Interactive Person-Retrieval in TV Series and Distributed Surveillance Video

Martin Bäuml, Mika Fischer, Keni Bernardin, Hazım K. Ekenel, Rainer Stiefelhagen
Institute for Anthropomatics
Karlsruhe Institute of Technology
Adenauerring 2, 76131 Karlsruhe, Germany
{baeuml, mika.fischer, keni.bernardin, ekenel, rainer.stiefelhagen}@kit.edu

ABSTRACT

Tracking and identifying persons in videos are important building blocks in many applications. For browsing of multimedia data or interactive investigation of surveillance footage it is not even necessary to uniquely identify a person. Rather it often suffices to find occurrences of a person indicated by the user with an exemplary image sequence. We present two systems in which the search for a specific person can be initiated by a sample image sequence and then be further refined by interactive feedback by the operator. In the first system, episodes of TV series have been processed offline and can be searched for occurrences of the different characters. The second system tracks people online in multiple cameras and makes the sequences immediately searchable from a central station.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Design, Performance, Human Factors

Keywords

Person Retrieval, Face Tracking, Face Recognition

1. INTRODUCTION

The massive explosion of multimedia content available on the internet, in our homes and on mobile devices calls for support in the search and retrieval of data of interest to the user. In the same way, video-based security systems monitor more areas than ever, generating vast amounts of surveillance footage, far too much to be looked at, searched and categorized by any human operator.

One common possible application in both multimedia and surveillance scenarios is the search for occurrences of specific persons. In many cases it is not necessary to actually identify the person or (TV-)character by name, but only find occurrences of the same person/character as given by an example sequence. For example, the viewer of a movie or TV show might be interested to see other scenes with his or her favorite character. In the surveillance domain, a human operator might want to know where a suspicious person has been during the last 10 minutes.

The general objective of our systems is to find occurrences of specific persons in videos. We use facial appearance as feature in order to be independent of a person's clothing. This is important in the surveillance domain when the timespan of videos to be searched is longer than one day. It is also relevant in the multimedia domain since usually characters change their clothes several times during one episode or movie.

The person to be searched for does not need to be known and trained in in advance, but rather is specified by presenting the system with an example sequence of the person in question. Taking into account the different characteristics of multimedia and surveillance data, our two systems employ different strategies and techniques for the retrieval [1, 4].

In the following, both systems and their usage will be briefly described.

2. SYSTEM OVERVIEW

The prerequisite for retrieving persons in videos is that they have been located beforehand. We employ detector-based face tracking to find faces in videos and connect them to longer face tracks. One characteristic of TV shows and movies is that in general there are much more profile than frontal faces (usually the actor never looks directly into the camera). For surveillance data, rather the opposite is true. Cameras are often placed such that they see persons from the front, walking towards the camera (e.g. in a hallway, on a sidewalk, or in front of the banking counter). We take this into account by using a tracker which can handle profile views but is slower and less accurate for the multimedia system, and a faster, more accurate tracker for the surveillance system.

In both systems, facial features of a local appearance based approach [2] are used for face representation. In short, for

MM'10, October 25–29, 2010, Firenze, Italy.

