Multi-Camera Based Object-Retrieving by Integrating Electronic Signals

Tingsong Chen[1], Cheng Zhang[1], Kyoko Yamori[2,3], and Yoshiaki Tanaka[1,3]
[1]Department of Computer Science and Communications Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
[2]Department of Management Information, Asahi University
1851 Hozumi, Mizuho-shi, Gifu, 501-0296 Japan
[3]Global Information and Telecommunication Institute, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
Email : tingsongchen@toki.waseda.jp, cheng.zhang@akane.waseda.jp, kyamori@alice.asahi-u.ac.jp, ytanaka.waseda.jp

Abstract: To ensure the safety of our society, video surveillance systems are widely deployed. However, video camera continually generates a huge amount of data, which makes object-retrieve time-consuming even if by latest pattern recognition algorithms. Nowadays electronic signals, such as Wi-Fi, Radio Frequency IDentification (RFID), are pervasive, and each of them has a unique number which can be taken as the identifier of the target object. In this paper, a new technology is proposed for object-retrieve from video data by integrating electronic signal of the object. We use the best match of Electronic Signal (ES) and Visual Signal (VS) to localize the position of object more accurately, and label the video segment to make object-retrieve faster.

Keywords—**Localization, Retreive, Electric, Visual**

1. Introduction

Nowadays, a lot of video surveillance systems are widely deployed by governments or individuals for public or personal security purpose. When crime happens, these video surveillance systems are helpful for positioning the criminals. However, it always failed to position the criminals in time due to a huge amount of video data to be searched and the high time complexity for pattern recognition.

Intelligent devices are increasing rapidly in recent years. Nowadays, almost every people owns an intelligent device, which is always connected to a cellular Base Station (BS) or WiFi Access Point (AP). Therefore, a people's intelligent device can be assumed as an identifer of the people.

To localize these electronic identifiers, people naturally use wireless technologies to capture and measure wireless signals [1][2][3] . Some technologies do not rely on special hardware or extensive measurement, such as Virtual Compass [4], it combines the blue-tooth and Wi-Fi Received Signal Strength Indicator (RSSI) readings to do relative positioning. However, they are easily affected by the electronic noises. Some technologies are in high accuracy, however, they use expensive hardware during the research, such as Pinpoint [5], this technology relies on time of arrival (TOA) to enable localization. There are also some related works in building a matching system to combine the visual signal (VS) and electronic signal (ES) [1], some use the flash light to track people [6], some use multiple visual signals but not for localization [7].

Our contributions are summarized as follows (Please refer to Figure 1 for the scenario that we consider).

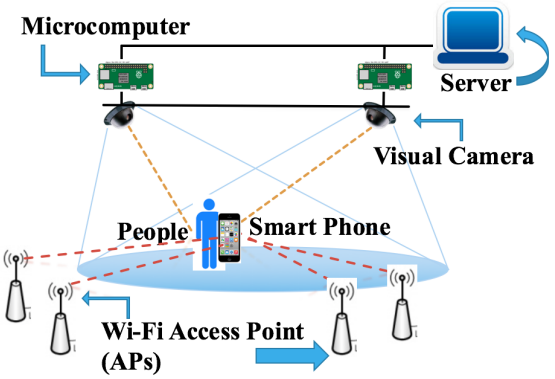

Figure 1. System Model.

(1) We propose to use multiple cameras cooperatively to more accurately localize objects in a much wider range.

(2) We propose to use distributed computing for the system, in which the time-consuming video data processing is done locally.

The rest of this paper is as follows. Section 2 introduces the system design in details. Section 3 mentions the process and results of experiment, and compares the performance in different situations. Section 4 describes the conclusion and future work.

2. System Design

This paper aims to improve the accuracy of object localization and retrieving. Visual localization has high accuracy, but when in dark night or in bad weather, it has low efficiency. Since mobile devices are widely used recently, if the electronic signal from mobile devices can be used along with visual signal, it is expected that the positioning accuracy can be largely improved.

