Interactive person re-identification in TV series

Mika Fischer Hazım Kemal Ekenel Rainer Stiefelhagen
CV:HCI lab, Karlsruhe Institute of Technology
Adenauerring 2, 76131 Karlsruhe, Germany
E-mail: {mika.fischer,ekenel,rainer.stiefelhagen}@kit.edu

Abstract

In this paper, we present a system for person re-identification in TV series. In the context of video retrieval, person re-identification refers to the task where a user clicks on a person in a video frame and the system then finds other occurrences of the same person in the same or different videos. The main characteristic of this scenario is that no previously collected training data is available, so no person-specific models can be trained in advance. Additionally, the query data is limited to the image that the user clicks on. These conditions pose a great challenge to the re-identification system, which has to find the same person in other shots despite large variations in the person's appearance. In the study, facial appearance is used as the re-identification cue. In order to increase the amount of available face data, the proposed system employs a face tracker that can track faces up to full profile views. A fast and robust face recognition algorithm is used to find matching faces. If the match result is highly confident, our system adds the matching face track to the query set. This approach help to increase the variation in the query set, making it possible to retrieve results with different poses, illumination conditions, etc. Furthermore, if the user is not satisfied with the number of returned results, the system can present a small number of candidate face images and lets the user confirm the ones that belong to the queried person. The system is extensively evaluated on two episodes of the TV series Coupling, showing very promising results.

1. Introduction

Person re-identification in videos is the task where the user selects a person in a video and the system then retrieves shots from the same or other videos, in which the same person is present. This application scenario can be very useful for several multimedia applications. For example the viewer of a video on YouTube or a show or movie on his TV might appreciate an intelligent fast-forward feature that can show him the scenes that contain a person of interest. Archive personnel also often have to go through large collections of video data in order to find shots of specific persons of interest. In the security domain, such a system could be used to find occurrences of certain persons in surveillance footage, for instance when a suspicious person in a railway station is detected, it might be crucial to quickly find out where the person was before and what he did. Finally, such a system could be used to quickly annotate large amounts of video data, for instance as preprocessing for video archives or for computer vision researchers. All these applications have in common that the users might be willing to put in a small amount of manual effort in order to get better results. In some cases, such as search in video databases, the manually provided information can even be saved, so that it has to be given only once. This fact strongly suggests that a person re-identification system should optionally allow the user to assist in the retrieval process in order to achieve better results.

In this paper, the task of person re-identification in TV series is addressed, using facial appearance as the re-identification cue. By this we mean that the user selects a person in an episode of a TV series by clicking on his face, and the system then retrieves shots from the same episode or other episodes, where the selected person also appears. As can be expected, this task poses many challenges for face recognition methods. There are no constraints on head pose, facial expression, illumination conditions, use of accessories, and other effects which lead to large variations in facial appearance. Additional difficulties in this scenario are that no training data is available for the queried persons, so that no person-specific models can be trained in advance. Furthermore, the query data is very limited, since only the image that the user selected to start the search can be used. This makes it very difficult to find other occurrences of the person because they might be taken under very different lighting conditions, view the face under a different pose, have different facial expressions, etc.

Most of the previous work on person re-identification was done in a security-related context, mainly for surveil-

lance scenarios [9, 10, 11]. In these applications, mainly color cues or local descriptors of the body appearance are used for re-identification. Facial appearance is rarely used. These approaches cannot be easily transferred to the multimedia domain, since in many videos, especially in TV series or movies, persons change their clothing often and sometimes only a small part of the body is visible. However, in these videos, the faces of the actors are well visible most of the time.

There has been significant research on using face recognition for video retrieval in multimedia applications. The published approaches can be divided into re-identification scenarios [2, 12, 14], where one or more video frames are used to start a query and scenarios that use (weakly) labeled training data to train person-specific models prior to the retrieval [1, 13, 15]. Most approaches, such as [2, 12, 14, 13] are limited to frontal face detections. Sometimes a tracker is used in addition to a frontal face detector. In [14] and [12], the additional non-frontal images are however not used for face recognition. Even if they are used [1, 13], it still does not solve the problem that a track is only started when there is at least one frontal face detection. This makes it impossible to retrieve content where persons are only seen in non-frontal poses, which is often the case. For instance, in the data used in this study, it was the case for one third of all occurrences. Only very recently have approaches been published, which take advantage of non-frontal views [15], however not in a re-identification scenario.

