Technical Disclosure Commons

Defensive Publications Series

October 16, 2018

GENERATING TRAINING DATABASES USED IN VECTOR BASED OBJECT RECOGNITION IN HYBRID CLOUD USING PUBLIC PROFILES

Ashutosh Malegaonkar

Ming'En Zheng

Haihua Xiao

Rizhi Chen

Li Kang

See next page for additional authors

Inventor(s)
Ashutosh Malegaonkar, Ming'En Zheng, Haihua Xiao, Rizhi Chen, Li Kang, and Stacy Ling

GENERATING TRAINING DATABASES USED IN VECTOR BASED OBJECT RECOGNITION IN HYBRID CLOUD USING PUBLIC PROFILES

AUTHORS:
Ashutosh Malegaonkar
Ming'En Zheng
Haihua Xiao
Rizhi Chen
Li Kang
Stacy Ling

ABSTRACT

Techniques are provided herein for generating a face data set which contains badge identifier photos and photos from social media. The faces are automatically tagged using facial recognition, text recognition, and human relationship mining.

DETAILED DESCRIPTION

In the last decade, machine learning has made great progress. One important factor is that a large amount of training data is available. But collecting, cleaning, and labeling this data requires an inordinate amount of human effort. Accordingly, a semi-automated method is provided to generate a large data set for training a facial recognition model in a network and cloud environment.

When using a collection of corporate identifier (ID) photos to train a facial recognition algorithm, the data set can be enhanced based on knowledge of the identity of the person in the ID photo. Knowledge of the person's identity enables acquiring additional photos from social media, intranet, and internet searches that are likely to be different views of the same person.

The acquired photos may be tagged as being of the known subject, thereby providing the machine learning algorithm a larger training dataset from which to learn.

Figure 1 illustrates five steps in the data set collection process.

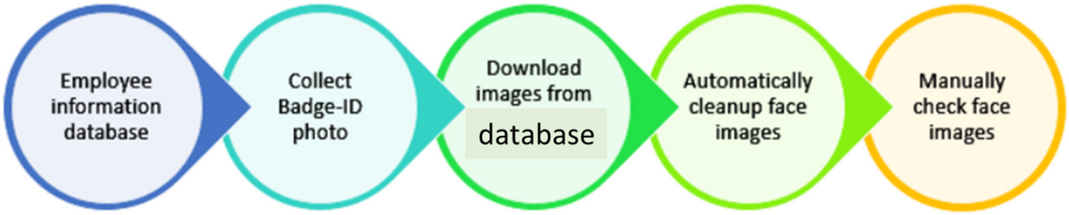
Employee information database
Collect Badge-ID photo
Download images from database
Automatically cleanup face images
Manually check face images

Figure 1: Workflow for generating large face data set

In the first step, an employee information database is created. The employee information may be obtained from the Information Technology (IT) department or the intranet. The information may include name, department, title of each employee, etc. stored in the database.

In the second step, the badge ID photo is collected for every employee. In general, the company takes a photo of each employee for the badge or employee directory. These photos are taken in a controlled environment (e.g., frontal face, neutral expression, controlled illumination, etc.). These photos are referred to as Badge-ID photos. Badge-ID photos may be obtained from the IT department or the intranet.

Badge-ID photos are important in many facial recognition use cases. In video surveillance, there is only one ID photo of a criminal that is available, and the facial recognition system needs to compare the ID photo with the faces captured by the surveillance camera. However, Badge-ID and ID photos are taken in controlled illumination environments, and the faces are frontward-facing and have neutral expressions. Thus, the data set consisting of Badge-ID photos and social media photos is closer to practical use cases, and can help train an algorithm with better generalization ability.

In the third step, images of every employee are downloaded from the Internet. In order to obtain more images which relate to the employee, multiple keywords are used to search for employee photos, including employee name, employee name and company name, employee name and title, etc.

Figure 2 below illustrates how at step four, the data set is automatically cleaned up.

