Call for Proposals

Human-Computer Interaction

HIRP OPEN 2017

Application Deadline: 09:00 A.M., 16th June, 2017 (Beijing Standard Time, GMT+8).

If you have any questions or suggestions about HIRP OPEN 2017, please send Email (innovation@huawei.com). We will reply as soon as possible.

Catalog

HIRPO2017150301: The low-cost roadside vision sensor for target localization and target tracking

1 Theme: Human-Computer Interaction

2 Subject: Low-cost Roadside Sensor Technology and Algorithm Research

List of Abbreviations

3 Background

Vehicle perception of the environment is the basis of intelligent driving，depending on the vehicle sensor, the sensing range and sensing capabilities are limited。In order to support advanced intelligent driving, the required vehicle sensor costs are also very high，the low cost roadside vision sensor can greatly enhance the perceptual range of vehicles, and reduce the cost of vehicle sensor.

4 Scope

This project focuses on the analysis of the latest research progress of roadside vision sensor in academia and Industry, put forward own solution of roadside vision sensor

- What are the current solutions of roadside vision sensor in academia and Industry, for these solutions, please describe the advantages and disadvantages, the application effect.
- Study on the roadside vision sensor hardware system: not limited to the use of standard camera, depth camera, infrared camera; not

limited to the use of fusion of vision sensors with other low-cost sensors (Such as fusion of camera with microwave or millimeter wave sensor)

- Study on visual recognition algorithm: multi object real-time tracking, target classification, distance, width, position and velocity information, not limited to the use of deep learning algorithm etc. The results shall meet the following requirements specifications：
  - target classification: pedestrian, slow moving object
  - target distance: pedestrian 0~50 meters, the error is less than 0.3m
  - target position: X, Y, Z coordinate information, the error is less than 0.3m
  - target speed: speed range 0~40km/hour, error less than 1m/s
  - Other information: target width, error less than 10%
  - Information reporting frequency: 20Hz
  - Continuous tracking target number: Up to 16
  - In the multi-target scene, the recognition rate is greater than 80%

Main application scenarios

- Study on visual tracking algorithms: the ability to track targets in multiple visual sensors

5 Expected Outcome and Deliverables

- Research Report - the current research status of the roadside vision sensor
- Solution design report and acceptance test report
- Provide the algorithm model and code which can verification the above algorithm

The following deliverables are only as a condition of preferred partners:

- (Optional)The roadside vision sensor prototype system and code
- (Optional)One patent

6 Acceptance Criteria

1、 The algorithm meet the needs of the scene, and can be implemented.

2、 The algorithm has been reviewed by experts.

3、 The prototype meet the needs of the scene, and can be implemented.

7 Phased Project Plan

Phase1 (~2 months): The latest research survey of roadside vision sensor in academia and Industry, delivery research report.

Phase2 (~2 months): Study on the solution of road object recognition and tracking using roadside vision sensor, delivery solution design report.

Phase3 (~8 months): Develop and provide prototype systems, algorithms, codes, simulation results, test reports and patents

