# Identifying Important People in Broadcast News Videos

Hua Gao, Hazım Kemal Ekenel, Rainer Stiefelhagen
Karlsruhe Institute of Technology
Adenauerring 2, Karlsruhe, Germany
{gao, ekenel, rainer.stiefelhagen}@kit.edu

## Abstract

*Automatic face identification in multimedia archives such as broadcast news videos is useful for indexing or retrieving documents based on important persons that appear in the video. In this paper, we propose a system which automatically detects a list of important targets such as anchor speakers or active politicians in broadcast news videos. This involves several steps including detecting faces in various conditions, associating faces to tracks and identifying whether a face track contains certain faces defined in a watch list. We evaluated this system on a database, which contains about 36 hours of broadcast news videos. Experiments show that our system achieves a very high precision with a reasonable recall rate.*

## 1 Introduction

Multimedia archives have been proliferated nowadays through the new media i.e. Internet and conventional media i.e. TV. Indexing or retrieving large scale archives is a tedious work if one needs to go through all the videos. Recently, content-based video analysis is becoming popular for these tasks. In particular, statistically, most of the pixels in video frames are occupied by human. Therefore, indexing videos according to the identities of the persons is of great interest. However, identifying people in large scale multimedia archives is not a trivial task, in which we confront several challenges including variation in pose, illumination, etc. For the videos on the Internet, one should also consider the variations in frame resolution and quality.

Fitzgibbon et. al. [7] addressed the "cast list discovery" problem in movies. In their work, faces are clustered by appearance, aiming to collect all faces of a particular character into a few pure clusters (ideally one), which can then be assigned a name manually. However, generating a few clusters per character is a challenging task. Multiple characters can be merged in a single cluster due to pose or illumination problems. Another work [8] has addressed finding particular characters by building a model of a character's appearance from user-provided training data, and efficient retrieval of characters based on example face images [6]. Fischer et al. [4] use a face tracker in order to generate longer face tracks. In an offline pre-processing step, frame-to-frame distances between all tracks are precomputed. Using the global nearest-neighbor map, retrieval and enlarging of the query-set are performed simultaneously.

Other cues such as text information has been investigated for searching specific person in news videos [9] or naming faces in broadcast news archives [10]. However, text information relies on automatic speech recognition (ASR) or optical character recognition (OCR), which might introduce additional error. It is also difficult to associate names if there are multiple faces appearing.

In this paper, we present a system for detecting and identifying important persons that appear in broadcast news videos. To avoid heavy nearest neighbor calculation for matching image sets such as the clustering-based system [7] or query by example-based system [4], we utilize available annotations for training discriminative models for each target person. The trained models can be used to retrieve a certain target in novel test video or automatic name assigning. In this system, human faces up to a certain size appear in the news videos will be detected with a robust face detector. An effective face tracker is employed to bring the gap between detections. To avoid across shot boundary tracking, shot boundaries are found by using a simple shot boundary detector. The system identifies a certain important person by verifying whether the appeared face is in the target list, and assign an identity if he or she is in the list.

We evaluate the proposed system on a large scale broadcast news video database which contains about 36 hours of video data. In the evaluation, we need to find out the time-stamp and duration of a specific target appears in the news videos. The performance of face tracking will be evaluated to check how much percent of error is caused by tracking. The performance of person retrieval is evaluated with the precision and recall metrics. An average retrieval precision of 0.926 at a recall rate of 0.754 is achieved.

The rest of this paper is organized as follows: In Section 2, we explain the detailed process of face detection and tracking, as well as the person identification approach. The experements and results of the proposed system is presented in Section 3. And finally we conclude the paper in Section 4.

## 2 Methodology

In the following subsections, the processing steps of the system are explained.

### 2.1 Face Detection and Tracking

The tracker builds upon a generic face detector using the modified census transform (MCT) as feature. The implementation of the face detector follows the approach in [5]. In order to associate face detections from the same person over time, the detector is embedded in a particle filter framework in [2], in which faces are tracked in camera network. Multiple detectors are trained so that we can detect faces in different yaw angles (0, 15, 30, 45, and 60 degrees). By mirroring the detectors we get corresponding detectors for negative yaw angles.

The detectors are integrated in a particle filter in order to evaluate them only at locations likely to contain

a face, i.e. around the locations where a face has been in the last frame. We use one particle filter for each tracked face (consisting each of a weighted set of 2000 particles in our experiments). The state

$$x = \{x, y, s, \alpha\}$$

of each particle consists of the location of the face $(x, y)$, its size in image pixels $s$ and its yaw angle $\alpha$.

