Software Proposal Document for project LipDrive

Ziad Thabet, Amr Nabih, Karim Azmi, Youssef Samy
Supervised by: Dr Mai El-Shehaly and Eng Silvia Soliman

November 14, 2017

Abstract

Lip-reading, according to the Cambridge dictionary, is to understand what someone is saying by watching the movements of their mouth. Lip-reading plays a vital role in human communication and speech understanding however, it is a difficult task to be done by humans. Lip-reading is so essential for hearing-impaired people as they, if taught, can recognize what a person is saying from the movement of lips. However, it takes years for a person to learn to read lips. Not only is it used for people who lost the ability to hear but also for biometrics identification, silent dictation or even on surveillance cameras. In addition, it can be used to communicate with autonomous vehicles. The main goal of this project is to develop a deep learning approach for the real-time detection of spoken words in autonomous vehicles setting. The training of a deep-layered CNN and RNN is being used to convert lips movement to written words.

1 Introduction

1.1 Background

Car accident rate has been increasing lately all over the world. According to the Association for Safe International Road Travel [12], nearly 1.3 million people die in road crashes each year, on average 3,287 deaths a day. In addition, 20-50 million are injured or disabled. There is a clear theme to the vast majority of these incidents: human error and inattention. Thus, the enhancement and working on Autonomous Vehicles is important to reduce these numbers. Autonomous vehicle (AV) is the ability to a vehicle to drive itself, which means that autonomous vehicle has to imitate the human driver's performance. However, relying on autonomous approaches alone is also error prone and can have severe consequences. Therefore, recent AV research aims to combine autonomous approaches with effective Human-Computer Interaction, in order to facilitate the collaboration between humans and vehicles. This project aims to help the driver in the vehicle, to deliver certain commands to the vehicle in a noisy environment, like passengers talking or radio is on, through lip-reading. This will increase the accuracy of speech recognition from the driver to the vehicle. These

commands could be as: Start Ignition, stop the car, or set direction to a certain destination. Most of the commands will need to be done instantaneously, like change lane or reduce the speed limit. This project would help AV to receive commands in real-time without switching off music or asking other passengers to stop talking.

1.2 Motivation

1.2.1 Market Motivation

Users have stated that the main problem, when communicating to an intelligent personal assistant like Siri or Google, is the vague detection of words due to the surrounding noise. On the 10th of November 2018, we have launched a questionnaire to identify users' problem with IPA. After collecting 130 responses, 79 of them use IPA and mostly 58.8 percent have faced problems ordering an IPA due to the surrounding noise. This leads to the dissatisfaction of some users that 74.5 percent of them stop using the application, while 17 percent of them ask people to stop talking/or reduce the noise.

