Design and Implementation of Chatbot Framework For Network Security Cameras

Truong Van Cuong
Faculty of Computer Engineering
University of Information Technology,VNU HCM
Ho Chi Minh City, Vietnam
cuongtv@uit.edu.vn

Tran Minh Tan
Faculty of Computer Engineering
University of Information Technology,VNU HCM
Ho Chi Minh City, Vietnam
15520771@gm.uit.edu.vn

***Abstract*— In recent years, with the development of science and technology Internet Protocol camera is getting and getting popular and widely used. In this paper, we present a chatbot framework to help user get the human detection information from the cameras via Facebook messenger instead of observing 24/7 called Security Bot (Sbot). To build Sbot, we design a system including camera network, Human Detection Server (HDS), and Sbot server. In the system, Sbot transfer information between user and HDS using Facebook Messenger Platform. In the human detection task, we use SSD-MobileNetV1 network architecture for detecting human in real-time and updating dataset for retraining using transfer learning method with more case taken from surveillance camera.**

***Keywords*— SSD-MobileNetV1, human detection, Sbot**

I. INTRODUCTION

Currently, home automation is applied widely and sharply. The surveillance camera, especially AI camera, is one of the important devices in smart home system which is getting popular and being used more and more in hospitals, airports, building etc. Appling tradition method, video could be recorded directly on itself, NRV or on cloud storage. As a result, the collected data from camera is numerous. And it is very difficult to extract essential information such as human behavior, license plate, car tracking, people counting and etc.

In addition, chatbot and virtual assistant (VA) are increasingly applied in many fields such as e-commerce, gaming or education sectors. For example: Siri and Google assistant - famous and intelligent VA in the world owned by Apple and Google. In smart home, some branch such as Google Home, Amazon, Xiaomi also introduce some virtual assistant applied in security. Researches about applying chatbot in managing and supervising CCTV system are limited. These researches focus on video search engine. However, these functions have not met the requirements of applications with high request on real-time, immediate information update.

In this paper, we present a framework based human readable. The essential information about the camera, the video and images of detected human will be sent to the user via Facebook messenger. Besides, the user also can communicate and get information about the camera by sending a message.

To communicate between cameras and user, we build two servers, the Human Detection Server (HDS), which processes information collected from the camera, and the Security Bot Server (SBS) to communicate with users. In particular, HDS integrates the function of detecting human, processing images collected through the camera and performing the functions required by users related to cameras. SBS sends messages to users via Facebook's Rest API, and uses socket to transmit information to HDS as text, images or a short video. The system supports: get camera information, video profiles, live streaming, set time for human detection and enable human detection.

In these current years, deep learning methods have been successfully applied in image classification, speech recognition and natural language processing. Researchers have come up with a number of Deep Convolutional Neural Network (DCNN) models such as R-CNN [1], Fast-R-CNN [2] or Faster-R-CNN [3] for object detection, locate and classify. However, these models have limited time for complex and lengthy training, training takes place over several phases, and data predictions over the network do not respond in real time. To solve this problem, in this paper, we present a method of detecting humans using the SSD-MobileNet model, the model is a combination of the Single Shot Multibox Detector (SSD) [4] network and the MobileNet [5] network, that help to speed up processing and real-time responding. Specifically, the architecture of the VGG-16 network will be removed, replaced by the architecture of the MobileNet network and connected to the rest of the original SSD network architecture. Replacing the VGG-16 network with MobileNet can reduce the number of parameters without significant reduction in accuracy.

We use SSD-MobileNet model for human detection. If a human is detected, a short video will be recorded and sent to users with images and URL for live streaming from the camera. In order to reduce errors for detecting human, the group adds the training data set from surveillance cameras in reality with a high angle of view.