Although there have been several studies on person retrieval for multimedia applications, the case of a re-identification scenario with unrestricted face pose, where no training data is available to train person-specific models, has not yet been addressed in the literature. Furthermore, none of the published studies incorporates user feedback into the retrieval system.

The system proposed in this work first analyzes a video, using a shot boundary detection and face tracking component, and then extracts features that can be used for retrieval. This is done offline and the results are stored in a database. The user watching the video can then, at any point in the video, click on an actor's face and the system retrieves shots from the same or a different video in which the same actor is present. If he is not yet satisfied with the number of returned results, he can assist in the retrieval process by confirming face images depicting the queried person among a small number of candidate images presented to him by the system. The query is then continued using the user-provided selections, yielding more results than before.

There are four main features of our system: first, we employ a face tracker that can successfully track persons across pose changes [8] and we use this information to extract frontal and non-frontal face images. The reason for this is that in most videos, there is a significant number of shots where only profile views of an actor are available. If we restricted our system to frontal faces, we could no longer retrieve these shots using the face information alone.

Second, we use a robust face recognition algorithm [6] to match tracks. The algorithm is local appearance-based and has been shown to perform very reliably in several real-world applications, such as door monitoring [16], a visitor interface [7] or smart environments [4].

Third, the retrieval process itself works differently from usual approaches. First, a query set is built by collecting all face images of the track the user selected by clicking on a face image. Then the system searches for very close matches to any image in the query set. The resulting matches and the tracks they belong to are then directly added to the query set. Then a new search is performed using the enlarged query set, and so on, until no more matches can be found automatically. The enlarged query set is then presented to the user. This approach allows the system to increase the variation in the query set and this way close the gap from the pose of the face in the query image to the different poses in the target images, by leveraging the temporal association provided by the tracker and by enlarging the query set automatically using results that have a high confidence.

Finally, the system allows the user to assist in the retrieval process. It presents the user with a small number of candidate images and lets the user confirm which ones depict the queried person. Using this additional query data, the search process is restarted and yields more results.

The proposed system is evaluated quantitatively in experiments on two labeled episodes of the TV series *Coupling*. Within-episode and cross-episode retrieval as well the effect of user interaction on the system performance are evaluated. The results indicate that the system works very well, given the difficult data. The fully automatic retrieval provides a reliable basis for further improvement through interactive user feedback. Even for small amounts of user feedback, very high recall rates can be achieved. For instance, with five rounds of user feedback, the recall already reaches over 80% at 95% precision. With ten rounds, the recall further increases to around 90%. This is the case both for within-episode and cross-episode retrieval.

The remainder of this work is organized as follows: in Section 2, a system overview will be presented and several system components will be described in detail. In Section 3, the experimental setup and the results of the experiments will be discussed. Finally, in Section 4, conclusions will be given.

2. System description

The structure of the proposed system is shown in Figure 1. As can be seen, it consists of an offline and an online

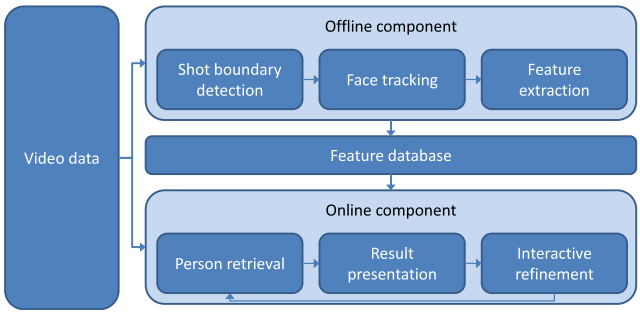

Figure 1. Structure of the proposed system

component. The offline component performs the following steps with a new video: first, a shot boundary detection system [3] is run on the video in order to segment it into separate shots. On each shot, the face tracker [8] is then run, resulting in multiple face tracks. For each face track, all the face images are extracted from the video and local appearance-based features suitable for face recognition are derived from them. These features are then stored in a database.