Click here to back to the Top Page

HIRPO2017150302: The creation and understanding of intelligent space

1 Theme: Human-Computer Interaction

2 Subject: The Creation and Understanding of Intelligent space

List of Abbreviations

IoT	Internet of things

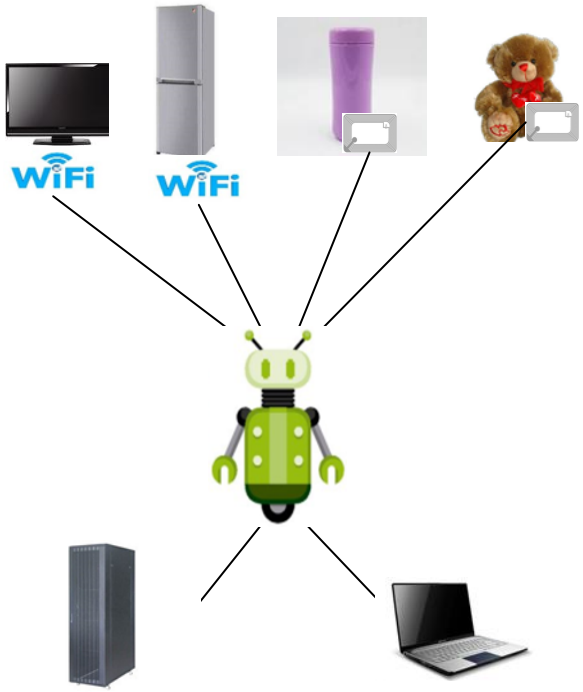
3 Background

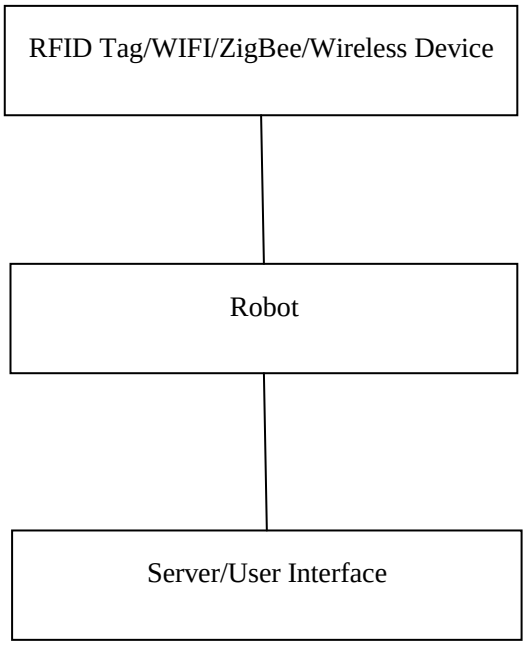
The intelligent space technology research of the service robot has important social significance and research value. It is a challenge leading direction in the field of service robotics research. This project focuses on the perception of intelligent space and the control of service robots in intelligent space.

In order to recognize the people, facilities, service robots in the intelligent space, the digital standard source technologies are studied. The information of each element in intelligent space is built. Then the virtual Space is constructed based on the mapping of space elements from physical dimension to information dimension. It can break the time and space constraints in the information dimension to "panoramic perception", real-time access, accurate, complete, global, intelligence space elements and the environment information, percept and understand the events occurred in the intelligence space. It also can be used to global optimization control the service robot scientifically to provide corresponding support and service for people.

4 Scope

System architecture:





a. RFID Tag/WIFI/ZigBee/Wireless Device：

In the home space, a variety of different IoT devices or furniture can be found through difference wireless positioning techniques

b. Robot：

a). Identification and management of different IoT devices or furniture within the intelligent space through robot

b). Percept and understand the events occurred in the intelligence space

c. Server/User Interface：

a). Intelligent space modeling and data storage through the server

b). Through the user interface you can see the indoor space layout and a variety of equipment and furniture location, running status and other information

Requirement:

a. Need algorithm design and theoretical proof.

b. Need modeling design and simulation

c. Survey of intelligent space techniques and systems

Simulation software:

a. Need to do:

a). Concise 2D UI for map of house，IoT device, furniture, etc.

b). Simulate discovery of different type of device, connection status of the device, running status of the device, etc.

c). Simulate events occurred in the intelligence space

b. Don't need to do:

Quasi physical robot

Performance:

a. Positioning distance: 1cm~5m

b. Positioning accuracy:1cm~1m

5 Expected Outcome and Deliverables

Proposal stage：

Submit proposal, include:The current research status; Scheme, design and data model; Background Intellectual property; Human resources; Budget (include purchase); Schedule.

1. **T+0 ~ T+2 (month)**
   - Data model (design document)
   - Positioning algorithm design document
   - Simulation design document
2. **T+2 ~ T+6 (end of project)**
   - Positioning algorithm prototype
   - Simulation
   - One patent and one report

6 Acceptance Criteria

4、 The intelligent space design specification has been reviewed by experts.

5、 Intelligent space system demonstration

7 Phased Project Plan

Phase1 (~2 months): Data model (design document), Positioning algorithm design document; Simulation design document.

Phase2 (~6 months): Positioning algorithm prototype, Simulation, One patent and one report

Click here to back to the Top Page

HIRPO2017150303: Personalized healthcare and early disease detection

1. Theme: Human-Computer Interaction

2. Subject: AI in Healthcare

List of Abbreviations

NA

3. Background

There is significantly increasing demand for people to monitor their health and fitness at home, on a daily basis. Such monitoring could enable early disease detection at home, so that people can seek for professional help from doctors as early as possible. The recent deep learning technology enables people to retrieve valuable information from big amount of data. The key components are the deep learning technology and the data. On the technology side, we have increasing amount of publications from academic and industries, open source platforms like Tensorflow and Caffe etc; However, on the data side, it's behind the wheels. In other words, relevant healthcare data is crucial for training the data model, which is of great value.

