Tracer: Interactive AI camera man that frees your hands and records your own cinematic footage

Heran Zhang, Shurui Li, Junjie Lu, Chengdong Sun, Zengyang Pan, Wenjia Luo, Kexin Li
Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ
Email: {hz4915, sl10215, jl15315, cs5015, zp815, wjl15, kl2215}@ic.ac.uk
Supervisor: Dr. Yiannis Demiris

Abstract

We are all aware of the explosive growth in popularity of social media globally, and video content has been a major driving force behind the scene. ***Tracer*** *is an interactive cameraman designed to help individuals or groups to film cinematic videos. The design report features the high-level design of* ***Tracer****, as well as the team's approach to evaluate its functionality in order to provide better user experience and produce better results.*

1. Introduction

Since the introduction of social medias like *YouTube*, *Instagram*, *Facebook* and *Twitter*, there has been an extensive growth in use of video contents. Views of branded video content have increased 99% on *YouTube* and 258% on *Facebook* between year 2016 and 2017 [1]. Moreover, a video Tweet is 6 times more likely to be shared than a photo Tweet. Just like *Mark Zuckerberg* said, "I see video as a megatrend".

Moreover, in September 2016, the release of a social media app *TikTok* better approve this growing trend. The app provides a short video platform for people to showcase their creativity. Up till June 2018, in China alone, the daily active users had gone over 150 million and over 300 million monthly active users [2]. In the first quarter of 2018, *TikTok* was officially downloaded by 4.58 million times in *iOS*, overtaking *YouTube* and become the world's most downloaded iPhone app [3]. These statistics defines the leading position of video contents on the social network.

However, there are many occasions when the individual or group require a photo or video being taken from a third-person point of view but there aren't any people around or don't want to bother strangers. Therefore, we introduce *Tracer* as a perfect hypothesized companion for video taking.

In this project, the *Tracer* robot will be equipped with several functionalities to improve user experience and explore its potentials. The robot's performance will be examined with various hypothesis we introduced later in the report.

2. Related Work

There are several self-flying drones in the market that tracks user to perform film recording, the most renowned ones are selfie-drones from *DJI* and *Skydio R1*. However, these drones are only capable of taking wide angle shots and not suitable for close-up shots. In addition, these drones are not the best choice when audio recording is required. In most cases, background music is imported and used rather than raw audio input recorded by the drone as the sound from the propellers will be recorded in. Moreover, when the filming needs to be taken indoor, drones are not preferable, since they require more active space in general and space limitation could also limit its capabilities.

On the other hand, *PhotoBOT* from the class 2017 is another perfect example. It is a robotic photographer on wheels which responsible for taking pictures of people. Furthermore, it can interact with the users via speech and mounted touch screen interface. It shares some similar functionalities with our proposal, such as simple movements on wheels, object avoidance, speech recognition and navigation [6]. However, *PhotoBOT* is only capable of taking photo shot but not video recording and we proposed to use gesture recognition for commanding the robot during shooting mode in order to develop novel features in human interaction.

3. Research Hypotheses

The purpose of this project is to investigate the possibility a robot has, to film video for individuals and groups and a feasible solution to solve the problem when people need a cameraman when they are alone. The robot can be qualified based on below hypotheses:

1. The robot should recognize individuals and tracks them to perform follow up shots, recognize surrounding objects to avoid collisions.
2. The robot should perform different functions upon receiving hand gestures or voice commands.
3. The robot should perform advanced filming techniques like portrait mode or long shot to close up shot automatically by novel algorithm.

Clips produced by the robot should be rated with different individuals and take in as a measurement for whether the robot can be qualified with the role of a dummy cameraman. Methodologies to validate the robot's abilities will be further discussed later in *Experimental Validation* section.

4. System Design

The *Tracer* robot is a complex system integrated with both ROS and multiple AI techniques, it is designed to act as a personal cameraman and supposed to do similar job. It can be fully controlled via voice and gesture, without using any physical controller or app, just like a real human.

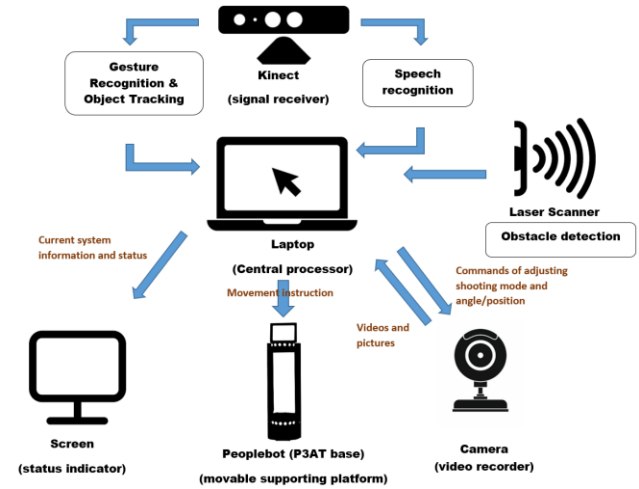

Fig 1: High Level System Overview

Fig 1 shows the high-level system overview for *Tracer*. The system can be broken down into several components where the laptop acts as the central processor where all the data from Kinect and laser sensor are fed into different algorithms and generate outputs i.e. movements for the Peoplebot, angle adjustments for the camera, status display for the screen.