When the user is watching a video, he can click on the face of an actor and the online component then uses the information stored in the database to retrieve shots from the same or a different video in which the same actor is present. The user can then either view any of the retrieved shots or, if he is not yet satisfied with the number of returned results, interactively help the system by selecting face images depicting the queried person from a small number of candidate images presented by the system. Using the user-provided selections, the system can then retrieve further shots[1].

2.1. Feature extraction

The tracker provides a number of tracks for each shot in a video, where a track is a sequence of face boxes in subsequent frames. In the next step, features suitable for face recognition are extracted from the face images.

Normally, alignment is a vital step in a face recognition system. However, robust facial feature localization, which is necessary for correct alignment, is a difficult task, especially if non-frontal faces are to be considered. There are, however, systems that could avoid the alignment step by providing a lot of data to the system [13]. The additional data is expected to cover at least partly the variations in scale, rotation and translation, that would be normalized out by the alignment step. So in this work, the alignment step was omitted, and the features were extracted directly from the cropped face images.

[1] A demonstration video of the system can be found at `http://cvhci.ira.uka.de/~mfischer/person-retrieval/`

For feature extraction, the method presented in [6] was used. It is based on local appearances, has been successfully applied in multiple real-world face recognition systems [5, 16, 7] and achieved the best face recognition results in the CLEAR 2007 evaluations [4].

The approach can be summarized as follows: first, the cropped face image is resized to 64×64 pixels and the rightmost and leftmost eight columns are removed to reduce the amount of background in the image. The remaining 48×64 image is then divided into non-overlapping 8×8 pixel blocks and the discrete cosine transform (DCT) is applied to each of the blocks. The obtained DCT coefficients are ordered using zig-zag scanning. From the ordered coefficients, the first one is discarded, and the next five are kept to build a local feature vector, which is then normalized to unit norm, in order to reduce illumination effects. Finally, the local feature vectors are concatenated to construct the global feature vector.

2.2. Person re-identification with interactive feedback

The retrieval process starts with a single face that the user selects by clicking on it. The track to which the selected face belongs is then added to the query set. The system then automatically enlarges the query set by searching for tracks with a distance to one of the tracks in the query set that is lower than some threshold, where the track distance is defined as the minimal L1-distance of any two feature vectors extracted from the tracks.[2] The found tracks are added to the query set and the process is repeated until no more tracks can be found. An illustration of this process is shown in Figure 2.

At that point, the enlarged query set is presented to the user. If he is not yet satisfied with the number of returned results, he can assist in the retrieval process, by manually confirming images of the queried person from 21 candidate images that are presented by the system. The candidate images correspond to the 21 tracks that are the closest matches to the tracks in the query set, while having a distance above the threshold for automatic addition to the query set. After the user confirms which images depict the searched person, the tracks corresponding to the selected images are added to the query set. The tracks corresponding to the other images are marked as not belonging to the correct person, so they will not be presented to the user as candidate images in further selections. Then the automatic query set enlargement process is restarted.

The automatic query set enlargement makes it possible to deal with different poses of the faces in the database. An

[2] Since the matching with a large number of feature vectors is slow, and since only tracks are compared with each other in this system, as another preprocessing step, the closest distances between face images of two tracks are precomputed and saved for later use.

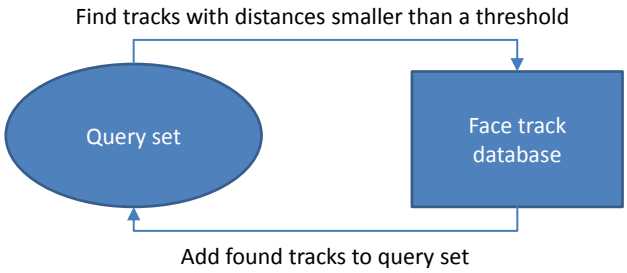

Figure 2. Query set enlargement process

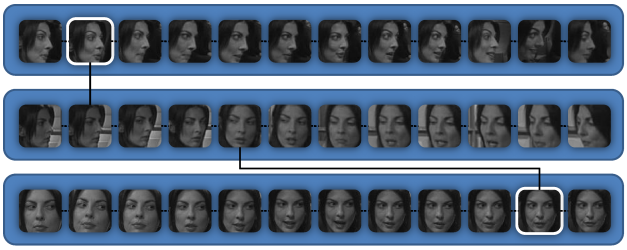

Figure 3. Matching different poses: Each row corresponds to a different track. The marked images with very different poses can be matched by exploiting the temporal association (horizontal, dotted) and the associations provided by the face recognition algorithm (vertical, solid)

example of this can be seen in Figure 3. The first track, which contains only profile face images, can be matched to the second track, which contains profile and frontal face images. By adding this found track to the query set, other tracks which might contain only frontal face images, such as the third track, can be found, as well. So the temporal association provided by the tracker together with the query set enlargement procedure make matching faces across different poses possible.