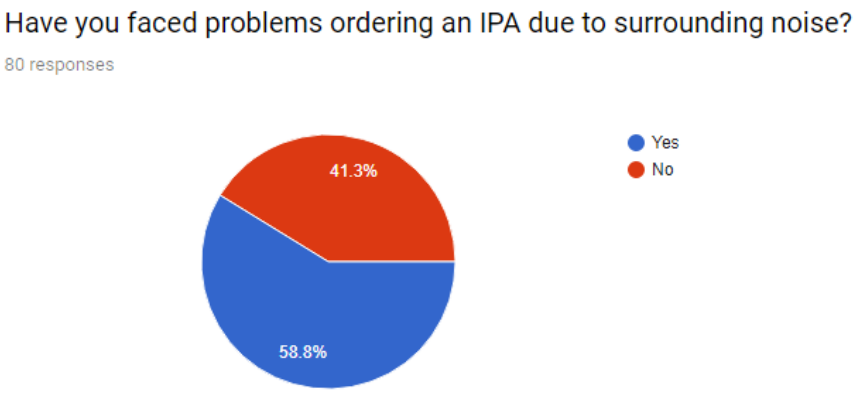


Figure 1: Statistics 1

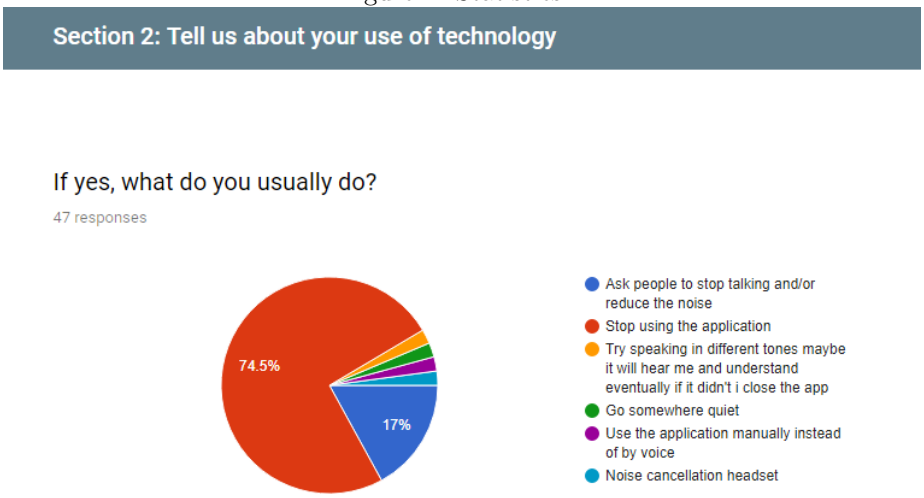


Figure 2: Statistics2 2

1.2.2 Stakeholders

The Dubai's Roads and Transport Authority (RTA), has carried out the first test run of an autonomous vehicle, it is expected to be set to launch operations soon. Dr. Ismail Hisham Zohdy, Chief Specialist/Program Manager of Self-Driving Transport in Roads and Transport Authority (RTA) in Dubai, has mentioned his concerns about this project, stating that he is going to be part of this project's stakeholder team.

1.2.3 Academic motivation

Our work is motivated by both the application domain and previous work in the literature. Xinjun Ma, Hongjun Zhang and Yuanyuan Li explored Lip reading under variant lighting conditions[2] [7]. Sonu Lamba, Neeta Nain and Harendra Chahar have developed a Multi-Model approach for face detection in crowd[9]. Meanwhile, Assael and Yannis M., et al. have improved the accuracy of detection and Real-time processing [6]. Mengchen Liu, Jiaxin Shi and Zhen Li have used Deep Convolutional Neural Networks Approach to detect lip-reading[8]. Not only CNN but also Amit Garg , Jonathan Noyola, Sameep Bagadia have used Deep Recurrent Neural Networks Approach.[10].

1.3 Problem Definitions

Enhance the accuracy of speech recognition through Lip-reading in order to detect certain commands in an autonomous vehicle.

2 Project Description

The LipDrive system aims to develop a deep learning approach for real-time detection of spoken words in an autonomous vehicle setting. Not only does the system work on the word level but also on the sentence level such as *"LipDrive turn left"*, *"LipDrive go home"*. The system will facilitate the detection of words at variant speeds of speech. This system aims to help a driver in an autonomous vehicle (AV), to deliver certain commands to the vehicle in a noisy environment, like passengers talking or radio is on, through lip-reading. This will increase the accuracy of speech recognition between the driver and the vehicle. These commands could be as: *"Start Ignition, stop the car"*, or *"set direction to a certain destination"*. Most of the commands will need to be done instantaneously, like *"change lane"* or *"reduce the speed limit"*. This project would help AV to receive commands in real-time without turning off music or asking other passengers to stop talking. The system is composed of a camera which is set in the AV, that will detect the driver's lips movement and based on a trained data, the system will be able to detect the spoken command.

