Poster Abstract: Towards Self-supervised Face Labeling via Cross-modality Association

Chris Xiaoxuan Lu
Department of Computer Science
University of Oxford

Xuan Kan
College of Software Engineering
Tongji University

Stefano Rosa
Department of Computer Science
University of Oxford

Bowen Du
Hongkai Wen
Department of Computer Science
University of Warwick

Andrew Markham
Niki Trigoni
Department of Computer Science
University of Oxford

ABSTRACT

Face recognition has become the de facto authentication solution in a broad spectrum of applications, from smart buildings, to industrial monitoring and security services. However, in many of those real-world scenarios, tracking or identifying people with facial recognition is extremely challenging due to the variations in the environment such as lighting conditions, camera viewing angles and subject motion. For most of the state-of-the-art face recognition systems, they need to be trained on a large dataset containing a good variety of labelled face images to work well. However, collecting and manually labelling such datasets is difficult and time consuming, probably more so than developing the algorithms. In this paper, we propose a novel framework to automatically label user identities with their face images in smart spaces, exploiting the fact that the users tend to carry their smart devices while seen by the surveillance cameras. We evaluate our method on 10 users in a smart building setting, and the experimental results show that our method can achieve > 0.9 f_1 score on average.

CCS CONCEPTS

• **Human-centered computing** → Ubiquitous computing;

KEYWORDS

Camera; WiFi; Cross-modality Association

ACM Reference Format:
Chris Xiaoxuan Lu, Xuan Kan, Stefano Rosa, Bowen Du, Hongkai Wen, Andrew Markham, and Niki Trigoni. 2017. Poster Abstract: Towards Self-supervised Face Labeling via Cross-modality Association. In *Proceedings of 15th ACM Conference on Embedded Networked Sensor Systems (SenSys'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3131672.3136991

SenSys'17, November 6–8, 2017, Delft, The Netherlands

ACM ISBN 978-1-4503-5459-2/17/11…$15.00
https://doi.org/10.1145/3131672.3136991

1 INTRODUCTION

Biometric-based authentication has emerged as one of the most promising options for recognizing individuals in recent years. Compared to other biometric modalities, face recognition has a number of strengths and hence has raised significant interest from various industries. Recent projections have the face recognition market valued at $2.19 billions by 2019 [1].

Face recognition has been studied extensively in the computer vision community, yet significant challenges still stand, especially when deployed in the wild. The performance of facial recognition algorithms is significantly impacted by uncontrolled lighting conditions, large pose variations and partial occlusions. State-of-the-art face representation approaches, like FaceNet [3], are able to learn good embeddings and robustly recognize faces drawn from environments similar to the training set, but generalize poorly in the wild, where facial images are captured without users' active participation. For instance, Fig. 1 shows a failed example of recognizing a user's face in surveillance video, where FaceNet falsely returns a smaller distance between two distinct faces rather than between two instances of the same face from different domains (surveillance video v.s. facebook albums). How to source more labeled face images in the target domain to fine-tune the original model is the biggest challenge for face recognition that restricts its wider deployment in smart spaces.

We note that in-situ identity labels linked to face images are always available, but in a hidden manner to smart space administrators. Besides cameras, various sensor modalities co-exist in the smart space. Some of these sensors are able to capture the emitted identity information from the users' devices. For instance, the wireless traffic in the environment conveys MAC addresses which are uniquely mapped to device owners. Such information can be easily collected with a WiFi router, and the mapping from MAC address to a particular user is in the database open to administrators of smart spaces.

In this paper, we propose a novel framework to associate uniquely identity labels to face images in smart spaces, which is inspired by recent work [2] on acoustic-based user identification. The labeling process is self-supervised and free of human intervention. The key challenge to overcome is the fact that the WiFi labels are noisy and do not always guarantee a one-to-one mapping. We thus cast this to an association problem.

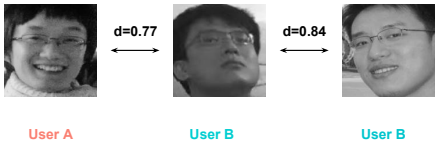

Figure 1: A failed example of face recognition when using the face images of the same user (user B) but from different domains (surveillance video v.s. facebook albums). It turns out the faces of two distinct users are predicted to be more similar. The distances d is calculated by FaceNet [3].

2 PROPOSED APPROACH

The key idea of self-supervised labeling is associating unique labels from an identifying sensor (e.g., a WiFi sniffer) to facial images captured by a camera. The identifying sensor and camera co-exist in the same physical location and therefore share the same context.