4. Scope

4.1 Provide functional databases relevant to the following scenarios:

- Environmental data: Air quality, temperature, humidity info of various geographical locations on a daily basis.

- Medical knowledge database: General knowledge regarding various disease introduction, symptoms, cause, treatment and prevention.
- Cancer knowledge database: General knowledge regarding various cancer introduction, symptoms, cause, treatment and prevention.

5. Expected Outcome and Deliverables

- Open access to a complete knowledge database as requested above.
- The knowledge database should be in Chinese language, and covers more than 90% disease and cancer types which can be found on Google or Baidu search engine.
- Provide API in certain programming language in order to access and retrieve information from the database.

6. Acceptance Criteria

- The complete knowledge database can be accessed, modified and maintained.
- Project receiver should be able to access the knowledge database in the preferred programming language.
- The database is accessible for the upcoming years.

7. Phased Project Plan

Phase1 (~4 months): State-of-the-art knowledge databases review on the

scope, content, front end, back end, necessary hardware and software.

Phase2 (~4 months): Creation or modification to the existing knowledge database according to the project requirements.

Phase3 (~4 months): Implementing software API to access, modify, retrieve certain information from the database.

Click here to back to the Top Page

HIRPO2017150304: Research on environment sensitive dialogue policy

1 Theme: Human-Computer Interaction

2 Subject: Dialogue System

List of Abbreviations

IA	Intelligent Assistant
DM	Dialogue Management
NLU	Natural Language Understanding
NLG	Natural Language Generation

3 Background

Dialogue systems have been widely researched and applied in intelligent assistants (IA) on smart devices. A typical dialogue system consists of natural language understanding (NLU), dialogue management (DM), and natural language generation (NLG).

Currently, most dialogue systems take recognized text of user speech as the only input from real world, and exploit conversation context to improve performance. However, the interactive experience of an automated dialogue system is still far from talking to human. An important reason is that current systems leverage little information about dialogue environment.

Environmental information such as time, location, people and objects can be easily obtained by embedded sensors and analyzed in the cloud, which provides meaningful context for human-machine interaction.

Thus, it is necessary and valuable to research how to utilize environmental information to enrich interactive context and build a more intelligent dialogue system.

4 Scope

Environmental context classification and representation

Investigate and classify typical indoor environmental factors that can be captained by commodity sensors and can be used to enrich dialogue context. Propose a flexible format for representing environmental context, and figure out how to transfer/synchronize such context between dialogue domains.

Utilize environmental context to refine dialogue policy

Investigate user intention recognition/understanding by utilizing environmental context. Explore new dialogue policies that can proactively extend and deepen dialogue topics based on available environmental context.

5 Expected Outcome and Deliverables

- Technical reports of context management and utilization in dialogue systems;
- Technical reports of how to classify and represent environmental context;
- Prototypes of utilizing environmental context to refine user intention understanding and extend dialogue topics
- 1 invention/patent with corresponding paper accepted by top conferences/journals;

6 Acceptance Criteria

- Technical reports cover typical approaches proposed in academic papers and public industrial solutions;
- Propose reasonable and workable approach for representing environmental context;
- A Demonstrated prototype that can effectively extend dialogue topics for more than 5 turns based on properly simulated environmental context;
- Invention idea accepted by patent review committee in Huawei, corresponding paper accepted by top conferences/journals.

7 Phased Project Plan

Phase1 (~3 months): Survey state-of-the-art approaches of context management and utilization in dialogue systems.

Phase2 (~3 months): Research on classifying and representing environmental context;

Design approaches of refining dialogue policies based on environmental context;

Invention idea proposal

Phase3 (~6 months): Prototyping, simulation, and paper writing

Click here to back to the Top Page

HIRPO2017150305: Dialog emotion detection based on context information

1 Theme: Human-Computer Interaction

2 Subject: Emotion Detection

3 Background

Emotion is one of the most important information during communication and the change of emotion states affects the expression and transmission information, which leads to the change of people's perception and decision-making. Usually, the information carrier of emotion includes speech, words, facial expression, body movements and other physiological processes. In recent years, emotion detection has become a research hotspot due to its wide range of application, and also become a key field of pattern recognition. Via emotion detection, human computer interaction can have emotion communication.

