Demo: Tagging Vision with Smartphone Identities by Vision2Phone Translation

Bryan Bo Cao
Stony Brook University
boccao@cs.stonybrook.edu

Abrar Alali
Old Dominion University
Saudi Electronic University
aalal003@odu.edu

Hansi Liu
Rutgers University
hansiiii@winlab.rutgers.edu

Nicholas Meegan
Rutgers University
njm146@scarletmail.rutgers.edu

Marco Gruteser
Rutgers University
gruteser@winlab.rutgers.edu

Kristin Dana
Rutgers University
kristin.dana@rutgers.edu

Ashwin Ashok
Georgia State University
aashok@gsu.edu

Shubham Jain
Stony Brook University
jain@cs.stonybrook.edu

***Abstract*—We demonstrate our system *ViTag* to associate user identities across cameras' and smartphones' multimodal data. *ViTag* associates a sequence of camera's bounding boxes with smartphones's Inertial Measurement Unit (IMU) data and Wi-Fi Fine Time Measurements (FTM). Our system translates one modality to another by a multimodal LSTM encoder-decoder network (*X-Translator*). Next, an association module finds camera and phone identity correspondences by matching the translated modality with the observed data for the same modality. Our system runs in real-world indoor and outdoor environments, achieving an average Identity Precision Accuracy (IDP) of 88.39% on a 1 to 3 seconds window. We demonstrate our system by visualizing the resulting camera-phone correspondences.**

***Index Terms*—Cross Modal, Fine Time Measurements, Inertial Tracking, Object Tracking, Association, Multimodal Learning**

I. Introduction

With the pervasive use of multimodal sensors of cameras and smartphones, a key application is associating camera detected persons with smartphone's sensor data, as depicted in Figure 1. Real world application includes sending alert messages to associated pedestrians' devices who have been visually detected from camera, distracted pedestrians at risk detected through an infrastructure mounted camera are alerted by voice or vibration on their smartphones and so forth; a particular use case is in facilitating exposure notifications to users in the same scene to prevent potential spread during the current COVID-19 pandemic.

To solve the multimodal association problem, we hereby propose and demonstrate *ViTag*, a system that associates data across camera and phone domains. In particular, vision tracklets are generated by a vision tracker from the camera stream, which are matched with IMU and FTM data obtained from the smartphones.

II. System Architecture

The workflow of *ViTag* consists of two steps shown in Figure 2: (1) cross-modal translation by an encoder-decoder network *X-Translator*; and (2) association by a bipartite matching algorithm. *X-Translator* employs bidirectional LSTM for sequential data extraction, as well as a joint representation layer between vision, motion, and WiFi data from two domains

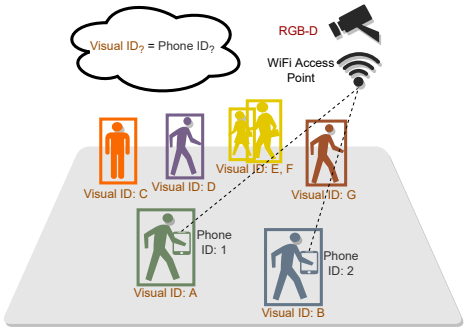

Fig. 1. Motivation: Visually detected subjects and corresponding phone identifiers association by multimodal data.

- camera and smartphone. *X-Translator* leverages the joint representation to reconstruct or translate one modality into the other. In the second step, the reconstructed data (e.g. reconstructed phone data) is matched with the observed data from the same modality.

X-Translator comprises three main modules: (1) an Encoder that learns each input unimodal representation, (2) a joint representation layer that learns the latent features across various modalities, and (3) a Decoder to reconstruct the other modality. *X-Translator* [1] is trained by multimodal reconstruction losses that enforce the network to reconstruct different modalities when not all input modalities are available.

III. Demonstration

A. System Setup

We first introduce our experimental setup as preparation for our sysem demonstration. An RGB-D camera with a WiFi access point device are installed in both indoor and outdoor environments shown in Figure 3. Users walk (in a random manner) with their smartphones in their hands and are captured by the camera. Readings from motion sensors including accelerometer, gyroscope, and magnetometer are captured by each device. Simultaneously, subjects' smartphones exchange FTM messages with the WiFi access point.

[1]Code is available at https://github.com/bryanbocao/vitag. Dataset can be downloaded at https://sites.google.com/winlab.rutgers.edu/vi-fidataset/home.