For this approach to be successful, it is important that the precision of retrieval stays very high. If face images of other persons than the originally queried person get in the query set, the approach usually fails because it will automatically extend the query set with more and more images of this wrong person. So the approach critically hinges on the accuracy of the tracker. It should ideally never switch from one person to another in one track. In addition, the threshold for automatic query set enlargement has to be set rather conservative, since there is currently no recovery from wrong matches, although a possibility for the user to manually remove tracks of incorrect persons that were added to the query set could be a useful further improvement.

3. Experiments

Two episodes of the British TV series *Coupling* are used as experimental data in this work. Each episode is about 30 minutes in length and contains around 44,000 frames. The resolution of the video data is 720×405 pixels. Both episodes were labeled with face positions and identities.

The face tracker extracted 129,352 face images grouped in 1,886 face tracks from the two videos. The width (and height) of the face images ranged from 22 to 304 pixels with a median of about 100 pixels.

For the evaluation of person re-identification, the precision and recall metrics are used. The tracks supplied by the tracker are used as the base set, so that the metrics are defined as follows:

$$\text{precision} = \frac{\text{\# retrieved tracks that are associated to the correct person}}{\text{\# retrieved tracks}}$$

$$\text{recall} = \frac{\text{\# retrieved tracks that are associated to the correct person}}{\text{\# tracks that are associated to the correct person}}$$

Two experiments were performed to evaluate the re-identification component. In the first experiment, retrievals within a specific video were evaluated and in the second one, cross-video retrieval was evaluated.

For the first experiment, all possible queries were performed, using varying thresholds for query set enlargement and a different number of manual selections (which were simulated automatically using the ground-truth labels). The precision and recall results were then averaged over all queries. The results of these experiments are shown in Figures 4 and 5. It can be seen that the manual confirmations improve the results considerably. For instance in episode 1, the automatic retrieval is able to achieve recall rates around 25% for reasonably high precisions. However even with only one round of manual confirmations, the recall increases to almost 50%. More manual confirmations lead to further increases and with ten selections, a very high average recall rate of over 91.2% is achieved, at an average precision of 97.6%. The results for episode 6 are similar.

Note that the interesting look of the graphs comes from the effects that the user feedback and the automatic query set enlargement have on the results. For instance, consider that one track of an incorrect person is added to the query set because of an increased distance threshold. Chances are, that the automatic query set enlargement procedure will then find more tracks of this incorrect person, and add them to the query set as well. So in effect, when the precision decreases, it often does so in a dramatic drop.

Another unusual effect is that the recall can actually decrease with increasing distance threshold, if manual confirmations are used. This counter-intuitive fact can however be explained easily. Consider, for instance, the case where only one track is in the query set before the manual confirmation. The tracks closest to this single track are then