Emotion detection attempts to recognize the emotion or attitude of the current object through speech, facial expressions and body movements. However, the emotion is not expressed separately. In multi-round human computer interaction，the specific scenario or context in which the dialog occurs is usually a key to the emotion. These information can't be obtained when only the current input is considered. So we should study emotion detection based on the context information. Current research focuses on using multi-modal data of the current state to improve the accuracy, but the multi-round dialog context in practical application has rarely been considered. In order to acquire good interaction experience, this project aims to study and solve this problem.

4 Scope

Context based emotion detection from text

1. In multi-round dialog, one sentence often do not contain enough information, such as elliptical sentence, co-reference, how to use the memory of previous utterances to full understand the user's semantic.

2. How to select and extract the context or scenario information that associate with emotion detection as complementary features; how to use these features to build up models.

Speech Emotion detection Based on Acoustic Context

1. How to combine the current utterance and previous utterances to extract acoustic features; how to select and deal with the features to build models.

2. Study the relevance of adjacent utterances, construct emotion transition probability and emotion reasoning rule, in order to assist detecting emotion.

The method to merge different context models

1. How to merge text and speech context information, how to merge the proposed model with current multi-modal (image and speech) models.

5 Expected Outcome and Deliverables

- Source code, algorithm description, as well as code documentation, etc.
- Technical reports or survey of the state of the art of emotion detection algorithm using context, including how to merge different models.
- Technical reports or survey of the method of extract and deal with context information.
- 1 patent;

6 Acceptance Criteria

In the multi-turn human-robot interaction, the algorithm can detect the emotion of user input under some specific scenario, the precision is higher than 80%

7 Phased Project Plan

Phase1 (~3 months): survey the state of the art of emotion detection based on context, analyze and provide the related technical report.

Phase2 (~6 months): research on how to extract, select and merge context features from text and speech, how to build up and merge models.

Phase3 (~3 months): research and provide related algorithms, results and patents.

Click here to back to the Top Page

HIRPO2017150306: Real-time multi-talker signal separation

1 Theme: Human-Computer Interaction

2 Subject: Speech Processing Technology

3 Background

Automatic Speech Recognition (ASR) is well known as an important technology in artificial intelligence area, ASR based applications and functions are already integrated into all kinds of smart devices to offer people a new generation of natural interaction.

However, ASR system is not perfect yet, its robustness limits its capability for application in complex environment. For example, performance of ASR always suffers in following scenarios:

- Strong background noise
- Far-field scenario
- Multi-talkers scenario

To solve this kind of problem, lots of algorithms and approaches are imported to make noise polluted or distorted speech signal recover as much as original one, some of them have got promising development：

- Adaptive noise filter and DNN based filter are used to get rid of the influence of background noise and already got outstanding performance.

- Microphone array depends on its sound source localization and beam forming ability to pick up speech from really long distance and keep available SNR.

Unfortunately, as far as we know, there are lack of efficient methods to deal with cocktail party effect problem, which we mentioned above as multi-talkers scenario.