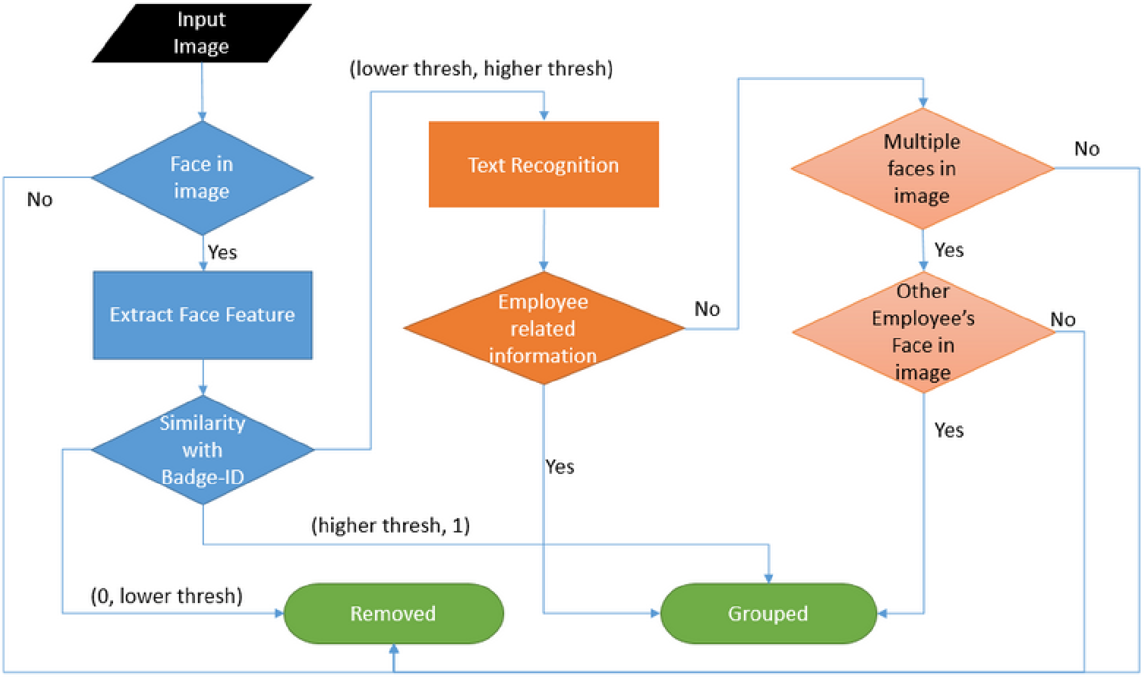

Figure 2: Workflow for automatically cleaning up data set

At the third step, dozens of images were obtained for each employee. It may be desirable to clean up the image data set, including removing the images having no face or without the employee's face.

The dlib face detector may be applied to detect the faces in the data set. If there are no faces in an image, that image may be removed.

Many photos collected from social media may not contain the corresponding employee's face. The faces that do not belong to the employee may be filtered out. In recent years, facial recognition technology has made great progress. Pre-trained facial recognition may be used to compare the Badge-ID photo with the photos downloaded from the Internet. There are two thresholds used in this method: lower threshold and higher threshold. If the similarity is smaller than lower threshold, then that face may be filtered out (i.e., the face is determined not to belong to that employee). If the similarity is greater than the higher threshold, then that face is determined to belong to that employee. If the similarity is between lower threshold and higher threshold, it is still not certain that the face belongs to the employee, and this image may be reserved for further processing.

In many cases, characters in an image are a descriptor of image content. If the employee's related information appears in the picture (e.g., employee name or corporate name) there may be more confidence that the face in the image belongs to that employee. As illustrated in Figure 3 below, if the facial recognition model does not recognize the face in the picture, but the "COMPANY" and "EMPLOYEE NAME" characters are extracted via a text recognition model, then this image may be reserved for further processing.

