACM, 1997: 31-40.

[9] N. Kaklanis, K. Votis, D. Tzovaras. Open Touch/Sound Maps: A system to convey street data through haptic and auditory feedback[J]. Computers & Geosciences, 2013, 57: 59-67.

[10] M. Eid, A. El Issawi, A. El Saddik. Slingshot 3D: A synchronous haptic-audio-video game[J]. Multimedia tools and applications, 2014, 71(3): 1635-1649.

[11] F. Galán, M. Nuttin, E. Lew, P.W. Ferreza, G. Vanackerc, et al. A brain-actuated wheelchair: asynchronous and non-invasive brain–computer interfaces for continuous control of robots[J]. Clinical Neurophysiology, 2008, 119(9): 2159-2169.

[12] R. Poppe. A survey on vision-based human action recognition[J]. Image and vision computing, 2010, 28(6): 976-990.

[13] C. Garre, M. A. Otaduy. Haptic rendering of objects with rigid and deformable parts[J]. Computers & Graphics, 2010, 34(6): 689-697.

[14] E. Richard, A. Tijou, P. Richard, J. L. Ferrier. Multi-modal virtual environments for education with haptic and olfactory feedback[J]. Virtual Reality, 2006, 10(3-4): 207-225.

[15] Y. Wang, X. Yang, J. Zou. Research of Emotion Recognition Based on Speech and Facial Expression[J]. TELKOMNIKA Indonesian Journal of Electrical Engineering, 2013, 11 (1): 83-90.

[16] K. Karpouzis, G. Caridakis, R. Cowie, E. D. Cowie. Induction, recording and recognition of natural emotions from facial expressions and speech prosody[J]. Journal on Multimodal User Interfaces, 2013, 7(3): 195-206.

[17] X. Teng, B. Wu, W. Yu, C. Liu. A hand gesture recognition system based on local linear embedding[J]. Journal of Visual Languages & Computing, 2005, 16(5): 442-454

[18] H. I. Suk, B. K. Sin, S. W. Lee. Hand gesture recognition based on dynamic Bayesian network framework[J]. Pattern Recognition, 2010, 43(9): 3059-3072.

[19] Z. Feng, B. Yang, Y. Chen, Y. Zheng, T. Xu, et al. Features extraction from hand images based on new detection operators[J]. Pattern Recognition, 2011, 44(5): 1089-1105.

[20] M. F. Ho, C. Y. Tseng, C. C. Lien, C. L. Huang. A multi-view vision-based hand motion capturing system[J]. Pattern Recognition, 2011, 44(2): 443-453.

[21] A. Just, S. Marcel. A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition[J]. Computer Vision and Image Understanding, 2009, 113(4): 532-543.

[22] G. Zhao. Research on key techniques of vision-based large head pose tracking[D]. Zhejiang University, 2009.

[23] Zappella L, Lladó X, Provenzi E, J. Salvi. Enhanced local subspace affinity for feature-based motion segmentation[J]. Pattern Recognition, 2011, 44(2): 454-470.

[24] P. Antonakaki, D. Kosmopoulos, S. J. Perantonis. Detecting abnormal human behaviour using multiple cameras[J]. Signal Processing, 2009, 89(9): 1723-1738.

[25] A. Mokhber, C. Achard, M. Milgram. Recognition of human behavior by space-time silhouette characterization[J]. Pattern Recognition Letters, 2008, 29(1): 81-89.

[26] T. B. Moeslund, A. Hilton, V. Krüger. A survey of advances in vision-based human motion capture and analysis[J]. Computer vision and image understanding, 2006, 104(2): 90-126.

[27] Ç. Çığ, T. M. Sezgin. Gaze-based prediction of pen-based virtual interaction tasks[J]. International Journal of Human-Computer Studies, 2015, 73: 91-106.

[28] J. Yang, X. Ling, Y. Zhu, Z. Zheng. A face detection and recognition system in color image series[J]. Mathematics and Computers in Simulation, 2008, 77(5): 531-539.

[29] J. Yang, C. Liu C, J. Yang. What kind of color spaces is suitable for color face recognition?[J]. Neurocomputing, 2010, 73(10): 2140-2146.

[30] K. M. Lee. Component-based face detection and verification[J]. Pattern Recognition Letters, 2008, 29(3): 200-214.

[31] S. Phimoltares, C. Lursinsap, K. Chamnongthai. Face detection and facial feature localization without considering the appearance of image context[J]. Image and Vision Computing, 2007, 25(5): 741-753.

[32] R. Poppe. Vision-based human motion analysis: An overview[J]. Computer vision and image understanding, 2007, 108(1): 4-18.

[33] K. S. Patwardhan, S. D. Roy. Hand gesture modelling and recognition involving changing shapes and trajectories, using a Predictive EigenTracker[J]. Pattern Recognition Letters, 2007, 28(3): 329-334.

