Face Verification and Open-Set Identification for Real-Time Video Applications

Jakob Grundström
pi07jg8@student.lth.se
Supervisor: Kalle Åström
kalle@maths.lth.se
Assisting Supervisors: Martin Ljungqvist, Jiandan Chen
{martin.ljungqvist,jiandan.chen}@axis.com

1 Introduction

As a popular topic in computer vision face recognition has been the subject of significant research efforts during the past several years. Focus of much of this research has been on how to address challenges such as changes in illumination, variation in head pose or facial expression, occlusion, translation and image quality related issues. This is typically done by designing or learning invariant features for the face representation and by designing robust face recognition algorithms.

Face recognition has many important applications in video analysis, in for example access control and person identification, and can constitute a key component in person re-identification systems. The primary advantage of face recognition compared to other biometrics is its non-intrusive nature and that it does not require active cooperation.

Two main tasks in face recognition are *verification* and *identification*. The purpose of *face verification* is to confirm or deny a claimed identity. In a pair-wise formulation the problem is to decide if two face images are of matching or non-matching identities. Identification concerns correctly identifying face images. Identification considering the possibility of impostors is called *open-set identification*.

The aim of this work was to develop a video-based face verification algorithm suitable for use in an embedded environment and to investigate face verification-based solutions to real-time video applications requiring open-set face identification. In particular, we built a prototype system for keeping track of the identities of persons currently inside a *closed area*. Essentially, this is done by matching identities between two separate video streams, verifying face images captured at the entry with images captured at the exit.

2 Description

A pair-wise face verification algorithm consist of feature extraction and a classifier. Given two face images it produces a similarity score for the identities of the face images. The similarity score can then be thresholded to reach a final decision if the two face images are of matching or non-matching identities.

2.1 Features

The aim of the feature extraction is to create a compact and discriminative representation for the identities in face images. We primarily investigated two feature types:

- Local Binary Pattern (LBP) (Ojala et al., 2002) features extracted from image regions around facial landmark points in multiple scales (method used in the closed area application). This feature extraction is roughly a reduced, light-weight version of (Chen et al., 2013) but with changed LBP operator radius to achieve scale invariance instead of using scale-space pyramids.

- Deep feature representations extracted from the last hidden layer activations of Convolutional Neural Networks (CNN). The CNNs are *transfer learned* from generic object recognition by using pre-trained weights from (Jia et al., 2014) and then fine-tuning for face identification with face image data. This is similar to how (Karayev et al., 2013) fine-tuned a generic CNN for Flickr style categorization.

2.2 Classification

To solve the face verification problem a binary decision if two face images are of matching or non-matching identities needs to be made. For this purpose we used a classifier, the *Joint Bayesian* (Chen et al., 2012). It defines a probabilistic similarity measure by modeling faces as the sum of two independent multi-variate Gaussian stochastic variables

$$x = \mu + \epsilon \tag{1}$$

and creates the log-likelihood ratio of the probability of a pair assuming matching identities and the probability of a pair assuming different identities

$$r(x_1, x_2) = \log \frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)}. \tag{2}$$

We compare the accuracy of this classifier to the accuracies of *Random forests*, *Boosted regression trees* and *Radial basis function support vector machines*.

2.3 Closed Area Application

With the video-based face recognition pipeline shown in Figure 2 faces are detected from individual video frames, turned into a sequence of feature vectors representing a face track of a single person and matched against a dynamic database. In the closed area application (see Figure 3) one instance of this pipeline runs for each video stream. Figure 1 shows a screenshot of the closed area application.

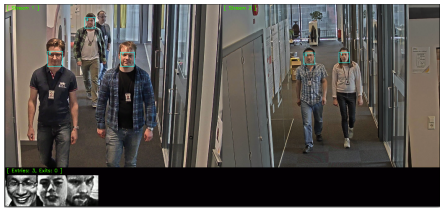

Figure 1: Screenshot of the closed area application showing the video streams from the entrance (left) and the exit (right) to a closed area. Persons currently inside the closed area are shown in the gallery (bottom).

The face recognition pipeline used includes:

- **Face alignment** A similarity transform is fitted between detected facial landmark points and a set of reference points, and is used to rectify the face images produced by the face detector.
- **Face tracking** groups face detections into temporally connected face tracks for each individual appearing in the video streams. The underlying tracking algorithm is optical flow-based tracking of point features.
- **Best-shot selection** is performed to reduce the number of images representing a track and keeping the ones with best quality. In a captured face track often a proportion of the images were of low quality due to motion blur, bad focus or low resolution.

3 Results and Discussion

We considered a pair-wise face verification setup that makes the binary decision if two face images are of matching or non-matching identities. The Joint Bayesian classifier was found useful for this purpose and we showed that it compares favorably with a set of standard classifiers. Face verification was evaluated using two distinct feature types: local feature representations around landmark points and deep representations extracted from Convolutional Neural Networks (convnets) trained for generic object detection and fine-tuned on face image data. With these face verification algorithms we implemented and evaluated open-set identification.

Combined with a Joint Bayesian classifier the deep representations showed good accuracy applied to face verification; we reached a face verification accuracy of 91.37% on the popular *Labeled Faces in the Wild* benchmark. The feature extraction with the deep representations was computationally demanding but may become applicable to embedded environments in the future. The local feature representations investigated required less computational resources but gave lower verification accuracy, 86.4% on the Labeled Faces in the Wild benchmark.