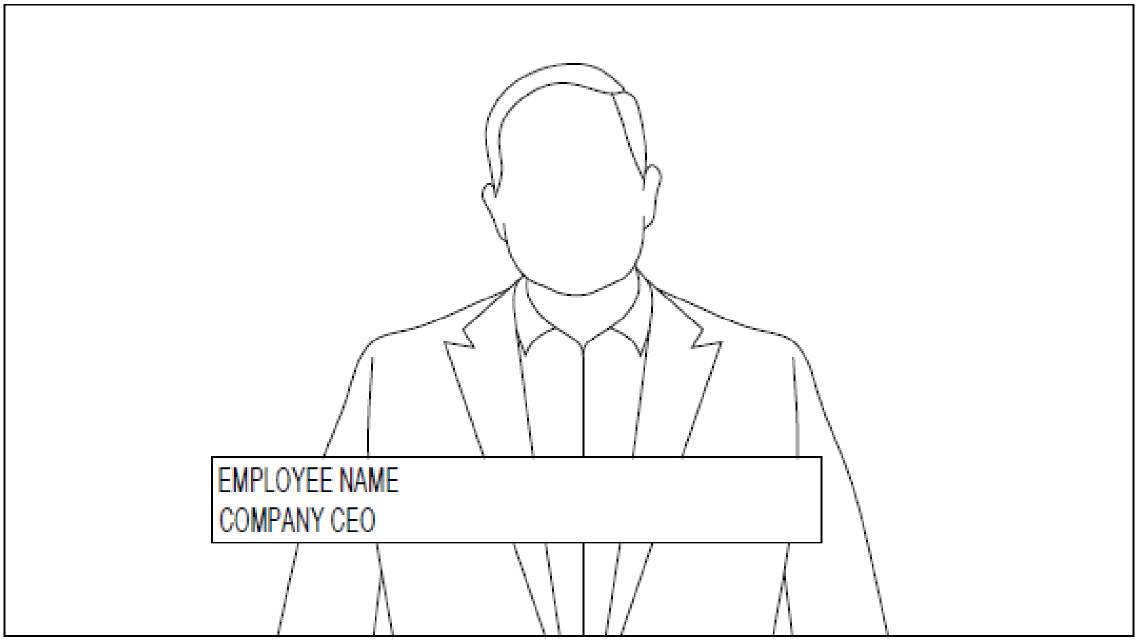

Figure 3: Text recognition to extract context information

In addition, there are many group photos on social media. If it is known that one of the people in a group photo is an employee of the company, then the other person in the group may also be the employee of this company. As illustrated in Figure 4 below, if the employee from Figure 3 is already known, then the other people in shared photos may have a high probability of being employees of the same company.