II. OUR PROPOSED SYSTEM

A. The topology of proposed system

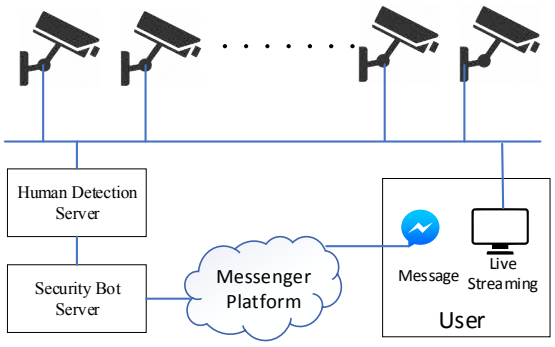


Fig. 1. The topology of proposed system

978-1-7281-0525-3/19/$31.00 ©2019 IEEE

Our proposed system is shown in Fig. 1. The system includes five main components: IP cameras, Human detection server (HDS), Security Bot Server, Messenger Platform, and User. Human detection server streams video from the camera to extract important essential information and sends to the Security Bot Server (SBS) by using Socket. The central unit of the system is the SBS, which transfers and processes information between User and HDS. The SBS text to user via Messenger Platform, which is the open framework provided by Facebook.

B. Human Detection Server (HDS)

In HDS, we implement SSD-MobileNet architecture for human detection algorithm, and send notifications to user when detect any people in the regions of camera. This function is controlled by User via Facebook messenger.

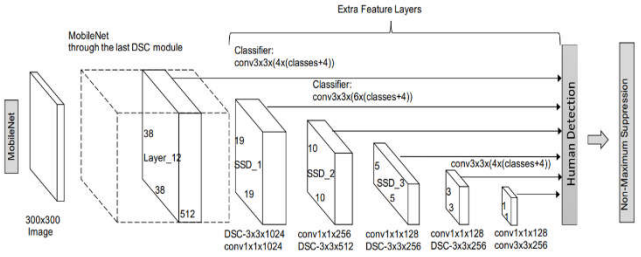


Fig. 2. SSD-MobileNet architecture for human detection, modified based on [4].

The SSD-MobileNet network architecture consists of three main components: the MobileNet network used to extract features, the SSD network to detect a human in different sizes and Non-Maximum Suppression (NMS) [Fig. 2]. Inside, The MobileNet Network is based on the concept of convolutional division in depth and is divided into two phases: a convolution of 3x3, followed by a convolution of 1x1. This reduces the number of parameters in the network, increasing the speed of the network compared to using VGG-16 in the standard SSD model. The SSD network is based on the idea of adding an auxiliary convolutional layer following the base network model to create feature maps of different sizes: 19x19, 10x10, 5x5, 3x3, 1x1. This allows to extract the characteristics of the image at various levels and reduce the size of the input for each subsequent layer, which helps the network to better detect the image in various sizes. And NMS is used to eliminate duplicate predictions on the same object. NMS retains only the predicted limit boxes with the greatest probability, eliminating the lower probability limits and the Intersection Over Union (IoU) values greater than the threshold value.

To improve the accuracy of human detection algorithm in case surveillance camera, firstly, we add more image training collected from many camera in different view and time Fig. 3. After that, we use transfer learning method for training and update our model.

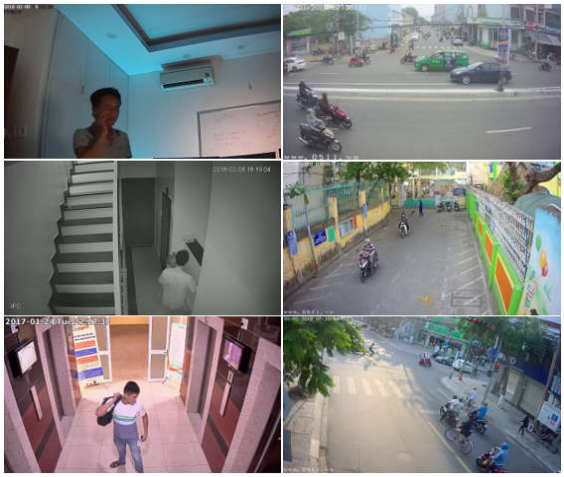


Fig. 3. Example of addition image for training step

C. Security Bot Server (Sbot)

Sbot plays a role as the communication mean between users and HDS, supporting users conducting some main roles such as: camera monitoring, receiving important notification. Via Facebook messenger, users can configure camera, controlling HDS. There are two main components in Sbot: database and messenger platform. Database stores information about users, configuration information, and cameras' status.