The process is shown in Figure 2, first is data collection. The visual signals are collected by cameras, which record the video stream without stop. The electronic signals, such as International Mobile Station Equipment Identity (IMEI) are collected by APs, which have signal collecting unit inside that can collect RSSI from a mobile device. Next the RSSI signals will be sent to data conversion module that can convert RSSI signals to distances, then the matching system in each microcomputer start to associate the electronic signals to

visual signals. By Hungarian Algorithm, the matching system will output the best match of difference between VS and ES, if there are multiple objects, the system will output when all the differences have a best match. Next the microcomputer will calculate locally and send data to server, finally the positions of target objects will be presented in the coordinate of this area.

At the same time when VS and ES are matched, the identifier will be labeled in the video segment. With the help of the electronic identifier, the retrieve of this object (or person) will be much easier.

The microcomputer includes three functions: VS-ES conversion, error modelling and result sending. Since the signal collected from camera and that from AP are in different units, we need to convert both to distance unit for comparison.

To introduce the ES-VS match engine, we first assume that all the signal collections are finished, no false appears, all the visual signals are distinct, can be identified separately.

Suppose that there are p people and p wireless devices, and it is assumed that each people has one device. The visual location descriptors from two cameras, which describe the coordinate of visual locations, are $\boldsymbol{v}^x = (v_1^x, ..., v_p^x)^T$, $\boldsymbol{v}^y = (v_1^y, ..., v_p^y)^T$, and the electronic location descriptors, which describe the coordinate of electronic locations, are $\boldsymbol{c} = (c_1, ..., c_p)^T$. s_i is a rearrangement of the series (1, 2, ..., p), where $i \in (1, p)$. $v_{s_i}^x = (v_{s_1}^x, v_{s_2}^x, ..., v_{s_p}^1)^T$ is an arrangement of the former $\boldsymbol{v}^x = (v_1^x, v_2^x ..., v_p^x)^T$. Based on these, the formulation of best-match problem is presented:

$$\arg_{s_i} \min \sum_{i=1}^{p} ||c_i - \frac{v_i^x + v_i^y}{2}|| \tag{1}$$

$$m_i = \alpha c_i + \beta(\frac{v_i^x + v_i^y}{2}) \tag{2}$$

The formula (1) is to find an arrangement of $\boldsymbol{v}$ to match $\boldsymbol{c}$, it can be solved with the standard Hungarian Algorithm [8], a combinatorial optimization algorithm to solve the assignment problem in polynomial time. After the arrangement s_i has been founded, we combine the location of $\boldsymbol{c}$ and $\boldsymbol{v}$ to $\boldsymbol{m}$ = $(m_1, ..., m_p)^T$. α and β are coefficients to reflect the confidence of the measurement. If the measurement is not accurate, which is, the standard deviation is big, the weight offered to the location estimate will be decreased, vise versa. At the same time, we define λ_1 and λ_2 as the standard deviation of every $c_i \in \boldsymbol{c}$ and $v_i \in \boldsymbol{v}$, respectively. λ_1 and λ_2 are determined by the equipment used, so they doesn't change during the experiment. The relationship of α, β and λ_1, λ_2 are: $\alpha = \lambda_1^{-2}/(\lambda_1^{-2} + \lambda_2^{-2})$ and $\beta = \lambda_2^{-2}/(\lambda_1^{-2} + \lambda_2^{-2})$.

To achieve these goals, we need to convert all measurements into signal strength or visual distance to perform matching. There is a relationship between signal strength and distance, which is expressed as EM-wave propagation model that introduces the transform between VS and ES. We denote the real distance of the *i*th device and the *j*th AP as d_{ij}. It is assumed that there are n APs in total without electronic interfrences, we denote e_{ij} as the correct RSSI reading of the *i*th device from the *j*th AP, due to [9]. The

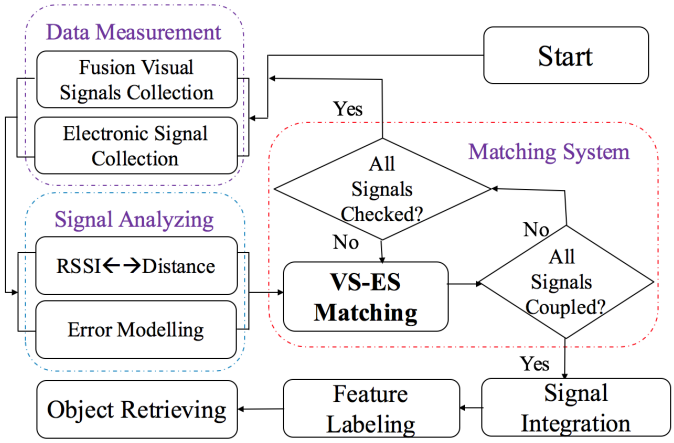

Figure 2. System Flow Chart.