The robot will automatically follow user when powered on. The algorithm will let the robot keep a roughly constant distance from user to ensure the user is at the center of the video while the actual distance is determined by the user's command or the robot's own filming algorithm.

However, voice commands are not suitable during filming since the user's voice commands will be recorded as well. Thus gesture commands are introduced to compensate this problem as using gesture to control the robot is more natural than voice.

ROS (Robot Operating System) will be integrated in the design and development of the Tracer. It's a framework that can ensemble different software modules that allows communications between different nodes and control the operation of robot.

4.1. Software Components

4.1.1 Computer Vision and Gesture Detection

Kinect V1 is used for detecting gestures and motion tracking by exploiting *OpenCV* library. The detected gestures will then be used as user input into the system and the robot will perform actions accordingly. Object recognition algorithm will be invoked to perform human tracking task and implement minor adjustments to shooting directions which enhances the recording result.

4.1.2 Speech Recognition

Speech recognition is vital for the robot to record our speech and transcribe into text accurately and clearly. Speech recognition function can be realized using python's built-in library *SpeechRecognition* that exists on *PyPI* [9]. Voice commands provide a robust and convenient user interface to control the robot when it's not in the filming mode. User can use voice commands to start/stop filming, take picture, change filming mode/angle and adjust distance to user. To avoid disturbing user's regular talking, the voice command function will be turn on only when the user says 'Tracer Tracer'.

4.1.3 Navigation

The autonomous navigation system of the robot mainly consists of two algorithms, the first algorithm is an obstacle detection algorithm, and the other is the object tracking algorithm. *LiDAR* laser scanner and *Kinect* are the receiver for the tasks respectively.

The live video data obtained from *Kinect* will be first converted into 3D point clouds then processed into depth data maps. The data of 3D point clouds contained coordination data in *Cartesian space* i.e. X,Y,Z-axis for the environment allowing shorter processing time [4]. By using the depth value and read the location of the target human from the *Kinect*, the robot can follow the target human as well as comparing the distance between the target human [8].

On the other hand, *LiDAR* scanner can create a 3D representation of the surroundings in the form of point cloud [7]. Thus it can acts as an ideal candidate for obstacle avoidance by assuring the robot to avoid close range objects.

4.1.4 Advanced filming algorithm

Novel algorithm will be designed to analyze the current filming environment based on average image depth, user's voice and body language. The algorithm will use pre-set model to suggest and pick the best filming mode.

4.2. Hardware components

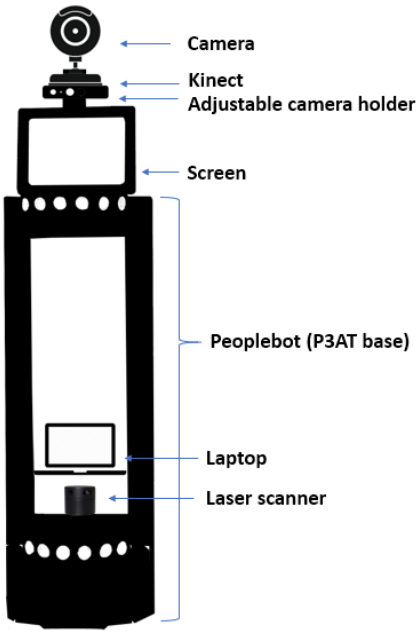

Fig 2: Structure of Tracer

Fig 2. presents the structure of the system and how components are physically placed in the system. The Camera, *Kinect* and screen will be placed above the *Peoplebot*. Laptop will be placed using a laptop holder. Laser scanner will be placed at the bottom platform of the *Peoplebot*.

4.2.1 Kinect (camera and microphone)

Kinect is a peripheral device which not only provides RGB image, but also provides a depth map. In order to detect the users' motion and create a physical image on the touchscreen, the *Kinect* contains an RGB colour VG video camera, a depth sensor and a multi-array microphone [5]. These components can be used to detect and track 48 different points on body. Putting software and hardware together, *Kinect* is able to generate 3D images and recognize users in front of the camera while data is constantly transferring back.

Kinect also has built-in microphone functionality which is the input source for the speech recognition algorithm. If the built-in microphone turns out to be not sufficient to do the speech recognition solely, extra microphone modules will be added to form a microphone array to increase recognition accuracy.

4.2.2 Peoplebot (P3AT base)

Peoplebot can be controlled by either directional keys or a joystick through included software. It can move and follow the users when recording videos.

4.2.3 Screen

Screen acts as information center and is used to display current status of device to users (e.g. whether filming or not, filming mode, whether user's face is recognized, whether obstacle is detected) and some useful information (e.g. distance to user, filming angle, whether user is in the center of the screen).