The Messenger Platform is the toolbox for building bots. Via Messenger Platform of Facebook, developers can create applications by using the features or existing data on their servers. Its syntax is also simple and easy to integrate with other applications. In Fig. 4, when the user sends a message to the bot (1), Facebook POST message to webhook for processing (2). After the process is complete, if you want to reply to the user, all messages are sent by sending a POST request to the Send API with a page access token. The content of the HTTP request is sent in the JSON format and three properties*: messaging_type* to identify the purpose of the message sent, *recipient* to identify the intended recipient of the message and *message* to define the message to be sent. The Messenger Platform supports sending many types of content in messages, so we can send a text message, image, and video or attached file (3). And then the user will receive the message you reply (4). Moreover, you can avoid being locked in your account when using the API provided by Facebook than using the emulator environment to login Facebook and send messages, increase stability for the application.

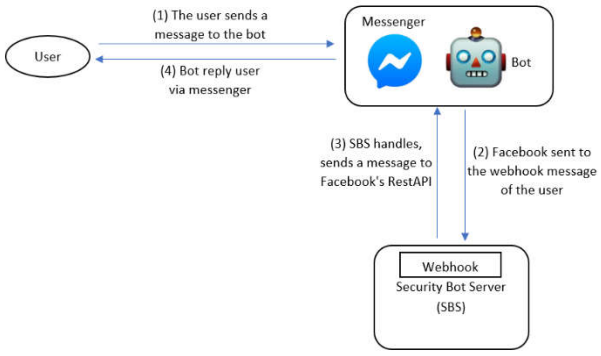


Fig. 4. The mechanism of the chatbot on facebook

TABLE I. COMMAND LINES INTERACTION BETWEEN USER AND SBOT

Command lines	Result returned	Description
Enable human detection	The camera you have assigned successes enable human detection function. Detecting human intrusion!	Enable function human detection, when human is detected, a short video will be recorded and send to the user along with image and URL for live streaming of IP camera.
Disable human detection	The camera you have assigned successes disable human detection function.	Disable function human detection, the user does not receive any information when human is detected.
Camera information	RTSP URL: Resolution: Codec: FPS: Human detection is Enable/Disable.	Get information about the camera such as state of the human detection function, FPS, RTSP URL, codec and resolution.
Capture image/Take a photo	Receive image	The user will be received image capture from the camera after call command.
Capture video	Receive video	The user will be received a short capture from the camera after call command.
Help	We support some functions: 1. Enable human detection. 2. Disable human detection. 3. Capture image / Take a photo. 4. Capture video. 5. Camera information.	The user will be received all command lines that the bot supports.
Other command lines	I don't understand what you mean. You can send 'Help' to get support from us.	The bot will reply to user "I don't understand what you mean. You can send 'Help' to get support from us." when the user input a command line other than above commands lines.

TABLE I describes the interactive commands between the user and the bot. The bot can perform simple functions via command lines sent from the user such as enable human detection, disable human detection, camera information, capture image, and capture video. The mode of operation of command lines is described in the Description column of the table. When the user sends command in the command lines column to the bot, the bot will perform the corresponding function and send the result to the user in the result returned column corresponding to the command line received.

a) Enable human detection: When the user send message "Enable human detection" to bot, message will be sent to SBS via webhook, SBS sends a request enable human detection to HDS. If activated successfully, bot will reply "The camera you have assigned successes enable human detection function.", "Detecting human intrusion!" to user. Then, the IP camera will collect the image, taking it into SSD-MobileNet model. If the human is detected, a short video will be recorded at that time. The result is sent to SBS along with image and URL for live streaming via socket. SBS will send message to Facebook's Rest API. Get the message, Bot will send the video and the URL for live streaming of the IP camera to them user.