2.1 Scope

The system will cover several things inside its scope:

1. Lips reading, under variant lighting conditions all the way from good subset of natural light to external dark light conditions in other words dimmed[2]
2. The system will have its own implemented recurrent neural network, containing a set of commands for the vehicle, it will be able to predict the upcomming letters and words through a training phase.
3. The system will work on the sentence level not the word level such as "LipDrive turn left, LipDrive go home .."
4. The system will also cover variant speed of speaking [10]
5. The system will not cover multi model approach as we have one driver and we are not working on crowded place

2.2 Project Overview

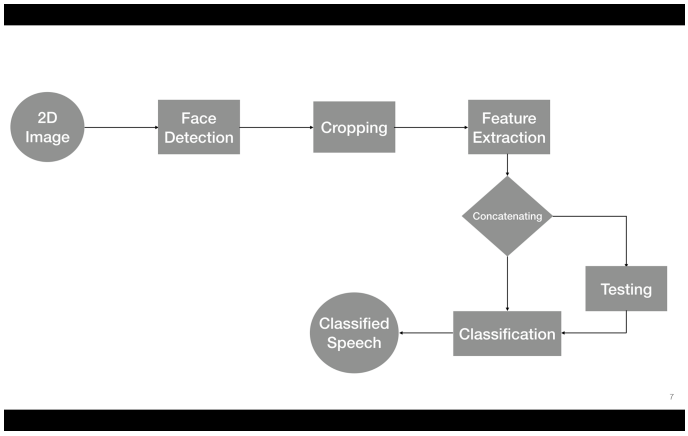


Figure 3: Flow Diagram

1. Face detection: the input will be a 2D image and the system will start by detecting all facial features
2. Cropping and adjusting image: the image will be cropped to mouth level avoiding any loss of important information in our case the lips, chin, nose and cheeks, then will be adjusted to a specific width and length to help in feature extraction similarity of points.

3. Feature extraction: the features will be extracted from the mouth image using a CNN structure as in Figure 4 into a vector of Points (x,y), the vector will consist of 20 points

Figure 4: Feature extraction

4. Concat: every vector will be added to the previous vector to create a sequence
5. Recurrent Neural Network training: the RNN will be earlier fed with efficient amount of different instances as in sequence form that will be given a label of the sequence's command
6. Classification: in this phase the system will be taking a real time camera footage and concatinating image by image to get the the speech classification
7. Measuring the accuracy: we will feed the RNN with instances of every command said by different poeple then we will calculate the precentage of how many time the system managed to succeed in labeling the sequence

3 Similar System Information

1. Lip-Reading by Surveillance Cameras [1]:
   (a) Researchers had written this paper to increase the safety of citizens among crime activities and terrorist attacks.

(b) The main problem is that microphones attached to these cameras are not able to analyze speech in noisy environments and the long distance between the microphone and speaker.

(c) Researchers had developed a prototype to record and understand movements of mouth under limited conditions using the Dutch language.

(d) Researchers had discussed different steps to read lips such as:

i. Data-acquisition
ii. Lip tracking
iii. Feature extraction
iv. Recognition

(e) This paper discusses the option to enable CCTV Cameras to read lips of persons on a large distance and to recognize abnormal behaviors.

2. Feature Extraction Method for Lip-reading under Variant Lighting Conditions [2]:

(a) Researchers had written this paper to increase the accuracy of detecting lip movement under variant lighting conditions.

(b) The main problem is that lip reading is mainly conducted under the ideal lighting conditions.

(c) Researchers had proposed a new method of lip feature extraction under variant lighting conditions. The method consists of a preprocessing chain of illumination normalization and improved LBP features, which can improve the recognition rate of lip-reading under variant lighting conditions from two aspects.

(d) This paper will help in detecting lips under different light conditions such as morning, afternoon, night, direct and indirect light

(e) According to the effects caused by variant lighting conditions on lip-reading recognition result, researchers had put forward a new method for extracting lip features. This method consists of a illumination preprocessing chain and a light invariant feature extraction operator.

3. A novel approach for lip Reading based on neural network [3]:

(a) Researchers had written this paper because lips reading has an advantage of high accuracy and noise independency.