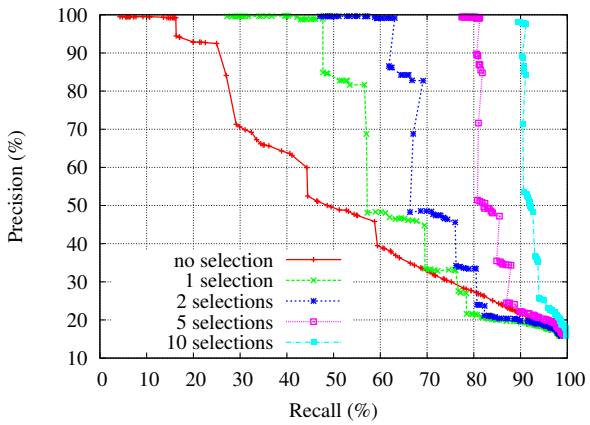

Figure 4. Retrieval results within episode 1

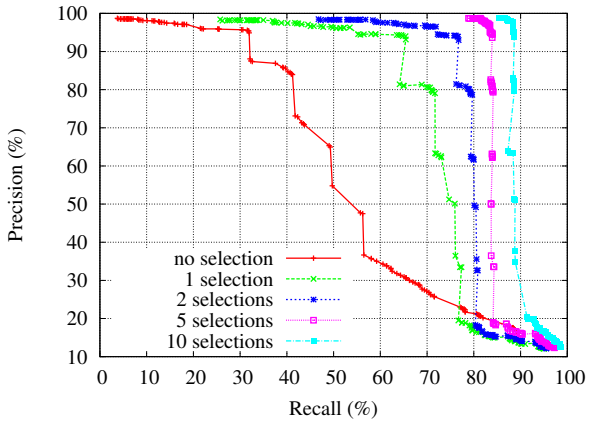

Figure 5. Retrieval results within episode 6

presented for manual selection. Assume that one of these tracks can be matched automatically to many other tracks, so it is very important because it increases the recall substantially. When the threshold is increased, more than the single track from before can be in the query set before the manual confirmation. Now, the closest tracks to *any* track in the query set are presented for manual confirmation. This means that our important track from before could be left out of the manual confirmation process, in effect possibly reducing the recall.

An example of a query image and some of the results that were automatically retrieved using it are shown in Figure 6. It can be noted that the system can retrieve shots with very different illumination conditions, poses and expressions, than in the image that was used for querying.

In the second experiment on inter-episode retrieval, one track from episode 1 was selected and then tracks from episode 6 were retrieved. This was done for all possible queries and the results were averaged as before. The results are shown in Figure 7 and they show that the matching between episodes is more difficult in the automatic case, yielding 19% recall at a precision of 90%, compared to 25% for retrieval within episode 1 and 32% for episode 6. But similarly to the within-episode experiment, user feedback improves the results significantly. Even after two rounds of user feedback, the results increase to 75%, which is comparable to the within-episode results.

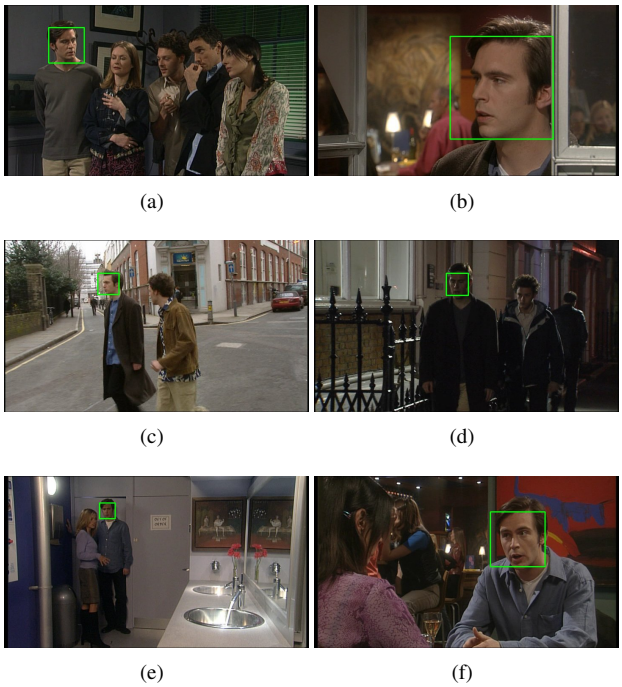

Figure 6. Example person retrieval results: (a) query face, (b)-(f) retrieved results with varying pose, illumination conditions, face sizes and facial expressions

Note that the results approach the within-episode results for episode 6 in Figure 5 with increasing amount of user feedback. This comes from the fact that more and more tracks from episode 6 are added to the query set by the user feedback process and so the task becomes more and more similar to within-episode retrieval.

4. Conclusions

In this work, a system for re-identification of persons in videos has been presented. The system incorporates a shot boundary detection module, as well as a face tracking module which can reliably track faces from frontal up to full profile poses. This is very important, since in many videos, there are shots where only non-frontal views of a person are available.