The particles are propagated by means of independently drawn noise from normal distributions. For updating the particles' weights $\omega_i$, we evaluate at each particle's location the detector that has the lowest angular distance between the particle's pose angle $\alpha$ and the detector's trained angular class. The detector provides a confidence value of the detection in form of its stage sums. These stage sums are directly used as weight update, but only if all stages of the detector cascade have been passed:

$$D_\gamma = \arg\min_\gamma(\alpha - \gamma), \gamma \in \{-60, -45, \ldots, +60\} \quad (1)$$

$$\omega_i = \begin{cases} 0 & \text{if } n < n_{max}, \\ \sum_{k=1}^{n_{max}} H_\gamma k(x) & \text{if } n = n_{max} \end{cases} \quad (2)$$

where $H_{\gamma k}(x)$ is the $k^{th}$ stage sum of the cascade of detector $D_\gamma$ and $n$ is the number of passed stages. By selecting the detector with the best matching yaw angle, the particles whose pose angles best describe the current face pose are assigned the highest weights.

To automatically initialize a track, we scan the whole frame with the frontal, $\pm 30$ and $\pm 60$ degree face detectors every $k$ frames ($k = 5$ in our experiments). The value of $k$ trades off the average time until a new face is detected versus the speed of the tracker. A new track is initialized if more than three successive detections are spatially close to each other, but far enough from any known face track. A track is terminated when no detection was achieved during particle scoring (i.e. at none of the particle's locations the detector passed all stages) for more than 5 frames.

If there is more than one person in the video, care has to be taken that particles from different trackers do not coalesce onto one target. As mentioned above, in the case of multiple persons, one particle filter is used for each person/face. We ensure that particles from one track do not accidentally get scored on the face of another track by making the track centers "repel" particles from other tracks: A particle's score is set to zero if its distance to another track's center $\bar{X}_i$ is smaller than a threshold:

$$|||x - \bar{X}_i||| < \theta.$$

This simple method works well in practice. However, if a face is largely occluded by another track, it will not be reinitialized until it is far enough from the occluding face. This results in two disconnected tracks from the same person. At the moment we do not merge these two disconnected tracks.

To avoid across shot boundary tracking, the shot boundaries that are provided by a simple shot boundary detector [4] are used. Tracks will be terminated at the end of a shot segment which potentially increases the precision of face tracking.

### 2.2 Feature Extraction

Given the location of a face in a frame as output of the tracker, we perform face alignment before extracting feature vector. We try to localize the eyes using an eye detector which is—similar to the face detectors—boosted cascades of MCT features. Using the localized eye centers, we apply a rigid transformation so that the eyes are located in a fixed position in the aligned face image.

From the aligned image, we compute a feature vector according to the method in [3] which has proven to provide a robust representation of the facial appearance in real-world applications [3]. In short, the aligned face is divided into non-overlapping blocks of $8 \times 8$ pixels resulting in 64 blocks. On each of these blocks, the 2-dimensional discrete cosine transform (DCT) is applied and the resulting DCT coefficients are ordered by zig-zag scanning (i.e. $c_{0,0}, c_{1,0}, c_{0,1}, c_{0,2}, c_{1,1}, c_{2,0}, \ldots$). From the ordered coefficients, the first is discarded for illumination normalization. The following 5 coefficients from all blocks, respectively, are normalized and concatenated to form the facial appearance feature vector ($5 \times 64 = 320$ dimensional). See [3] for details.

### 2.3 Assigning Names to Face Tracks

We implemented face indexing by detecting and recognizing persons in a predefined target list. By doing this, we can find when a specific target appears in the broadcast news videos. This can be used for indexing videos according to anchor speaker or a certain politician. Since the broadcast news videos usually contain unknown persons which we are not interested in, we need to exclude the non-targets by biometric verification in addition to identification. This leads to an open set person identification problem.

Open set face recognition is different from the traditional face identification in that it also involves rejection of impostors in addition to identifying accepted genuine members that are enrolled in the database. We formulate the open-set face recognition as a multiple verification problem as proposed in [3]. Given a claimed identity, the result of an identity verification is whether the claimed identity is accepted or rejected. Given a number of positive and negative samples, it is possible to train a classifier that models the distribution of faces for both cases. Based on this idea, we trained an identity verifier for every one of the $n$ known subjects in the gallery using support vector machine (SVM) classifiers. Once a new probe is presented to the system, it is checked against all classifiers; if all of them reject, the person is reported as unknown; if one accepts, the person is accepted with that identity; if more than a single verifier accepts, the identity with the highest score wins. Scores are linearly proportional to the distance to the hyperplane of the corresponding SVM classifier.

Since a person's identity does not change within a face track if there is no track switching error, we can enforce temporal consistency. In order to make it possible to revise a preliminary decision later on, instead of relying on a single classification result for every frame an $n$-best list is used. $N$-best lists store the first $n$ highest ranked results. We choose $n = 3$ in this work.