(b) The main problem is that speech recognition systems are not accurate as much as lips reading system due to the surrounded noise.

(c) The researchers had presented an algorithm for automatic lip reading. The algorithm consists of two main steps:

i. Feature extraction
ii. Classification for word recognition.

The lip information is extracted using lip geometric and lip appearance features. The words are recognized throughout neural network.

(d) The accuracy achieved by proposed approach is 97% . The proposed algorithm is applied for recognition of ten words of Hindi language and can be easily extended to include more words of other languages

(e) This algorithm is important due to its accuracy and it can help in developing an automatic lip reading system for practical applications.

4. Development of Novel Lip-reading Recognition Algorithm [4]:

   (a) Researchers had written this paper to find a way to reconstruct total laryngectomy people's basic communication function

   (b) The main problem is that people affected by "Total laryngectomy" lose their natural voice and their communication functions.

   (c) A novel speech recognition algorithm was developed to extract the features of lip contour and real-time recognize each English vowel when speaking. Here, the criteria settings were developed to improve the stability of detecting the mouth ROI and lip contour.

   (d) The performance of lip-reading recognition under different backgrounds and individual differences was also tested, and the accuracy of the proposed algorithm on lip-reading recognition was over 80% and the total initialization time and the total tracking time were 0.5 seconds and 0.27 seconds respectively.

   (e) This algorithm is important due to its accuracy and its time efficiency.

5. Lip Reading in the Wild [5]:

   (a) Researchers had written this paper to recognize the words being spoken by a talking face, given only the video but not the audio.

   (b) The main problem is having a video without its audio file.

   (c) Researchers had collected their data-sets over TV broadcasts, they developed CNN architectures that are able to effectively learn and recognize hundreds of words from these data-sets and they demonstrate a recognition performance.

   (d) This algorithm was tested on 333-word test set and it achieved 65.4% and then achieved 92.3% .

   (e) This paper is important because it shows the way of collection data and the algorithm has high performance (accuracy).

6. LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING [6]:

   (a) Researchers had written this paper to decode text from the movement of a speaker's mouth and to create a lipreading system that is working on the sentence-level.

(b) The main problems are designing or learning visual features, prediction and the existing work on models trained end-to-end perform only word classification not sentence-level sequence prediction. Studies have shown that human lipreading performance increases for longer words, indicating the importance of features capturing temporal context in an ambiguous communication channel.

(c) Researchers had introduced LipNet which is a model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. LipNet is the first end-to-end sentence-level lipreading model that simultaneously learns spatiotemporal visual features and a sequence model.

(d) LipNet had achieved 95.2% accuracy in sentence-level.

(e) This paper is important because it shows how to read lips over a huge sentence not just word-by-word.

7. Mobile Phone Security using Automatic Lip Reading[13]:

(a) Researchers had written this paper because mobile phones usage and their applications in the everyday life has increased dramatically.

(b) The main problem is that the usage of smart phones has increased and the appearance of mobile banking applications, so it is important to find new different methods to secure those phones.

(c) Researchers had introduced a new security method in which password is received using the phone camera that track users lip movement using lip reading techniques to identify the security words like passwords.

(d) A mobile banking application has been implemented and it was tested. The algorithm has successfully recognized the password with about 70% of accuracy.

(e) This paper is important because it shows the usage of lip reading in different fields such as security with a very good percentage of accuracy.

8. Read My Lips, Login to the Virtual World[14]:

(a) Researchers had written this paper to find a new method to login into smart devices to obtain high level of security and to be user friendly at the same time.

(b) The main problem is that the login systems in smart devices require many factors to get the highest level of security and at the same time those systems should be easily used by the user.

(c) Researchers had introduced a security application that satisfies those requirements using the neural network approach (LSTM) with good accuracy, it uses the lip reading technique to transfer visual movements into the password.