relationship between RSSI signals and distance are shown as follows equation [3].

$$e_{ij} = P_0 - t \ln \frac{d_{ij}}{d_0} \tag{3}$$

P_0 is the original transmission power (known), t is the attenuation coefficient (known), d_0 is a reference distance (known).

We denote d_{ij}^w to the actual distance from visual side, which is calculated by the mean of distances detected by two visual cameras, and denote e_{ij}^l as the actual signal strength from electronic side. And in this experiment, there is no effect from electronic noises. We can get equations like below.

$$d_{ij}^w = d_{ij} + 0 \tag{4}$$

$$e_{ij}^l = e_{ij} + 0 \tag{5}$$

In the notation, d is for distance, w is for visual, e is for RSSI, l is for electronic. Then we can estimate the signal strength e_{ij}^l and distance d_{ij}^w from each other.

$$e_{ij}^w = P_0 - t \ln \frac{d_{ij}^w}{d_0} \tag{6}$$

$$d_{ij}^l = d_0 \cdot \exp(\frac{P_0 - e_{ij}^l}{t}) \tag{7}$$

Next is about the VS-ES match algorithm.

The signals collected by APs and cameras are used in the estimated distance calculation, and we use a cost matrix to represent the similarity between each pair of converted distance from ES and VS. Based on it, the best match which has the highest similarity between each pair will be found out by Hungarian Algorithm, a combinatorial optimization algorithm to solve the assignment problem in polynomial time.

Now it is able to use visual distance as location descriptors, which is, $c_i = (d_{i1}^l, ..., d_{in}^l)$ for the *i*th electronic device and $v_h = (d_{h1}^w, ..., d_{hn}^w)$ for the *h*th visual object(people).

$$\Delta_{ih}^d = ||c_i - v_h||^2 = \sum_{j=1}^{n} (d_{ij}^l - d_{ij}^w)^2 \tag{8}$$

If c_i and v_h are collected from the same device, the distance between them, Δ^e_{ih} or Δ^d_{ih}, should be bounded within a limitation. If Δ^e_{ih} or Δ^d_{ih} is larger than the limit, the matching has high confidence to be wrong.

3. Experiment

This section makes experiment analysis to validate our analytical results. It includes the following aspects:

(1) Compare the localization accuracy of multi-camera and that of single-camera, to verify that multi-camera covers a wider range and provides higher accuracy than single-camera.

With one camera, it has to be settled in the centre of an area, in that case, the localization of object in the edge of that area may have bigger error than object in the centre. Multiple cameras are able to be settled discretely, even the object is in the edge, it can be localized more accurately.

(2) Compare the time complexity of distributed computing case with that of centralized computing case, to verify that while analyzing huge video data, distributed computing provides a higher efficiency than centralized computing.

In the centralized situation, every video signal will be sent to the central server, they may increase the operation pressure of centre server. In the distributed situation, the video signals will be analyzed in local microcomputer, only result will be sent to the central server, it is possible to decrease the operation pressure in central server.

In order to get the visual signal of object, two 720p surveillance cameras are settled on the top of experiment area, shooting from top and covering the same area from different position. The picture of cameras is shown in Fig. 3. The parameters of two cameras are known, such as focal length, lens distortions and so on. Each camera is connected to a microcomputer (Raspberry Pi) which has been installed with Python 2.7-based OpenCV libraries, and using the HoG pedestrian detector [10], the correspondences across frames are established using the detection scheme in [11]. Python 2.7-based OpenCV is used to find out the distance between human and camera, cv2 is for the OpenCV bindings, and an extension called Numpy is imported for numerical processing, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays. We use computer vision to exploit the basic structure of human that is seen from the top, These features can be passed on to machine learning models, which detect and track humans in images and video stream. OpenCV also ships with a pre-trained HOG + Linear SVM model that can be used to perform pedestrian detection in both images and video streams. In this experiment, humans plane coordinates in the area will be calculated by the pixel location of the bottom coordinate system.