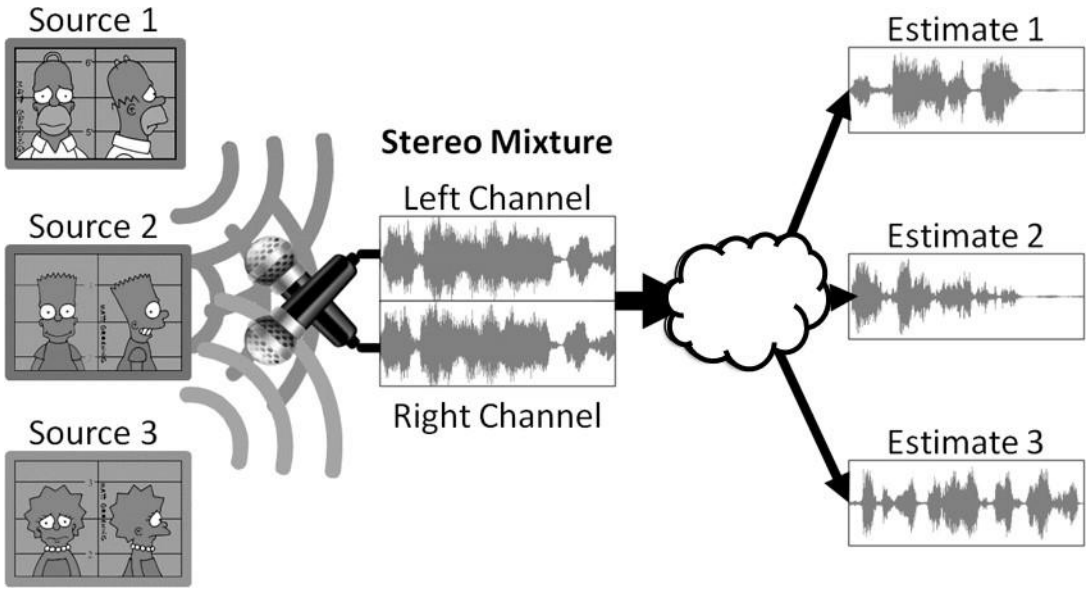


Fig.1 Scheme of multi-speaker separation system

Cocktail party effect is a phenomenon of being able to focus one's auditory attention on a particular stimulus while filtering out a range of other stimuli, much the same way that a partygoer can focus on a single conversation in a noisy room. Human can filter out the speech they are interest in because they can use binaural processing to make complex noised speech simple and clean, then use their brain, the neural system, to track target speech signal.

Once we enable machine accomplishes similar function, ASR system could help people in more and more situations. As demonstrated in Fig.1, supposing we receive a signal composes speech signals of lots of speakers, our final

target is to separate speech of each speaker from original mixed signal, then build a solution which is able to deal with cocktail party problem.

4 Scope

Identifying target speaker in multi-talker scenario, tracking and extracting speech signal of target speaker, guarantee the extracted signal behaves good quality for ASR system (no auditory quality standard required). All requirements are summarized as follow:

- Speaker identification function
- Target speaker tracking function
- Target speaker speech extraction function

5 Expected Outcome and Deliverables

- The state-of-the-art investigation report of speech signal separation, including normal background noise environment and multi-talker scenario,;
- Technical reports of real-time multi-talker speech separation system, including solution design, architecture creation and core algorithm development;
- Source code of multi-talker speech separation system;
- 1~2 Invention/patents/papers;

6 Acceptance Criteria

Hardware limitation of solution:

- Maximum number of microphones: 3

Performance required:

- Background speech supervision: > 15dB
- Word error rate enhancement: > 50%

7 Phased Project Plan

Phase1 (~2 months): survey the state of the art of real-time speech signal separation for normal background noise environment and multi-talker scenario, identify the problems, metrics and requirements in this topic, forms technical reports.

Phase2 (~6 months): Research on real-time multi-talker signal separation, including solution design and architecture creation. Experiment report based on proposed algorithm and solution.

Phase3 (~4 months): Algorithm and solution optimization, make sure system performance achieves target criteria. Submission of real-time multi-talker signal separation system, including source code, module description, processing introduction and all test reports.

Click here to back to the Top Page

HIRPO2017150307: Multi-modal emotion detection

1 Theme: Human-Computer Interaction

2 Subject: Emotion Detection

3 Background

Emotion is one of the most important information during communication and the change of emotion states affects the expression and transmission information, which leads to the change of people's perception and decision-making. Usually, the information carrier of emotion includes speech, words, facial expression, body movements and other physiological processes. In recent years, emotion detection has become a research hotspot due to its wide range of application, and also become a key field of pattern recognition. Via emotion detection, human computer interaction can have emotion communications.

Emotion detection attempts to recognize the emotion or attitude of the current object through speech, facial expressions and body movements. Among all these modalities, Speech and facial expressions are the most important ones and such has always been of great concern in emotion detection technology. Single modal emotion detection, especially on speech and image (facial expression) has achieved some good results. But in many cases, single modality may miss some effective information or contains too much noise that leads to detection failures. So, emotion detection via multimodal fusion including video, speech, image and even physiological features, is an effective way to solve the problem of low accuracy of single modal emotion recognition.