(d) The implemented application was used to improve security accuracy as it achieved 93.8% by single iteration from the first result.

(e) This paper is important because it shows the importance of high accuracy in the field of lip password.

9. A Survey of Visual Lip Reading and Lip-Password Verification[15]:

   (a) Researchers had written this paper to combine between the two techniques: visual lip reading and lip motion password verification, those techniques are very powerful and flexible to use in many applications especially in the security field. This paper includes the double check process which is verifying both the speaker and his/her password.

   (b) Sometimes the normal sound is not available or the speech signal is incomplete, for example: in noisy environment.

   (c) Researchers had reviewed the process of lip reading and lip password verification passing by different steps:

      i. Feature Extraction.
      ii. Classification Schemes.
      iii. Localization and Segmentation.

   (d) This paper is important because it shows the different techniques used to build a security application that depends on two important techniques which are the lip reading and the lip password verification.

10. A Person Identification System Combining Recognition of Face and Lip-Read Passwords[16]:

   (a) This paper is for person identification system which is a combination of facial feature recognition as well as word recognition through lip movement. It combines a face recognition algorithm to identify the person, followed by words recognition through lip reading for password authentication, those two techniques added together are most powerful and flexible than to use in many applications especially in the security field. This paper includes both verification and authentication.

   (b) Sometimes specification isn't just enough so adding an authentication to our system would make it more secure and reliable.

   (c) This paper has used some techniques such as:

      i. Facial recognition PCA for feature extraction
      ii. To reduce dimensions from feature extraction KNN
      iii. ASR for lip movement recognition
      iv. Lip movement is recognized to its coresponding word by DWT

   (d) The importance of this paper is enhancing accuracy of security, facial recognition alone is 90% and spoken words by lip reading password is 72% combining them would lead to 98% accuracy

11. Automated Lip Reading Technique for Password Authentication[17]:

    (a) This system for extracting and transforming movements of individual lips into a pattern already recognized password which is set for the system using lip recognition image processing techniques. the breakthrough in this system is to provide motor-disability people a fast and simple way to protect their data. to capture and trace the successive sequence of lip movement during speech, the corresponding word can be recognized

    (b) Bio-metric-security is the most secure to authenticate in high security zones and a war is growing intense every hour for protecting the confidentiality of data from being intercepted.

    (c) This paper has used some techniques such as:

        i. Syllable tracking
        ii. Personalizing
        iii. Block Diagram
        iv. Simulation Procedure
        v. RGB Filter
        vi. Image encryption

    (d) The importance of this paper is that the usage of lip reading in passwords has proven to be much more convenient, secure and user-friendly technique for user authentication

3.1 Similar System Description

Lip Reading is the task of decoding text from the movement of a speakers mouth. LipNet is facing two major problems which are designing or learning visual features, and prediction. LipNet is a neural network architecture for lip reading and it is the first model to apply deep learning to end-to-end learning of a model that maps sequences of image frames of a speakers mouth to entire sentences. LipNet is based on high sentence level. Working on the sentence-level instead of working on the word level, LipNet had successfully achieved a very high accuracy 95.2%.

3.2 Comparison with Proposed Project

Previous System	Our System
Day light	Under variant light condition
Using tensorflow	Using Dlibrary
Accuracy is 95.2	Accuracy is almost the same or more
No specific application	Our application willl be on autonomous vehicle
Small scale of dataset	Large scale of dataset