Identifying Sensor Collection: The identification sensor observations are harvested from the ambient WiFi traffic. Through a WiFi sniffer, the wireless network traffic generated by a user's mobile device can be intercepted and logged. A user's identity is uniquely mapped to the MAC address of his carried devices and RSS can be used to determine user's context of location. We therefore employ the MAC address as the identity label I and collect its observations across different contexts. We then assign to each identity label I_m a membership vector E^I_m. $E^I_m(j)$ is set to 1 only if its respective MAC address has been captured in the j-th context.

Face Extraction: To retrieve all faces in the video, the first step is to detect faces in each frame. We use MTCNN [4], which jointly learns face detection and alignment using Cascaded Convolutional Networks. Each face image captured by MTCNN is then cropped and a descriptor computed for it. Next, we employ the FaceNet [3] descriptor, which adopts a triplet-based loss and is able to map from face images to a compact Euclidean space, where distances directly correspond to a measure of face similarity. Once this space has been produced, clustering can be easily implemented using standard techniques with FaceNet embeddings as feature vectors.

Face Clustering: Based on the extracted feature vectors, the second step is to group faces into a set of clusters F depending on how similar they look. We use agglomerative clustering to group all the face images recorded across all contexts, and the number of clusters is determined by the number of distinct MAC addresses found in WiFi observations, and is equal to 5 in Fig 2. After this step, we assign a membership vector E^F_k to the k-th final cluster F_k; $E^F_k(j)$ is set to 1 only if this cluster contains face images from j-th context. The set of contexts contributing to cluster F_k can be represented as: $C_{F_k} == \{j | E^F_k(j) = 1\}$.

Cross-modality Association: The last step is to associate biometric observations to identity observations, which completes the face labeling. We formulate this association task as a bipartite graph. Formally, for a given identity label I_m, an edge is created between I_m and F_k if I_m is observed to appear in any context within C_{F_k}, according to the intersection between biometric membership vector E^F_k and identity membership vector E^I_m. The weight of this edge is determined by the number of contexts in C_{F_k} that I_m has participated in. Then associating identities with final clusters is equivalent to solving a combinatorial optimization problem on the weighted bipartite graph, e.g. using the Hungarian algorithm. Fig. 2 illustrates the above association problem, when we are interested in 3 user identities, participating in 4 different contexts, in which a total of 5 distinct user identities (Wifi addresses) have participated.

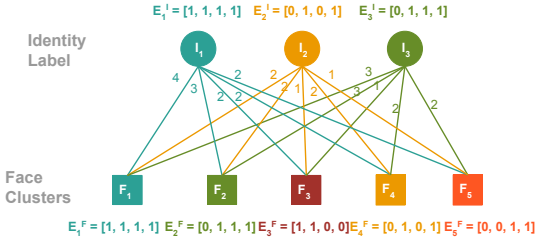

Figure 2: Illustration of the association problem formulation via bipartite graph.

3 PRELIMINARY RESULTS

In our experiments a total of 10 people participated in the experiments, for a total of 10 meeting events. Each user carried a WiFi-enabled mobile device. A camera was placed in the meeting room, facing the entrance door. Each meeting involved two to four participants, and a total of 28,697 face samples (2k samples per user on average) were detected and associated to 10 identities.

The average precision, recall and f_1 scores were 0.92, 0.89 and 0.95 respectively. Except for one user (user 6), whose face images are completely incorrectly associated (due to the inherent representation limits of FaceNet), the performance on other users' images is very competitive.

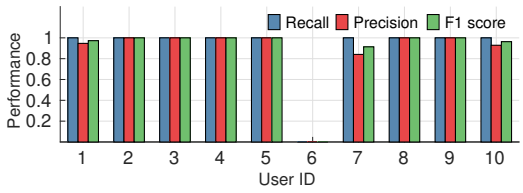

Figure 3: Association results for different users.

In summary, the technique we have presented here can allow users to be tracked robustly inside a smart space through the use of pervasive infrastructures such as CCTV. To increase accuracy, we are investigating modifications to FaceNet to improve discrimination ability over these large datasets.

REFERENCES

[1] Ian Koskela. 2015. Top 5 Applications of Facial Recognition. http://www.biometricupdate.com/201503/2014-year-in-review-top-5-applications-of-facial-recognition. (2015).
[2] Chris Xiaoxuan Lu, Hongkai Wen, Sen Wang, Andrew Markham, and Niki Trigoni. 2017. SCAN: learning speaker identity From noisy sensor data. In *IPSN*.
[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.
[4] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* (2016).