b) Disable human detection: When the user send message "Disable human detection" to bot, message will be sent to SBS via webhook, SBS sends a request disable human detection to HDS. If disabled successfully, bot will reply "The camera you have assigned successes disable human detection function.", and the user does not receive any information when human apperance.

c) Camera information: Similar to the way send and receive of "Enable human detection" and "Disable human detection", but the result returned to the user is camera information (FPS, codec, human detectection state, resolution, RTSP URL, HLS URL).

d) Capture image/Take a photo and Capture video: The user will be received a short video capture from the camera after send messege "Capture video" and received a image after send "Capture image" or "Take a photo".

e) Help: The use will be received all command lines that the bot supports.

f) Other command lines: The bot will reply to user " I don't understand what you mean. You can send 'Help' to get support from us." when the user input a command line other than above commands lines.

In addition, if the user registers more than one IP Camera when executing IP Camera related commands, the system will list which IP Cameras the user has registered. User can manipulate any IP Camera just by sending the serial number of the IP Camera that the system listed previously.

III. EXPERIMENTAL RESULT

A. Human detection

We use the data taken from the actual camera to test the model. The total number of frames included in the model is 110,052 with 48,020 images of the indoor environment under full light conditions and 62.032 images of the outdoor environment under daylight conditions. Experimental results show that in the indoor environment, the accuracy of the model was 95.7% and in the outdoor environment the accuracy of the model was 91.2% and the average GPU processing speed on the Intel Core ™ i7-7700HQ computer, Nvidia GTX 1050 Ti reaches 58.5 FPS [Fig. 5].

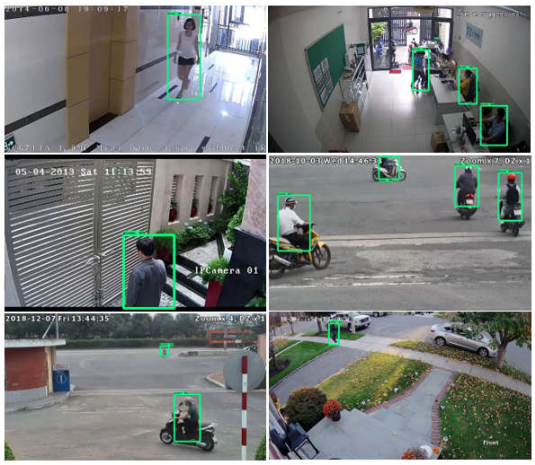
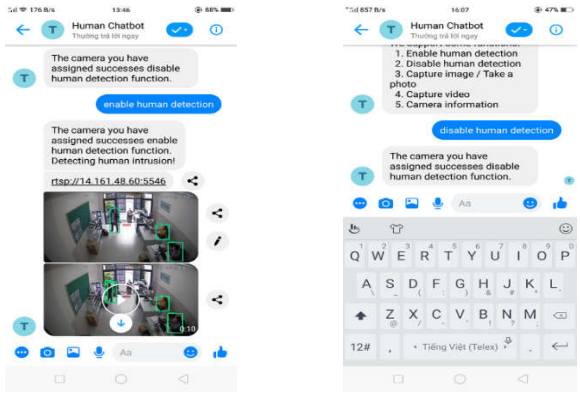


Fig. 5. Human detection result, in outdoor and indoor environment

Fig. 6. The wrong result of SSD-MobileNet model.