3.3 Screen Shots from previous systems

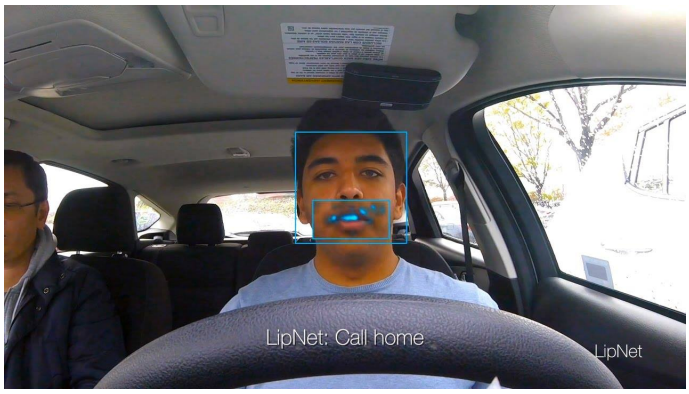


Figure 5: LipNet

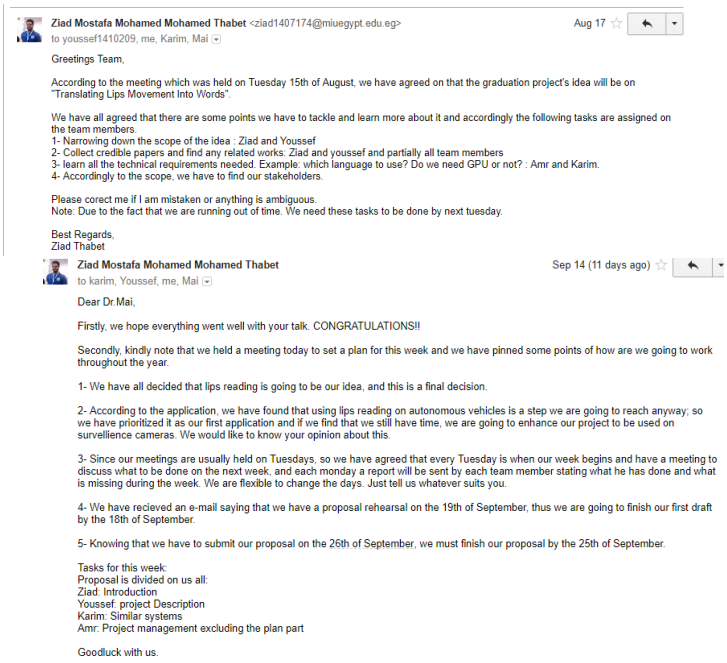
Figure 6: Other system

4 Project Management and Deliverables

4.1 Tasks and Time Plan

Task Name	Start	Finish
Idea discussion	07/26/17	08/11/17
Idea Research	08/15/17	09/15/17
Survey and proposal	09/16/17	09/25/17
Proposal presentation	09/26/17	09/26/17
Designing application	10/03/17	10/06/17
Implementing prototype	10/07/17	10/10/17
Implementing GUI desig	10/11/17	10/13/17
Designing database	10/14/17	10/17/17
Class diagram	10/18/17	10/20/17
SRS writing	10/21/17	11/08/17
SRS presentation	11/09/17	11/09/17
Data set collection	11/10/17	11/17/17
Data classification	11/18/17	11/24/17
Implementing applicatio	11/25/17	12/22/17
SDD writing	12/23/17	01/16/18
SDD presentation	01/19/18	01/19/18
Validation and testing	01/31/18	03/27/18
Implementation evaluati	03/29/18	03/29/18
Writing paper	03/31/18	04/10/18
Delivering the paper	04/12/18	04/12/18
Writing thesis	04/21/18	05/31/18
Final presentation	06/26/18	06/27/18

Figure 7: Timeplan



Ziad Mostafa Mohamed Mohamed Thabet <ziad1407174@miuegypt.edu.eg> Aug 17

to youssef1410209, me, Karim, Mai

Greetings Team,

According to the meeting which was held on Tuesday 15th of August, we have agreed on that the graduation project's idea will be on "Translating Lips Movement Into Words".

We have all agreed that there are some points we have to tackle and learn more about it and accordingly the following tasks are assigned on the team members.
1- Narrowing down the scope of the idea : Ziad and Youssef
2- Collect credible papers and find any related works: Ziad and youssef and partially all team members
3- learn all the technical requirements needed. Example: which language to use? Do we need GPU or not? : Amr and Karim.
4- Accordingly to the scope, we have to find our stakeholders.