[34] T. N. Tran, W. Cowley, A. Pollok. Automatic adaptive speech separation using beamformer-output-ratio for voice activity classification[J]. Signal Processing, 2015, 113: 259-272.

[35] J. M. Górriz, J. Ramírez, E. W. Lang, C. G. Puntonetc, I. Turiasd. Improved likelihood ratio test based voice activity detector applied to speech recognition[J]. Speech Communication, 2010, 52(7): 664-677.

[36] T. Oonishi, K. Iwano, S. Furui. A noise-robust speech recognition approach incorporating normalized speech/non-speech likelihood into hypothesis scores[J]. Speech Communication, 2013, 55(2): 377-386.

[37] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni , F. Piazza. An integrated system for voice command recognition and emergency detection based on audio signals[J]. Expert Systems with Applications, 2015, 42(13): 5668-5683.

[38] B. Yin, W. Wang W, L. Wang. Review of Deep Learning[J]. Journal of Beijing University of Technology, 2015, 1: 011.

[39] W. Hu, Y. Qian, F. K. Soong, Y. Wang. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers[J]. Speech Communication, 2015, 67: 154-166.

[40] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[41] C. C. Cheng. Online learning of large margin hidden Markov models for automatic speech recognition[J]. 2011.

[42] T. P. James, J. J. Pearlman, A. Saigal. Predictive force model for haptic feedback in bone sawing[J]. Medical engineering & physics, 2013, 35(11): 1638-1644.

[43] M. K. X. J. Pan, J. McGrenere, E. Croft, K.E. MacLean. Exploring the Role of Haptic Feedback in Enabling Implicit HCI-Based Bookmarking[J]. Haptics, IEEE Transactions on, 2014, 7(1): 24-36.

[44] X. Li, X. Gao, S. Gao. A hybrid brain-computer interface system based on two different paradigms[J]. Chinese Journal of Biomedical Engineering, 2012, 31(3): 326-330.

[45] U. Chaudhary, N. Birbaumer, M. R. Curado. Brain-Machine Interface (BMI) in paralysis[J]. Annals of physical and rehabilitation medicine, 2015, 58(1): 9-13.

[46] D. Lalanne, L. Nigay, P. Robinson, J. Vanderdonckt, J. F. Ladry. Fusion engines for multimodal input: a survey[C]. Proceedings of the 2009 international conference on Multimodal interfaces.

[47] Khaleghi B, Khamis A, Karray F O, et al. Multisensor data fusion: A review of the state-of-the-art[J]. Information Fusion, 2013, 14(1): 28-44.

[48] J. F. Allen. Maintaining knowledge about temporal intervals[J]. Communications of the ACM, 1983, 26(11): 832-843.

[49] M. Johnston, S. Bangalore. Finite-state multimodal parsing and understanding[C]. Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000: 369-375.

[50] L. Nigay, J. Coutaz. A generic platform for addressing the multimodal challenge[C]. Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co., 1995: 98-105.

[51] A. D. Milota. Modality fusion for graphic design applications[C]. Proceedings of the 6th international Conference on Multimodal interfaces. ACM, 2004: 167-174.

[52] M. L. Bourguet M L. A toolkit for creating and testing multimodal interface designs[C]. Proceedings of UIST'02. 2002: 29-30.

[53] M. Munz, M. Mählisch, K. Dietmayer. Generic centralized multi sensor data fusion based on probabilistic sensor and environment models for driver assistance systems[J]. Intelligent Transportation Systems Magazine, IEEE, 2010, 2(1): 6-17.

[54] C. Liang, Z. Chen, Y. Huang, J. Tong. A Method of dispelling the absurdities of Dempster-Shafer's rule of combination[J]. Systems Engineering - Theory & Practice, 2005, 25(3): 7-12.

[55] S. Reuter, K. Dietmayer, S. Handrich. Real-time implementation of a random finite set particle filter[J]. Sensor Data Fusion: Trends, Solutions, Applications (SDF 2011), 2011.

[56] S. Reuter, K. Dietmayer. Pedestrian tracking using random finite sets[C]. Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on. IEEE, 2011: 1-8.

[57] A. K. Dey. Understanding and using context[J]. Personal and ubiquitous computing, 2001, 5(1): 4-7.

[58] M. Glodek, F. Honold, T. Geier, G. Krelld, F. Nothdurft, et al. Fusion paradigms in cognitive technical systems for human–computer interaction[J]. Neurocomputing, 2015, 161: 17-37.

[59] F. Vernier, L. Nigay. A framework for the combination and characterization of output modalities[M]. Interactive Systems Design, Specification, and Verification. Springer Berlin Heidelberg, 2001: 35-50.

[60] D. L. Hall, M. McNeese, J. Llinas, T. Mullen. A framework for dynamic hard/soft fusion[C]. Information Fusion, 2008 11th International Conference on. IEEE, 2008: 1-8.