For each hypothesis a cumulated score is stored that develops over time.

## 3 Experiments

We evaluated our system on a video database which consists of 59 videos of French TV broadcast news. The database contains two months (February and March in 2007) of evening broadcast news on the French Channel "France 2". The resolution of the videos is $352 \times 288$ pixels in a framerate of $25fps$. The average length of each broadcast news video is about 36 minutes, so in total the database contains about 36 hours of video data.

There are totally 24 important persons defined as targets which we would like to find out when they appear in the video. The target persons include *four* anchor speakers, 18 politicians (mainly French), and *two* sportsmen. Sample frames which contain different targets are shown in Fig. 1.

(a) (b) (c)

Figure 1. Sample targets appear in the broadcast news video. (a) Anchor speaker, (b) Politician, (c) Sportsman

There are usually several faces appear in one frame in news videos. To be able to evaluate both face tracking and person identification, face bounding boxes and identities are annotated. Since it is a large scale database, annotating every frame is not possible. For the moment only one frame per second is annotated. Furthermore, as we are only interested in faces up to a certain scale, faces which are larger than $20 \times 20$ pixels size are labeled. We also ignore the faces with severe occlusion since they are usually hard to detect and also cause problem for recognition. An example of one annotated frame is shown in Fig. 2.


Figure 2. Sample frame in the broadcast news video annotated with face bounding boxes and identities.

### 3.1 Experimental Setup

The broadcast news database is split evenly in training set and test set. Here "evenly" means the appearance frequency of each target person is approximately same in both sets. The training set contains 30 videos and the test set contains 29 videos. For training, we use the annotated frames to train individual models for each target. All sample images are used as positive samples for training each model. Some targets have very few training samples (less than 100). We randomly select $2n$ ($n$ = number of corresponding positive samples) face images of non-targets as negative samples, these negative samples are used for training all target verifiers. The manual eye landmarks are used for registering the face images for training if available. During testing, we first extract face tracks in the test video. The face registration is done by using automatically localized eye centers. Open-set face recognition is performed on each detected face track.

We evaluated both face detection / tracking and face identification in the experiment. For the moment, the non-targets are assigned with an universal "Unknown" ID. Due to this, we are not able to evaluate exact face tracking because we can not assign different track-IDs to the tracks of different non-targets. The precision and recall rate is used as evaluation metrics for face tracking. For the evaluation of open-set face identification, we use the precision and recall rate as well.

### 3.2 Experimental Results

There are several parameters which effect the performance of face tracking. Here we tuned two parameters i.e. minimum Track Length (TL) and Automatic Detection Frequency to initialize a tracker (ADF). High value for the parameter TL means that short tracks will not be evaluated under the assumption that a face usually appears in several consecutive frames. This assumption removes shot tracks that might not contain faces. The parameter ADF is a trade-off between recall and tracking speed. Fig. 3 plots the overall precision and recall rates of face tracking on the test video with several combination of TL and ADF values. We can see from this figure that the precision values are very high, but the recall rates are relatively low. Note that the recall rate is calculated on all the persons who appear in the test video. Non-targets in the broadcast news video are usually less cooperative than the targets. Their faces might be full-profile which can not be detected in the current system. Nevertheless, they

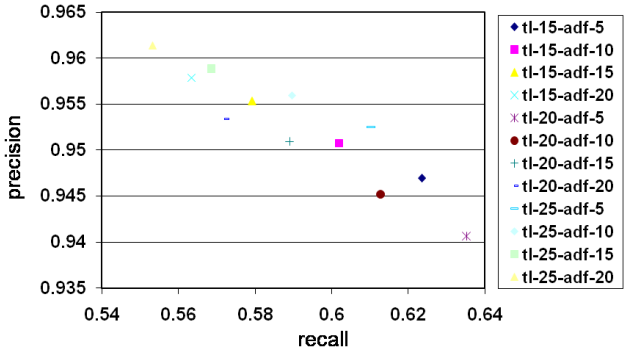


Figure 3. Performance of face tracking.

are all annotated as shown in Fig. 2. We select the $TL = 25$ and $ADF = 5$ which is a trade-off between precision and recall.

The recall rate of the targets are more important for us. As shown in Table 1, the tracking recall rate for all targets is 0.88 which is much higher than the total recall rate. The average recall rate of the anchor speakers is 0.97, which is relatively an easier task, since anchor speakers usually face directly to the camera in a close distance. 75% of the other targets are detected and tracked with a standard variation of 0.11.