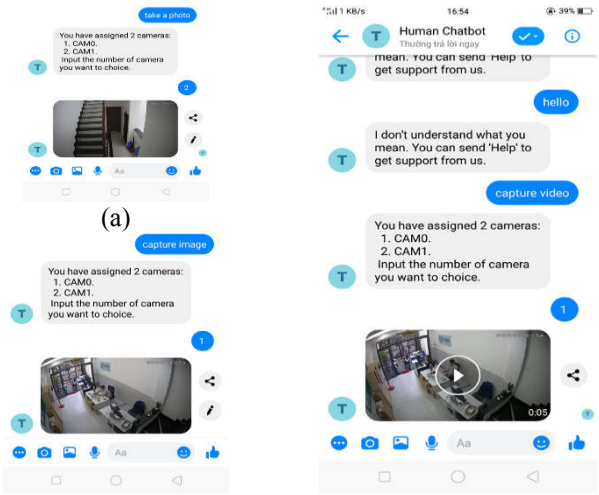
However, the limitation of the model is that it does not work well in night conditions and high false positive rate in the outdoor environment and cannot detect human too far away [Fig. 6].

B. *The interaction between the user and bot*



(a) (b)

Fig. 7. The result of human detection command: (a): enable human detection (b) disable human detection

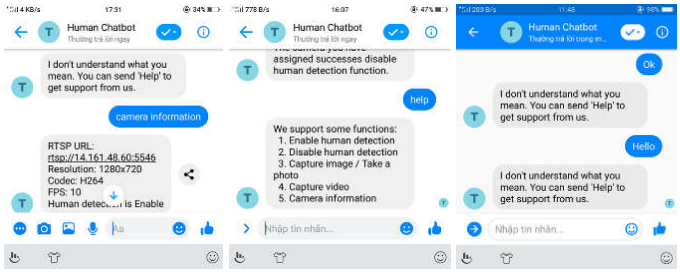


(a) (b) (c)

Fig. 8. Example for capture video/image command: (a), (b), take a photo and capture image are same meaning command, (c) capture a 5 seconds video

To control HDS for human detection task, Fig. 7 show the messages of Human detection task between User and Sbot. In case enable human detection, the system will check which IP cameras that registered by the user is disabled the human detection function. If more than one camera is disabled, the system will send the user a list of ones. A user can manipulate any IP Camera just by sending the serial number of the IP Camera that the system listed previously. Conversely, if none of the cameras are deactivated, the system will announce "None of the cameras you have assigned is disable human detection function.". With disable human detection command, the system will all cameras that registered by the user is enabled the human detection function. If more than one camera is enabled, the system will send the user a list of ones. A user can manipulate any IP Camera just by sending the serial number of the IP Camera that the system listed previously. Conversely, if none of the cameras are activated, the system will announce "None of the cameras you have assigned is enable human detection function.".

In Fig. 8 is example of messages for Capture image/Take a photo and Capture video command: As with similar the above functions, the system will send to the user a list of cameras registered by the user and user can select the camera that they want to capture image or video by sending the camera number in the list received previously. Some of other command is illustrate in the Fig. 9



(a) (b) (c)

Fig. 9: (a) Camera information command, (b) help command, (c) unsupported command and instructions for type right command

IV. Conclusion

In this paper, we propose a method to help the user easily and promptly getting significant information in combination with using traditional methods. Not with standing, in the first release of our framework, there is only feature about human detection which will be analyzed and sent to use through Facebook messenger. In the future work, instead of human detection server, we build a AI server which has more effective functions such human behaviors recognition and so on. Besides, we expect to develop a Chabot which can process natural language and friendly with human. In this research, we propose to use SSD-Mobilenet network for AI application because the next generation of IP camera can process AI algorithm instead of using server.

Acknowledgement

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number **D1-2017-12**

References

[1] J. D. T. D. J. M. Ross Girshick, "Rich feature hierarchies for accurate object detection and semantic segmentation," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580-587.

[2] R. Girshick, "Fast R-CNN," The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448.

[3] K. H. R. G. a. J. S. Shaoqing Ren, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.

[4] D. A. D. E. C. S. S. R. C. Y. F. A. C. B. Wei Liu, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, European Conference on Computer Vision, 2016, pp. 21-37.

[5] A. G. H. M. Z. B. C. D. K. W. W. T. W. M. A. H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.