We identified local features as the currently most applicable feature representation for the considered real-time video applications in the immediate future. For this reason a local feature scheme based on Local Binary Patterns (LBP) was used in our prototype system. For

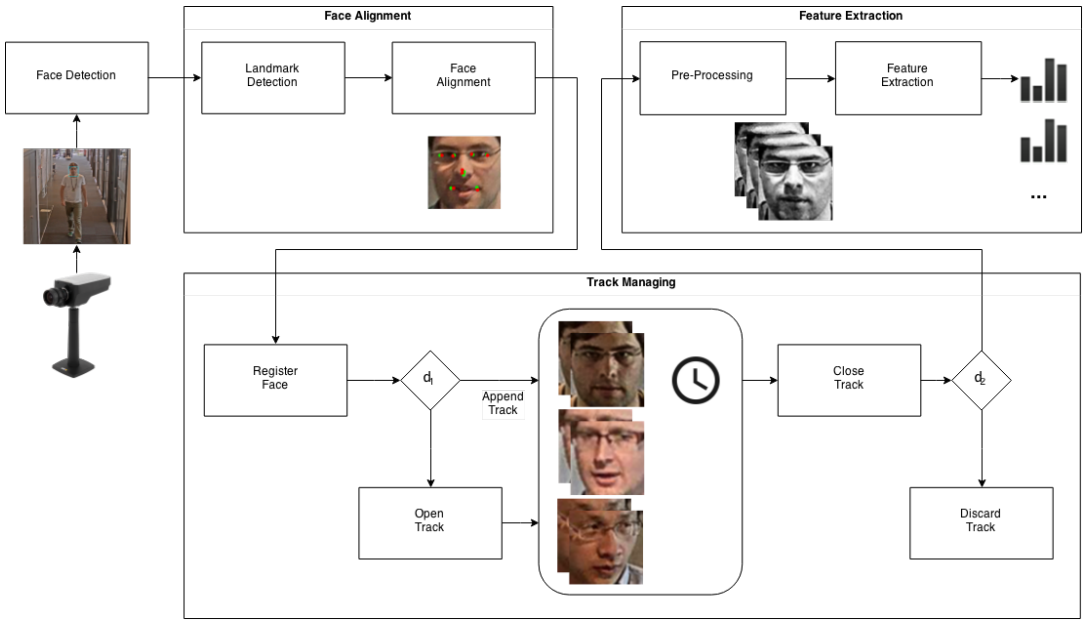

Figure 2: The video-based face recognition pipeline for feature extraction used to implement the closed area demo. For each video frame faces are detected and registered to the track managing. Track managing decides whether to open new tracks or append the face images to existing ones. A track is closed if it remains inactive for a short time period. Features are extracted from the image sequences of closed tracks. Finally a track is matched with the dynamic gallery (or database) of persons currently inside the closed area and decisions are made whether to add or remove the track from the database, or if nothing should be done.

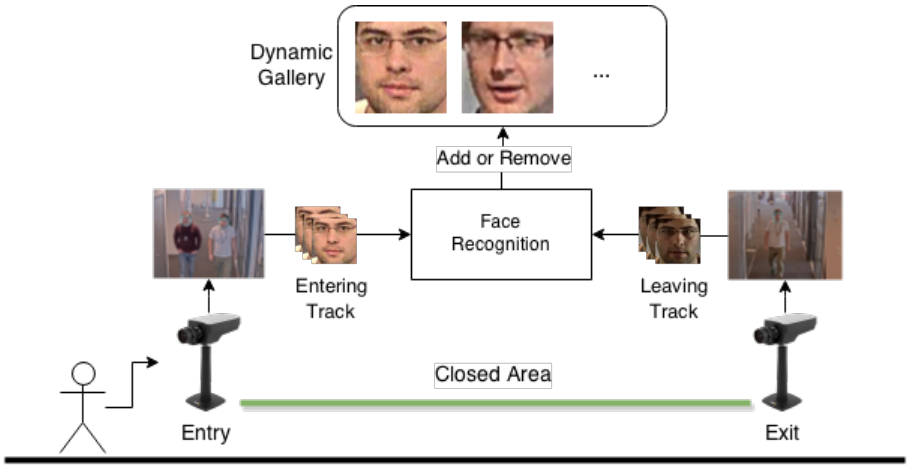

Figure 3: Persons are recorded while entering and leaving a closed area. Images from entry and exit is used to maintain information of who is currently inside the closed area; this is registered in the dynamic gallery.

the demo prototype we set up a video-based face recognition pipeline where *tracks* of face images were matched, instead of matching *individual* images. The demo prototype, running on a desktop computer, processes two simultaneous video streams in real-time either from

recorded files or live from network video and maintain a dynamic gallery of the persons that are inside the closed area.

References

[Chen et al.2012] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. 2012. Bayesian face revisited: A joint formulation. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 566–579, Berlin, Heidelberg. Springer-Verlag.

[Chen et al.2013] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. 2013. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. *Computer Vision and Pattern Recognition (CVPR)*.

[Jia et al.2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

[Karayev et al.2013] Sergey Karayev, Aaron Hertzmann, Holger Winnemoeller, Aseem Agarwala, and Trevor Darrell. 2013. Recognizing image style. *CoRR*, abs/1311.3715.

[Ojala et al.2002] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7*.