Please corect me if I am mistaken or anything is ambiguous.
Note: Due to the fact that we are running out of time. We need these tasks to be done by next tuesday.

Best Regards,
Ziad Thabet

Ziad Mostafa Mohamed Mohamed Thabet Sep 14 (11 days ago)

to karim, Youssef, me, Mai

Dear Dr.Mai,

Firstly, we hope everything went well with your talk. CONGRATULATIONS!!

Secondly, kindly note that we held a meeting today to set a plan for this week and we have pinned some points of how are we going to work throughout the year.

1- We have all decided that lips reading is going to be our idea, and this is a final decision.

2- According to the application, we have found that using lips reading on autonomous vehicles is a step we are going to reach anyway; so we have prioritized it as our first application and if we find that we still have time, we are going to enhance our project to be used on surveillience cameras. We would like to know your opinion about this.

3- Since our meetings are usually held on Tuesdays, so we have agreed that every Tuesday is when our week begins and have a meeting to discuss what to be done on the next week, and each monday a report will be sent by each team member stating what he has done and what is missing during the week. We are flexible to change the days. Just tell us whatever suits you.

4- We have recieved an e-mail saying that we have a proposal rehearsal on the 19th of September, thus we are going to finish our first draft by the 18th of September.

5- Knowing that we have to submit our proposal on the 26th of September, we must finish our proposal by the 25th of September.

Tasks for this week:
Proposal is divided on us all:
Ziad: Introduction
Youssef: project Description
Karim: Similar systems
Amr: Project management excluding the plan part

Goodluck with us.

4.2 Budget and Resource Costs

- NVIDIA TITAN X 12000 EGP
- Camera 500 EGP

4.3 Supportive Document

Stakeholder Invitation

2 messages

Ziad Mostafa Mohamed Mohamed Thabet <ziad1407174@miuegypt.edu.eg> Sun, Sep 24, 2017 at 6:25 PM
To: Ismail.zohdy@rta.ae
Cc: karim hani <kariimazmi@gmail.com>, Youssef Samy Mounir Aziz <youssef1410209@miuegypt.edu.eg>, amr1410718@miuegypt.edu.eg, Mai Elshehaly <maya70@vt.edu>

Dear Mr.Ismail,

I am Ziad Thabet, 4th year Computer Science student in Misr International University in Egypt. This year we are working on our graduation project, under the supervision of Dr. Mai El-Shehaly, which is about using "Lip-Reading" on autonomous vehicles.

Our idea mainly is to develop a deep learning appropach for the real-time detection of spoken words in an autonomous vehicle setting. According to our research, we have reached to solve the problem by training a deep-layered CNN (Convolutional neural network) and RNN (Recurrent neural network), and to improve real-time performance by implementing on a GPU.

Meanwhile, we would like to know more about the usages of lip-reading on autonomous vehicles.
Hence, we would honorably like to invite you to be our main stakeholder to our project. Your expertise and experience in the field of work will be an excellent addition to the project.

Best Regards,
Ziad Thabet

Ismail Hisham Zohdy <Ismail.Zohdy@rta.ae> Mon, Sep 25, 20
To: Ziad Mostafa Mohamed Mohamed Thabet <ziad1407174@miuegypt.edu.eg>
Cc: karim hani <kariimazmi@gmail.com>, Youssef Samy Mounir Aziz <youssef1410209@miuegypt.edu.eg>, "amr1410718@miuegypt.edu.eg" <amr1410718@miuegypt.edu.eg>, Mai Elshehaly <maya70@vt.edu>

Dear Ziad and Team,

Thank you for reaching out. It will be my pleasure to participate in this exciting project and be part of your stakeholder team. We can setup a meeting to discuss more your thoughts and plan for the project