There is an urgent need for multimodal emotion recognition in robot interaction. When talking with the robot, we want the robot can identify the current

emotional state of the speaker, so as to select the corresponding response mode, which is an important quality to enhance the robot experience. Specifically, the robot needs to identify the current emotional state of the speaker via speech, facial expression, communication contents and body movements, then provide it to the decision making module. This is a typical multimodal emotion detection scenarios, requires a complete set of multi modal pattern recognition system to integrate speech, image and other information mode.

This project aims to study how to leverage the multi modal information of human-computer interaction to implement real-time emotion detection. We hope be able to develop a practical multi modal emotion detection for different specific human populations and specific scenarios. Since we want the system can be easily adapted to different populations and scenarios, the system should not only be a trained model, it's important to find out how to do the feature selection, model selection and model fusion as a package. Unlike other modes, speech is quite different for different populations, which means that the system should select a whole set of new speech features for new scenarios. So another focus of this program should put on the speech feature selection.

4 Scope

The research plan of this project includes the following two aspects:

1） The emotion detection system via image and speech

Image and speech are the two most important modalities for emotion detection. Single modality emotion detection based on image or speech has got some results but the single modal information is not enough because there is complementarity between different modalities of information. Therefore, on

the basis of single modal, this project focuses on how to fuse the information of image and speech to improve the accuracy of the final system identification.

2） Feature selection for speech

The speech modality is different from other modes, the differences of language, age, gender and other factors will cause significant differences among different populations. However, there are no specific features of speech found out yet that can be used to express emotion. Therefore, this project hopes to develop a method that can extract speech features for specific populations to implement emotion detection. Thus we can quickly select the best features of speech when constructing new scenario or application to reduce the workload of development.

5 Expected Outcome and Deliverables

- Dual modal emotion detection algorithms (image and speech). Source code, algorithm description, as well as code documentation, etc.
- Feature selection algorithms for speech. Source code, algorithm description, as well as code documentation, etc.
- 1 patents;

6 Acceptance Criteria

The system can detect the emotion of user input during human-robot interaction under some specific scenario with a precision higher than 80%.

7 Phased Project Plan

Phase1 (~3 months): survey the state of the art of multi-modal emotion detection, provide the related technical report and implement a baseline system.

Phase2 (~6 months): research on how to improve the system and optimize the feature selection algorithm.

Phase3 (~3 months): research and provide related systems, algorithms, results and patents

Click here to back to the Top Page

HIRPO2017150308: Emotional state analysis based on speech audio signal

1 Theme: Human-Computer Interaction

2 Subject: Future HMI-Speech Audio Signal Analysis

List of Abbreviations

NA

3 Background

Human can express their emotions to the outside world through various ways. We as humans have got used to percept other's feelings through their moves, demeanors, speeches and intonations. However, Automatic sentiment extraction from natural audio streams containing spontaneous speech is a challenging area of research that has received little attention. If computers are successful at recognizing human emotions, we can take counter measures, for example, playing smoothing music to an anxious driver to alleviate road rage.

4 Scope

1) Build the in-vehicle speech utterance dataset for sentiment analysis and define the set of class labels for different types of emotional states, e.g., angry, sad and happy.

2) Build the model for sentiment classification based on features extracted from speech utterance. The model can be a supervised, unsupervised or a deep learning model.

3) Validate and test the model against a specific test set. Compare the model with at least one well-known baseline model and discuss and advantages and

disadvantages of the model. Also suggest any extension or future work needed to be done.

5 Expected Outcome and Deliverables

1. Literature review on the state-of-the-art on emotion state analysis based on speech audio signal.

2. In-vehicle Audio dataset containing spontaneous utterances of representative emotional states.

3. Source code containing the configuration of the model and corresponding training algorithms.

4. Documentation or research paper detailing the design and implementation of the core algorithm\system.

5. 1~2 Invention/patents.

6 Acceptance Criteria

Project proposal is accepted by the evaluation team, Huawei.
Project deliverables are accepted by the evaluation team, Huawei, as follows:
Speech audio dataset should collect no less than 1000h utterance samples from different subjects.
Model accuracy > 80% or relative accuracy improvement > 10% against well-known baseline model

7 Phased Project Plan

Phase1 (~3 months): Survey the state of the art on emotional state analysis based on speech audio dataset, and propose the overall algorithmic and architectural design.