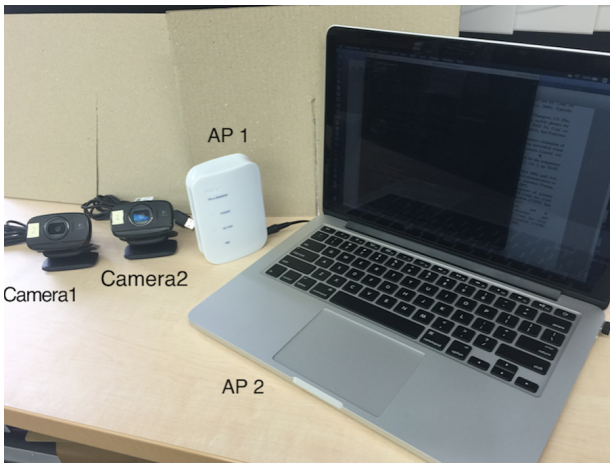

Figure 3. Cameras and APs.

In order to get the electronic signal of device, we put several APs on the ground to detect Wi-Fi MAC address of device. An application named RSSI Reader, which is Android based, is developed to detect RSSI strength between AP and device. Android Application Programming Interface (API) named Wi-FiManager is used to catch the information of RSSI strength. The Wi-Fi switch of the device is always on, so it can be continuously detected.

In the preparation of the experiment, 2 APs are placed in the edge of the area, to capture the electronic signals, to capture the electronic signals, which are shown in Fig. 3. We plan to have 3 people walking on the experiment area, every one has a phone in their pocket. We will take photos of people while they are moving randomly in several time points. Each time when photos are token, the participants stay still. At the same time, the APs collect Wi-Fi signal from the phones for every 100ms.

To compare with one camera situation, another two situations will also be settled, situation A is everything the same with proposed scheme, except two cameras has been changed to one camera in the centre top of the area. Situation B only use the electronic localization. The process of the experiment is totally the same.

Each of the three experiments will collect about 20 frames, followed by the comparison. From the previous work, we already knew that situation B is much worse than situation A. In this experiment, we are planning to prove that the proposed scheme shows better performance than the other two situations. We expect the result as, situation A and B still the same performances as before, proposed scheme will have a faster decrease of error than situation A, and the error of the final frame will be lower than situation A, which proves it is better than the one camera combined system.

The experiments above are based on distributed computing, which is using the microcomputer connected to the camera to analyze the visual signal locally, and send the result data to the centre server to do the calculation. The pisture of them is shown in Fig. 4. To prove this method is faster than directly send the pictures to centre server, another situation C will be considered, situation C has the same configuration with proposed scheme, the only difference between them is in situation C, the two cameras are directly connected to the centre server. The experiment process will also be the same as above.

After the experiment finished, the comparison of

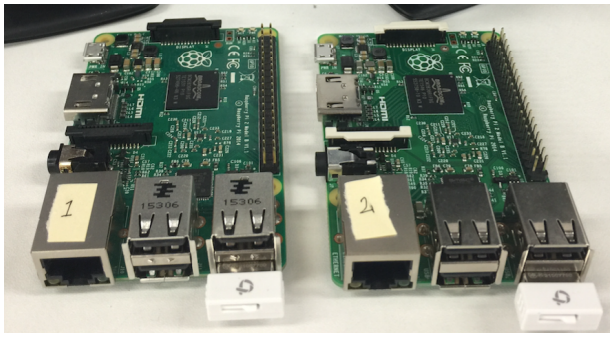

Figure 4. Computers (Raspberry Pi 2) for Distributed Computing.

processing time between the proposed scheme and situation C will be analyzed, and the expected result of them is, the speed of uploading photos are much slower than the data of numbers. Start from the first frame, the time spent by situation C will be much more than proposed scheme, with the increase of photos' number, the speed will be slower, till the last frame, the time spent by situation C will be much more than proposed scheme, which will prove that distributed computing with analysis locally is better than centralized computing.