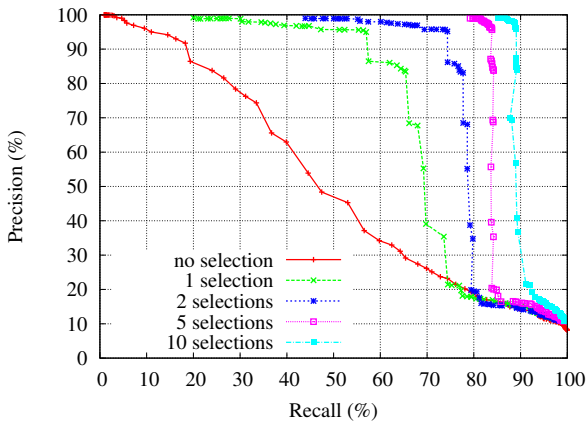

Figure 7. Inter-episode retrieval results

The following features of the proposed person re-identification system allow it to overcome the problems of unavailability of prior training data, very limited amount of query data and possibly severe mismatch of head pose, illumination conditions, facial expressions:

First, the system exploits the temporal association provided by the tracker in order to work on whole tracks instead of images for query as well as target, which increases the amount of data that can be used for matching. Second, the system uses a robust local appearance-based face recognition algorithm, which has been shown to perform very reliably in real-world scenarios. Third, the retrieval process works by iteratively enlarging the query set with tracks that match tracks already present in the query set. And finally, interactive user feedback was integrated into the system. The user is presented with candidate face images and has to confirm which ones depict the person the user is searching for. These features help to increase the variation in the query set, making it possible to retrieve faces with different poses, illumination conditions, facial expressions, etc.

The performed experiments for within-episode and cross-episode re-identification show that the system works very well, considering the difficult data. Automatic retrieval provides a reliable basis for further improvement of the results through interactive user feedback. It could be shown that even small amounts of user feedback improve the results significantly, yielding results of over 80% recall at 95% precision after only five rounds of manual confirmations, and around 90% recall at 95% precision after ten rounds, both for within-episode and cross-episode retrieval.

5. Acknowledgements

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

References

[1] N. Apostoloff and A. Zisserman. Who are you? - real-time person identification. In *British Machine Vision Conf.*, 2007.

[2] O. Arandjelović and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[3] H. K. Ekenel, M. Fischer, H. Gao, K. Kilgour, J. S. Marcos, and R. Stiefelhagen. Universität Karlsruhe (TH) at TRECVID 2007. In *Proc. of the NIST TRECVID Workshop*, Nov. 2007.

[4] H. K. Ekenel, Q. Jin, M. Fischer, and R. Stiefelhagen. ISL person identification systems in CLEAR 2007. In *Proc. of the CLEAR Evaluation Workshop*, 2007.

[5] H. K. Ekenel, J. Stallkamp, H. Gao, M. Fischer, and R. Stiefelhagen. Face recognition for smart interactions. In *Proc. of the IEEE Int'l. Conf. on Multimedia and Expo*, July 2007.

[6] H. K. Ekenel and R. Stiefelhagen. Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. In *Proc. of the CVPR Biometrics Workshop*, June 2006.

[7] H. K. Ekenel, L. Szasz-Toth, and R. Stiefelhagen. Open-set face recognition-based visitor interface system. In *7th Intl. Conf. on Computer Vision Systems*, Oct. 2009.

[8] M. Fischer. Automatic identification of persons in TV series. Diplomarbeit, Interactive Systems Labs, Universität Karlsruhe (TH), May 2008.

[9] T. Gandhi and M. M. Trivedi. Person tracking and reidentification: Introducing panoramic appearance map (PAM) for feature representation. *Machine Vision and Applications*, 18(3):207–220, Aug. 2007.

[10] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[11] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ACM/IEEE Int'l. Conf. on Distributed Smart Cameras*, pages 1–6, Sept. 2008.

[12] P. Li, H. Ai, Y. Li, and C. Huang. Video parsing based on head tracking and face recognition. In *ACM Int'l Conf. on Image and Video Retrieval*, 2007.

[13] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *Proc. of the IEEE 11th Int'l. Conf. on Computer Vision*, Oct. 2007.

[14] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *Proc. of the 4th Conf. on Image and Video Retrieval*, 2005.

[15] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – learning person specific classifiers from video. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[16] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *Proc. of the IEEE 11th Int'l. Conf. on Computer Vision*, Oct. 2007.