4.2.4 Filming HD Camera

For a filming robot, video quality is an extremely important factor and should not be compromised. Thus HD camera is used to ensure a good video quality. This component is connected directly to a computer, so the video can be easily stored on the computer.

4.2.5 LiDAR Laser Scanner

LiDAR laser scanner is used for monitoring obstacles around the robot and collision avoidance. The principle of LiDAR is shine a small light at a surface and measure the time it takes to return to its source. Since light speed is constant, the distance can be measured easily [7]. This sensor will co-operate with obstacle detection algorithm using computer vision to provide robust collision avoidance functionality.

4.2.6 Adjustable Camera Holder

Since the HD camera and Kinect camera cannot adjust shooting angle by themselves, an adjustable camera holder is required to adjust the direction and shooting angle of both cameras.

5. Experimental validation

In order to validate the four robot's abilities discussed in the *Research Hypothesis* section, the following methodologies are applied respectively.

Task 1: To validate tracking function of the robot, a relatively large and open-space environment is required, lecture rooms, corridors or outdoor are considered as the optimum choices. A person will stand in front of the robot in the beginning. Once the robot confirms the filming subject and starts film, he/she will begin to walk away from the robot slowly, the success of this experiment is indicated by the movement of the robot in the same direction as the person. In addition, the distance between the robot and person are supposed to be constant. The reaction time of the robot will be recorded each time. At later stage of the project, this experiment will be carried out within an environment contains multiple people, the walking pace will increase as well.

For obstacle avoidance, the corresponding experiment will be carried out in the room only with stationary obstacles at the beginning. Once the robot can

successfully avoid these collisions, moving obstacles such as chairs, human etc. will be used to carry out further tests.

Task 2: Individuals with different figure will demonstrate before the robot, at the same time perform a corresponding hand gesture to convey the adjustment instruction to the robot. The robot is supposed to understand the signal and adjust the filming angle accordingly. This experiment will first be carried out in front of a white wall, under which condition is easier for the robot to recognize the gesture. Experiments in the real environment such as corridor, lab, garden etc. will be carried out based on the success of the previous one.

Task 3: Since video quality is not quantitative, the only methodology that can reflect this is a customer survey about satisfaction with video shot by *Tracer*. The result will be used to make improvements to the filming algorithm.

After the achievement of success on individual experiments, the final experiment is designed to contain all conditions suggest above, which is letting people with different figure demonstrate before the robot in a real environment with multiple walking people each time. The person be filmed will randomly walk and perform hand gestures, the robot is supposed to adjust its camera orientation while following the person and avoid all possible collisions.

6. Conclusion

The initial design of the *Tracer*, which is an interactive cameraman, is presented in this report. This robot aims to investigate to what extent the robot can reproduce the work of cameraman when users are alone. The final specifications and the design details will be discussed in the final report.

7. References

[1] Wyzowl. (2018). Why Video is Exploding on Social Media in 2018. [online] Available at: https://www.wyzowl.com/video-social-media-2018/ [Accessed 29 Oct. 2018].

[2] Sohu.com. (2018). Tiktok user. [online] Available at: http://www.sohu.com/a/235554198_601684 [Accessed 29 Oct. 2018].

[3] Shen, K. (2018). Tik Tok is surging social network ecology — the brand-new interaction among the 00s generations. [online] Medium. Available at: https://medium.com/@kaichenshen/tik-tok-is-surging-social-network-ecology-the-brand-new-interaction-among-the-00s-generations-e7bfa572e829 [Accessed 29 Oct. 2018].

[4] Zainuddin, N., Mustafah, Y., Shawgi, Y. and Rashid, N. (2014). Autonomous Navigation of Mobile Robot Using Kinect Sensor. 2014 International Conference on Computer and Communication Engineering, [online] p.30. Available at: http://irep.iium.edu.my/41594/7/41594.pdf [Accessed 30 Oct. 2018].

[5] Cong, R. and Winters, R. (2018). How It Works: Xbox Kinect. [online] Jameco.com. Available at: https://www.jameco.com/jameco/workshop/howitworks/xboxkinect.html [Accessed 29 Oct. 2018].

[6] Balassa, L., Karandejs, N., Karolcik, S., Matas, J., Olexa, P. and Pulmann, T. (n.d.). Human Centered Robotics Final Report: PhotoBOT. pp.1-4. [Accessed 30 Oct. 2018].

[7] 3D Laser Mapping. (2018). What is LiDAR | What does LiDAR stand for | 3D Laser Mapping. [online] Available at: https://www.3dlasermapping.com/what-is-lidar-and-how-does-it-work/ [Accessed 31 Oct. 2018].

[8] Agarwal, P., Gautam, P., Agarwal, A. and Singh, V. (2017). Human Follower Robot Using Kinect. International Research Journal of Engineering and Technology (IRJET), [online] 04(04), p.1. Available at: https://www.irjet.net/archives/V4/i4/IRJET-V4I4335.pdf [Accessed 31 Oct. 2018].

[9] PyPI. (2018). SpeechRecognition. [online] Available at: https://pypi.org/project/SpeechRecognition/ [Accessed 31 Oct. 2018].