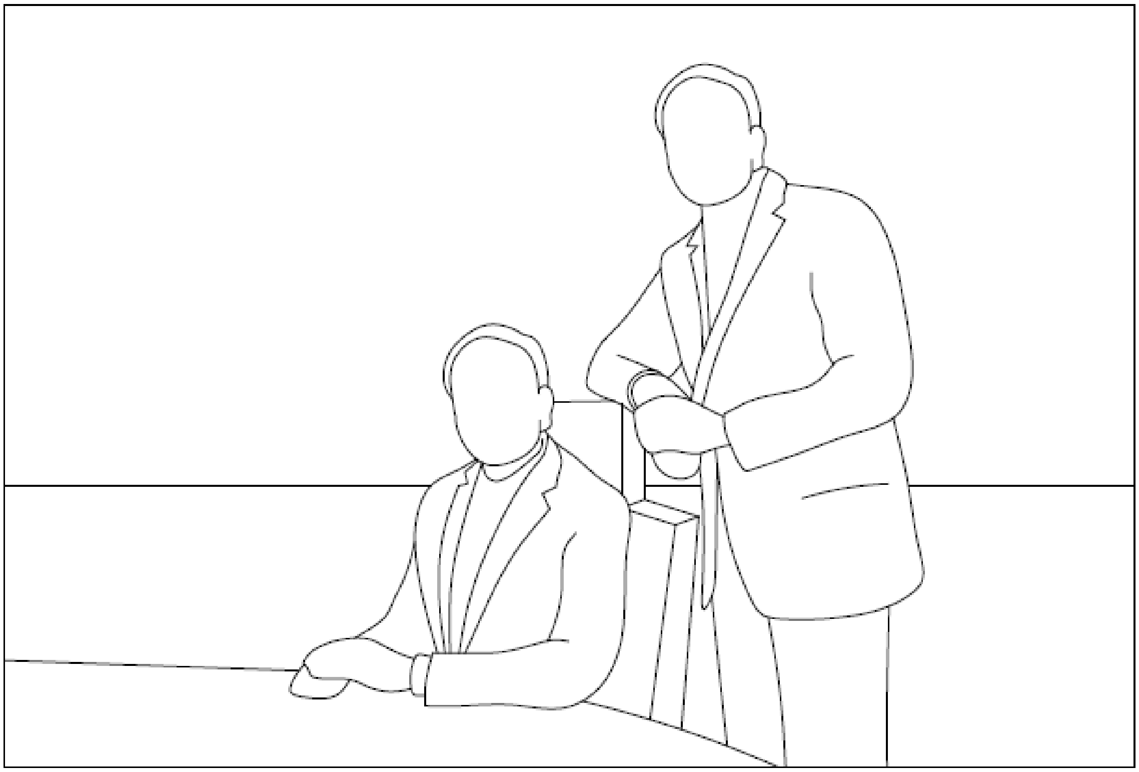

Figure 4: Human relationship

At step five, the data set is manually cleaned up. After the previous steps, there are still some faces downloaded from the Internet that do not belong to the employee. In order to improve the purity of the dataset, the data set may be manually checked one by one, to make sure the faces have the correct tag.

Described herein is a method to generate a facial image data set for training a facial recognition model. Many of facial image data sets are available. In this data set, each individual has several distinct photos. The ID photo may not be treated as special. The ID photo is taken in a constrained environment (e.g., uniform illumination, indoor, frontal view, neutral facial expression, and not occluded).

To build a machine learning model for future prediction, it may be assumed that unseen (test) data comes from the same distribution as the training data. When the distribution on training and test sets do not match, a covariate shift is present.

When a covariate shift issue is present, the learned model performs well on the training data set, but performs poorly on the test data set. In video surveillance applications, one ID photo of a suspect is present along with several surveillance images, and the ID photo and surveillance image are compared for similarity. The ID photo is captured in a constrained environment, while the surveillance images are taken in unconstrained environments and are subject to variations in illumination, pose, facial expression, etc. In order to avoid the covariate shift issue, it is better to have the training examples and testing examples in the same distributions. Therefore, in the data set, each individual has an ID photo and several daily life photos. The distribution of the training data set is close to the testing data set. This data set is not only used for training the facial recognition model in video surveillance, but also may be used in every ID photo and several unconstrained photo scenarios, such as telepresence applications.

Even though there are many photos, there may not be exactly one ID photo for every individual. The photos are uploaded by users. There is no other data set that tags the ID photos as special.

It is complex to generate a face data set for training facial recognition models. All of the examples in the training data set should be labeled correctly. If some examples are not labeled correctly, then there will be noise in the data set. A model trained on a noisy data set will not perform well. The facial recognition technique cannot tag every example correctly. As such, different information needs to be combined together. A pipeline for tagging photos is presented (e.g., as illustrated in Figure 2).

In this data set, there is one ID photo and several photos collected from the Internet. The data distribution in this data set is close to several real applications, such as video surveillance and telepresence. There are many efforts to clean up the photos downloaded from the Internet. The corporate photos are not enough for training a facial recognition model. A pipeline is described to clean up the images download from the Internet.

In summary, techniques are provided herein for generating a face data set which contains badge identifier photos and photos from social media. The faces are automatically tagged using facial recognition, text recognition, and human relationship mining.