4. Conclusion

In this paper, a new technology is proposed for object-retrieve from video data by integrating electronic signals of the object. We will use the conversion system to unify different units of VS and ES, to simplify the comparison. Then the matching system based on Hungarian Algorithm is presented. In experiment, on the one hand, we compare the situations of multi-camera and single-camera, and we expect that multi-camera combined system provides higher accuracy than one-camera system. On the other hand, while analyzing huge visual data, we plan to compare the performance of distributed computing and centralized computing. In distributed commuting, visual signal are captured and processed locally in microcomputer and only the result is sent to the centre server, while in centralized computing, all the collected signals are sent to the centre server for data processing. It is expected that the distributed computing provides a faster analysis speed since the process of sending huge visual information to the centre server in centralized computing wastes a lot of time.

In the future, the work will be concentrated on adding more cameras to build a visual network system, to locate not only indoor but also outdoor objects. Nowadays cameras become affordable, and they equipped with high focal length, able to detect in a wider range. We plan to improve this system to adapt more situations such as big square and tall buildings, to approach to the real world.

References

[1] J. Teng, B. Zhang, J. Zhu, X.F. Li, D. Xuan, and Y.F. Zheng, "EV-Loc: Integrating electronic and visual signals for accurate localization," *IEEE/ACM Trans. Netw.,* vol. 22, no. 4, pp. 1285-1296, August 2014.

[2] Z. Liu, M. Dong, B. Gu, C. Zhang, Y. Ji, and Y. Tanaka, "Fast-start Video Delivery in Future Internet Architectures with Intra-domain Caching," *ACM/Springer Mobi. Netw. and App.,* vol.21, no.105, pp. 1-15, February 2016.

[3] Z. Liu, M. Dong, B. Gu, C. Zhang, Y. Ji, and Y. Tanaka, "Inter-Domain Popularity-aware Video Caching in Future Internet Architectures," *Proc. 11th EAI Int. Conf. on Heterogeneous Netw. for Quality, Reliability, Security and Robustness (QSHINE 2015),* Taipei, Taiwan, pp.404-409, August 2015.

[4] N. Banerjee, S. Agarwal, P. Bahl, R. Chandra, A. Wolman, and M. Corner, "Virtual compass: Relative positioning to sense mobile social interactions," *Proc. 8th Int. Conf. Pervasive Comp. (Pervasive 2010),* Helsinki, Finland, pp. 1-21, May 2010.

[5] M. Youssef, A. Youssef, C. Reiger, A. Shankar, and A. Agrawala, "Pin-point: An asynchronous time-based location determination system," *Proc. 4th Int. Conf. Mobi. Sys. App. and Serv. (MobiSys 2006),* Uppsala, Sweden, pp. 165-176, June 2006.

[6] F. Yang, Q. Zhai, G.X. Chen, A.C. Champion, J.D. Zhu, and D. Xuan, "Flash-Loc: Flashing mobile phones for accurate indoor localization," *Proc. IEEE Int. Conf. Comp. Commun. (IEEE INFOCOM 2016),* San Francisco, CA, USA, April 2016.

[7] T. Hatanaka and M. Fujita, "Cooperative estimation of averaged 3-D moving target poses via networked visual motion observer," *IEEE Trans. Automatic Control,* vol. 58, no. 3, pp. 623-638, March 2013.

[8] H.W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.,* vol. 52, no.1, pp. 7-21, February 2005.

[9] S.Y. Seidel and T.S. Rappaport, "914 MHz path loss prediction models for indoor wireless communications in multifloored buildings," *IEEE Trans. Antennas Propag.,* vol. 40, no. 2, pp. 209-217, February 1992.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. IEEE Comp. Soc. Conf. Comp. Vision and Pattern Recognition (CVPR),* San Diego, CA, USA, pp. 886-893, June 2005.

[11] M. Andriluka, S. Roth, and B. Schiele, "People-tracking -by-detection and people-detection-by-tracking," *Proc. IEEE Comp. Soc. Conf. Comp. Vision and Pattern Recognition (CVPR),* Anchorage, AK, USA, pp. 1-8, June 2008.