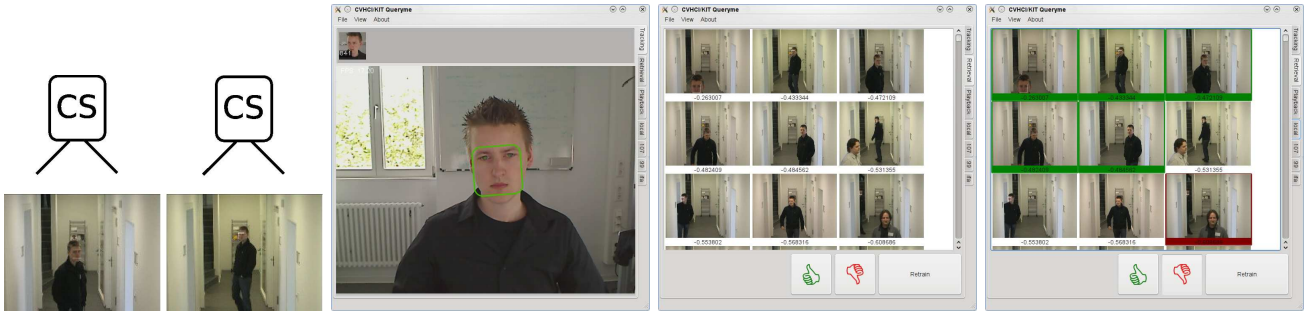

Figure 1: Our surveillance footage retrieval system. (a) The person is tracked by one or more cameras/capture stations. (b) For querying tracks of oneself, the user steps in front of the retrieval station. (c) The system reports and ranks the found sequences. (d) The user can give positive and negative feedback to improve the retrieval results.

the feature extraction the face is divided into local cells and the discrete consine transform (DCT) is applied on each of the cells. The feature vector consists of a concatenation of selected DCT coefficients from all cells.

2.1 Character retrieval in TV series

The video data of a TV series usually available long before a user starts the first retrieval. This allows for offline preprocessing: After extracting face tracks and computing facial feature vectors for each tracked face in each frame, we compute a global nearest-neighbor index of the facial appearance distances between all tracks.

A user can query the system by clicking on any tracked face in the running video. The full track associated with this face is used as the initial query set. The precomputed nearest-neighbor index allows to immediately report those tracks whose distances to the query set are smaller than a threshold.

If the user is not satisfied with the number of results, he can give feedback on a number of tracks suggested by the system. The user is asked whether these tracks show the person in question or not. The resulting positive tracks are used to enlarge the initial query set and repeat the search. [1]

2.2 Person retrieval in surveillance data

In contrast to the multimedia scenario, person retrieval in surveillance data should work without time-expensive preprocessing of the data. After all, an operator might want to query the system to determine where a suspicious person has been during the last 5 minutes. Hence it is critical that (i) the tracking and feature extraction work in real time and (ii) results to a query are available within 5 to 10 seconds. Due to this additional requirement, we cannot pre-compute a nearest-neighbor index as in the multimedia system.

Instead, we use the given example sequence to train a support vector machine (SVM) similar to approaches used for openset identification [3]. The SVM is then used to score each frame of each of the sequences in the database in order to find tracks close to the query track. This is very fast in practice (it takes less than 100 ms to search 1000 of possible target tracks with a linear SVM, and about 3 seconds with a polynomial SVM).

The demo system consists of two capture stations and one retrieval station (cf. Figure 1). The capture stations are equipped with one or more cameras to track persons in real-time. These correspond to the surveillance cameras in a surveillance deployment. The retrieval station can be used by the user to find sequences of himself in the database, as captured before in the capture stations.

A user can start the query by stepping in front of the retrieval station. The retrieval station automatically tracks the person standing in front of it and accumulates the data necessary to search for other tracks of the same person. In this way, we restrict the sequences a person can retrieve in the demo to the ones of himself, and deny random access to sequences of other persons.

As in the multimedia system, the user can refine the search by giving feedback on the results. Both positive and negative feedback is possible. The additional information is used to retrain the SVM and then retrieve more accurate results.

[1] A demonstration video of the multimedia retrieval system can be found at
`http://cvhci.anthropomatik.kit.edu/projects/person-retrieval/`

3. ACKNOWLEDGMENTS

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

4. REFERENCES

[1] M. Bäuml, K. Bernardin, M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In *Proc. 7th Intl. Conference on Advanced Video and Signal-Based Surveillance*, Boston, USA, 2010.

[2] H. K. Ekenel and R. Stiefelhagen. Analysis of Local Appearance-Based Face Recognition: Effects of Feature Selection and Feature Normalization. In *Proc. of CVPR Biometrics Workshop*, New York, USA, 2006.

[3] H. K. Ekenel, L. Szasz-Toth, and R. Stiefelhagen. Open-set Face recognition-based Visitor Interface System. In *Proc. of Intl. Conf. on Computer Vision Systems*, Liege, Belgium, 2009.

[4] M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Interactive Person Re-Identification in TV Series. In *Proc. of Intl. Workshop on Content-based Multimedia Indexing*, Grenoble, France, 2010.