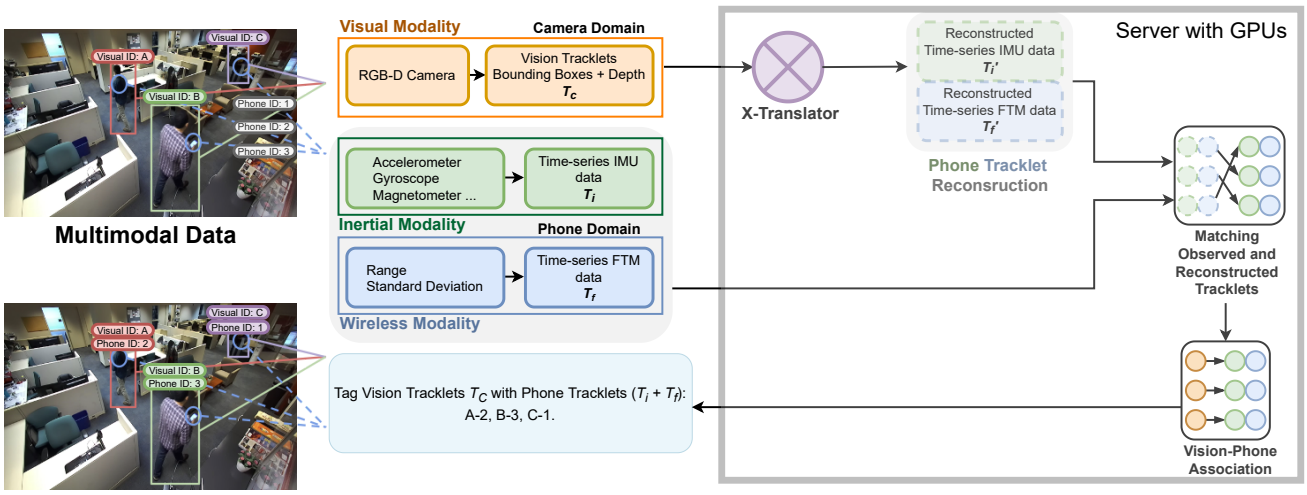

Fig. 2. *ViTag* System Demo Overview. With *X-Translator*, our system first performs data translation from camera to phone domain, followed by a step that finds the correspondences between the reconstructed and observed phone data. We feed vision tracklets (T_c) into *X-Translator* for reconstructing the corresponding IMU (T_i') and FTM (T_f') phone tracklets. We demonstrate *ViTag*'s association capacity by visualizing the outputs of Vision-Phone correspondences in the video stream. Processes that run on the server equipped with GPUs are denoted in the rectangle on the right during demonstration.

Fig. 3. Experiment Setup. In the *Indoor* setup, we mount a Google Nest WiFi AP and a StereoLabs ZED2 camera next to each other on the ceiling, while for outdoor setup, the camera is mounted on the handle of a roof-mounted bike. This setup simulates WiFi-enabled cameras that are becoming increasingly common.

All modalities are synchronized before they are fed into the model. Network Time Protocol (NTP) is used to synchronize the camera and phone data. The sampling rate for camera frames, IMU and FTM readings are 30 fps, 100 Hz, and 3-5 Hz, respectively. Camera data is downsampled to 10 fps and used as an anchor to resample other modalities.

Our model *X-Translator* runs on a server equipped with a NVIDIA GPU RTX 2080 Super. Ubuntu 18.04 LTS as well as the required drivers and frameworks are installed on the server, including CUDA drivers, TensorFlow and Keras.

B. System Demonstration

We demonstrate our system by first visualizing vision tracklets from camera domain. To be specific, subjects are detected and tracked in the camera stream, which is decorated by subjects' vision tracklets in different colors. Each tracklet has a unique ID displayed. Our system assigns subject's phone data with another set of IDs. A script synchronizes and feeds vision tracklets to our pre-trained model *X-Translator* to reconstruct the corresponding IMU and FTM data. Then the system applies maximum bipartite matching (Hungarian Algorithm) to the reconstructed phone data with phone received data for ID matching. Lastly, we demonstrate the Vision-phone ID matching results by displaying phone IDs next to their corresponding vision tracklets shown in Figure 2.

IV. Evaluation

Comparison of *ViTag*'s and baseline methods is summarized in Table I. Our system *ViTag* achieves the highest association performance with an average IDP of **88.39%** in all datasets, exceeding PDR+PA (38.41%) and Vi-Fi (82.93%).

Method	PDR+PA [1], [2], [4]	Vi-Fi [3]	**ViTag (Ours)**
Avg. IDP	38.41%	82.93%	**88.39%**

TABLE I
Summary of Online Association Performance in All Datasets

V. Conclusion

This work demonstrates a practical solution to cross-modal association. We designed and demonstrated *ViTag* to associate pedestrians visually detected from a camera stream with corresponding smartphone data in real world scenarios.

VI. Acknowledgement

This research has been supported by the National Science Foundation (NSF) under Grant Nos. CNS-2055520, CNS-1901355, CNS-1901133.

References

[1] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.
[2] W. Krzanowski. *Principles of multivariate analysis*, volume 23. OUP Oxford, 2000.
[3] H. Liu, A. Alali, M. Ibrahim, B. B. Cao, N. Meegan, H. Li, M. Gruteser, S. Jain, K. Dana, A. Ashok, et al. Vi-fi: Associating moving subjects across vision and wireless sensors.
[4] B. Wang, X. Liu, B. Yu, R. Jia, and X. Gan. Pedestrian dead reckoning based on motion mode recognition using a smartphone. *Sensors*, 18(6):1811, 2018.