Phase2 (~3 months): Collect utterance data and build the emotional speech database.

Phase3 (~3 months): Build and fine-tune the model and compare with at least one baseline model to show performance improvement. Paper/patent write-up.

Click here to back to the Top Page

HIRPO2017150309: User intention prediction model research

1 Theme: Human- Computer Interaction

2 Subject: Future HMI- Proactive Service Interaction

List of Abbreviations

HMI	Human Machine Interface
AI	Artificial Intelligence

3 Background

If there is a system that can keep perceiving and predicting the driver's intention and provides proactive service, it will give better user experience. Based on the technology of AI and deep learning, it should be possible to perceive the user's personal real time data stream from multiple resources, such as from cameras and microphones inside the car cockpit, the information from GPS and maps, from daily routine activities, from living preferences, from calendars, and any other possible resources, for example:

User status may include: facial expressions, motion, gesture, and so on.

Environmental semantics may include: car temperature, speaker volume, seat parameters, traffic data, etc.

User behavior intention may include: want to turn left/turn right, want to stop, want to turn on air conditioning, want to listen to music, or want to answer the phone, and so on.

It’s valuable to establish relational model from the user state and context to the user's behavior intention.

4 Scope

Focus on specific scene (e.g. car driving/using)

Through the analysis of a large number of car driving/using samples, establish relational model from the user status and context to the user's behavior intention and provide the specific parameters of the model to determine the principles and scoring.

This topic can be decomposed into many specific sub research topics, that depending on the researcher’s specification and interests. The scope can variant into different dimensions. Any suggestions are welcome.

5 Expected Outcome and Deliverables

- 1 paper/patent.
- The final delivery is the model and document of decision rules, without the need for code implementation.

6 Acceptance Criteria

The model should cover most car driving and using use case.

7 Phased Project Plan

Phase1 (~3 months): survey the car driving and using samples, and provide the related survey report.

Phase2 (~3 months): establish relational model from the user status and context to the user's behavior intention and provide the specific parameters of the model to determine the principles and scoring.

Click here to back to the Top Page

HIRPO2017150310: Automatic rule extraction from textual data

1 Theme: Human- Computer Interaction

2 Subject: Future HMI- Natural Language Processing

List of Abbreviations

NLP: Natural Language Processing

NLU: Natural Language Understanding

3 Background

The state-of-the-art technique for spoken language understanding involves matching user utterance against a lexical pattern, or the so called grammar rule, to derive user intention. The grammar rule, or simply rule, is usually in the format of a regular expression, which is a symbolic representation of many representative real-world utterances. Currently, each rule is created manually by human experts, who derive the rule by summarizing over many textual utterances. The process of manually creating the rules is labor intensive. Furthermore, the rules thus built suffer from many problems such as validity, the conflict between rules. It is highly valuable if the rule can be automatically created and validated from a large text corpus in the target domain.

4 Scope

1) Create the text corpus in the in-vehicle conversation dialog domain. Specific domain could include navigation, music, phone and radio, etc.

2) Design and implement the algorithm for automatically extracting grammar rules from the text corpus. State-of-the-art algorithms from the areas of information retrieval and text mining might be helpful.

3) Validate the generated grammars and evaluate against several well-used criteria such as accuracy, coverage, etc.

5 Expected Outcome and Deliverables

1. Literature review on the existing approaches on automatic grammar rule extraction from text and its evaluation methods.

2. A text corpus in the domain of in-vehicle conversational dialogue.

3. Detailed design and documentation of the proposed algorithm including source code, configuration scripts, etc.

4. Research results should be published in a highly-cited conference proceeding or journal.

5. 1~2 Invention/patents;

6 Acceptance Criteria

Project proposal is accepted by the evaluation team, Huawei.
Project deliverables are accepted by the evaluation team, Huawei, as follows:
Rule coverage > 90%, individual rule accuracy > 85%

7 Phased Project Plan

Phase1 (~3 months): Survey the state of the art of automatic rule extraction from text and propose overall algorithmic and architectural design.
Phase2 (~2 months): Collect in-vehicle conversational dialog text corpus.

Phase3 (~3 months): Design and implement the proposed algorithm for rule extract. Conduct comparison with baseline algorithms. Evaluate experiment results.

Phase4 (~1 month): Paper/Patent write up

Click here to back to the Top Page