Best,

Ismail

Dr. Ismail Hisham Zohdy د.اسماعيل هشام زهدي

Chief Specialist أخصائي رئيسي

Self-driving Vehicles المركبات ذاتية القيادة

CEO Office Licensing Agency مكتب المدير التنفيذي مؤسسة الترخيص

Roads and Transport Authority HQ –Um Rummol – Dubai - UAE

هيئة الطرق والمواصلات
المبنى الرئيس لهيئة الطرق– أم الرمول – دبي – إ.ع.م

M:+971 50 708 8075 ت : +971 4 290 2504 م : +971 50 708 8075

T: +971 4 290 2504

www.rta.ae |ismail.zohdy@rta.ae ismail.zohdy@rta.ae | www.rta.ae

From: Ziad Mostafa Mohamed Mohamed Thabet [mailto:ziad1407174@miuegypt.edu.eg]
Sent: Sunday, September 24, 2017 8:25 PM
To: Ismail Hisham Zohdy <Ismail.Zohdy@rta.ae>
Cc: karim hani <kariimazmi@gmail.com>; Youssef Samy Mounir Aziz <youssef1410209@miuegypt.edu.eg>; amr1410718@miuegypt.edu.eg; Mai Elshehaly <maya70@vt.edu>
Subject: Stakeholder Invitation

[Quoted text hidden]

References

[1] Leon Rothkrantz *"Lip-reading by surveillance cameras"*,13 July 2017 in IEEE

[2] Xinjun Ma, Hongjun Zhang, Yuanyuan Li from Harbin Institute of Technology, Shenzhen, China *"Feature Extraction Method for Lip-reading under Variant Lighting Con- ditions"*,February 24 - 26, 2017 in ACM

[3] Neeru Rathee *"A novel approach for lip Reading based on neural network"*,18 July 2016 in IEEE

[4] Bor-Shing Lin, Yu-Hsien Yao, Ching-Feng Liu *"Development of Novel Lip-reading Recognition Algorithm"*,09 January 2017 in IEEE

[5] Chung, Joon Son, and Andrew Zisserman. *Lip Reading in the Wild.* Springer-Link, Springer, Cham, 20 Nov 2016.

[6] Assael, Yannis M., et al. *LipNet: End-to-End Sentence-Level Lipreading.* , 16 Dec. 2016

[7] Xinjun Ma, Hongjun Zhang, Yuanyuan Li *"A Lip Localization Algorithm under Variant Light Conditions"*,2017 in ACM

[8] Mengchen Liu, Jiaxin Shi, Zhen Li *"Towards Better Analysis of Deep Convolutional Neural Networks"*, 09 August 2016 in IEEE

[9] Sonu Lamba, Neeta Nain, Harendra Chahar *"A Robust Multi-Model Approach for Face Detection in Crowd"*,24 April 2017 in IEEE

[10] Amit Garg , Jonathan Noyola, Sameep Bagadia *"Lip reading using CNN and LSTM"*, 2016 in stanford.edu

[11] Global Political Trends Center Istanbul Kltr University *"SECURITY IN EGYPT"* in April 2015

[12] Association for safe international road travel*"Annual Global Road Crash Statistics"* http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics

[13] Fatemeh Sadat Lesani, Faranak Fotouhi Ghazvini, Rouhollah Dianat *"Mobile phone security using automatic lip reading"*,16 July 2015 in IEEE

[14] Daehyun Lee, Kyungsik Myung *"Read My Lips, Login to the Virtual World"*,30 March 2017 in IEEE

[15] Seksan Mathulaprangsan, Chien-Yao Wang, Aufaclav Zatu Kusum *"A Survey of Visual Lip Reading and Lip-Password Verification"*,23 June 2016 in IEEE

[16] N. Radha, A. Shahina, A. Nayeemulla Khan *"A person identification system combining recognition of face and lip-read passwords"*,18 February 2016 in IEEE

[17] N. Radha, A. Shahina, A. Nayeemulla Khan *"Automated Lip Reading Technique for Password Authentication"*, September 2012 in ijais.org