Table 1. Target Tracking Performance

| Targets | Recall |
|---|---|
| All targets + Unknowns | 0.61 |
| All targets | 0.88 |
| Anchor speakers | $0.97(\sigma = 0.005)$ |
| Other targets | $0.75(\sigma = 0.11)$ |

The evaluation of the identification is conducted on the extracted face tracks. As different targets has different number of training samples, to balance the number of negative samples and positive samples for each target, we generate several virtual samples by perturbing the location of the eye pairs one pixel in the neighbourhood. The virtual samples are randomly sampled and we use again up to 3000 samples as positive samples for each target. The identification performance is listed in Table 2, which compares with and without using virtual samples for training. It is observed that both recall and precision rates are improved. Fig. 4 plots the precision-recall curve of the system with virtual samples used. The recall rate of the system is limited due to the recall of face tracking (0.88). However, the precision of the identification is very high with a moderate recall.

Table 2. Target Identification Performance

| Training method | Recall | Precision | F-measure |
|---|---|---|---|
| No virtual sample | 0.734 | 0.762 | 0.748 |
| With virtual sample | 0.754 | 0.926 | 0.831 |

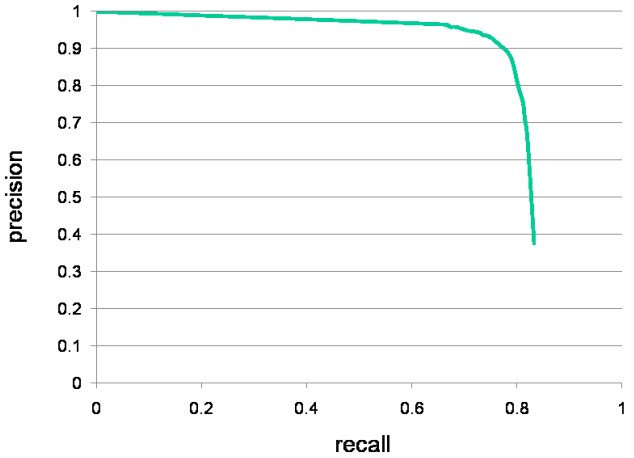


Figure 4. Precision-recall curve of the overall identification performance. Note the recall rate is limited due to the recall of tracking.

## 4 Conclusions

This paper presents a system which is able to detect and identify important persons in broadcast news videos. The main challenge of the task is the variations of face appearance which we usually have to confront in real-world face identification applications. However, important persons in news videos are more cooperative than other unknown persons. This fact makes it possible to detect and identify the targets with a reasonable performance. By using the virtual samples in the training stage, we achieved 0.926 precision at a recall rate of 0.754. The system can be applied for automatic naming in further broadcast news videos or indexing a large scale video database according to the person appeared.

To improve the performance of the current system, one can combine other modalities such as acoustic information. Full profile face detectors can also be applied to improve the recall of the face tracking as well as the recall of face identification.

## Acknowledgment

The authors would like to thank French National Audiovisual Institute (INA) for providing the corpus used in Quaero evaluations. This study is funded by OSEO, French State agency for innovation, as part of the Quaero Programme.

## References

[1] O. Arandjelovic and A. Zisserman, Automatic face recognition for film character retrieval in feature-length films, In IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 860-867, 2005.

[2] M. Bäuml, K. Bernardin, M. Fischer, H.K. Ekenel, R. Stiefelhagen, Multi-Pose Face Recognition for Person Retrieval in Camera Networks, $7^{th}$ Intl. Conf. on Advanced Video and Signal-Based Surveillance, 2010.

[3] H.K. Ekenel, L. Toth and R. Stiefelhagen, Open-Set Face Recognition-based Visitor Interface System, in Proc. of $7^{th}$ ICVS, 2009.

[4] M. Fischer, H.K. Ekenel, R. Stiefelhagen, Interactive person re-identification in TV series, In Proc. of Intl. Workshop on Content-Based Multimedia Indexing, 2010.

[5] C. Küblbeck and A. Ernst, Face detection and tracking in video sequences using the modified census transformation, Image and Vision Computing, 24(6):564-572, 2006.

[6] J. Sivic, M. Everingham, and A. Zisserman, Person spotting: video shot retrieval for face sets, in Proc. of CIVR, pp. 226-236, 2005.

[7] A.W. Fitzgibbon, A. Zisserman, On affine invariant clustering and automatic cast listing in movies, in Proc. of the $7^{th}$ ECCV, vol. 3, pp. 304-320, 2002.

[8] M. Everingham, A. Zisserman, Identifying individuals in video by combining 'generative' and discriminative head models, in Proc. of the $10^{th}$ ICCV, pp. 1103-1110, 2005.

[9] J. Yang, A. Hauptmann, M.-Y. Chen, Finding person X: correlating names with visual appearances, in Proc. of CIVR, pp. 270-278, 2004.

[10] S. Satoh and T. Kanade. Name-it: Association of face and name in video, in Proc. of CVPR, pp. 368-373, 1997.