Facial Reconstruction on the Basis of Video Surveillance System for the Purpose of Suspect Identification

Damian Pęszor[1,2(✉)], Michał Staniszewski[2], and Marzena Wojciechowska[1]

[1] Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland
{damian.peszor,mwojciechowska}@pja.edu.pl
[2] Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
michal.staniszewski@polsl.pl

Abstract. Growing importance and commonness of video surveillance systems brings new possibilities in the area of crime suspect identification. While suspects can be recognized on video recordings, it is often a difficult task, because in most cases parts of suspect's face are occluded. Even if there are multiple cameras, and the recordings are long enough to expose entirety of suspect's face, it is challenging for an observer to accumulate information from different cameras and frames. We propose to solve this problem by reconstructing a three-dimensional mesh that could be presented to an observer, so he could identify suspect based on accumulated information rather than fragmented one, while choosing any angle of observation. Our approach is based on extraction of anthropological features, so that even with imperfect recordings, the most important features in terms of facial recognition are preserved, while those not registered might be supplemented with generic facial surface.

Keywords: Facial reconstruction · Suspect identification · Facial composite · Surveillance

1 Introduction

The advent of video-based surveillance systems provided new tools for forensic investigators. In the past, when surveillance systems were few and far between, facial composites were constructed on the basis of eyewitness' memory with the help of trained professional artist. Because of the stressful nature of the experience, imperfection of perception and memory as well as inability to unambiguously convey information about human appearance, the margin of error using such approach is relatively high. Accessibility of recordings obtained through video-based surveillance systems made it possible for an eyewitness to recognize the suspect on a recording which provides the investigators with more reliable data. However, such recognition is not as certain as one might think. Even

N.T. Nguyen et al. (Eds.): ACIIDS 2016, Part II, LNAI 9622, pp. 467–476, 2016.
DOI: 10.1007/978-3-662-49390-8_46

though eyewitness does not have to verbalise his recollections and might not be consciously aware of his memories, he still might not recognize the suspect. Some of the reasons that the chances for successful recognition are lowered, are related to the fact, that eyewitness experience recordings in specific way, such that:

- Recordings from different cameras are viewed sequentially, which means that data from different recordings are processed separately. The effect of synergy is therefore greatly diminished, as information from one camera is not perceived while eyewitness watches another recording.
- Each surveillance camera is installed in a fixed point of space and records the scene from fixed perspective. This means, that eyewitness is not able to change perspective to the one that better corresponds with his observations and therefore his certainty is reduced.
- In most cases, suspect will not stand still while being recorded. Facial features will be visible clearly for several frames, but then suspect will change his pose, or occlusion will occur. Since eyewitness does not see all recorded features at once, but only partial image of face at any given time, his certainty will be reduced.

To resolve these problems, we employ a part of a method originally designed for the purpose of facial animation. This method, designed to retarget mimicry obtained from a video-based performance capture to different facial structure involves reconstruction of facial mesh from multiple recordings on the basis of positions of fiducial points characteristic from anthropometric point of view. Quality of reconstruction depends on the number of cameras which recorded the suspect, their positions, parameters and movement of suspect and occlusions which makes it hard to estimate the error. This solution is not intended to be an automatic facial recognition software, which would require much more reliable data and could be used mostly in case of identification of previously apprehended criminals. Rather than that, the authors intend to assist eyewitness in suspect identification at the stage of facial composite creation, when the data is present, but it is too distributed for eyewitness to properly analyse. With our approach, following statements hold:

- Data from different recordings are combined, so that eyewitness perception benefits from all of them at once.
- Eyewitness is able to see suspect without partial occlusions, as entire face is reconstructed. Even in case of part of face not being present on any recording, it is reconstructed from generic model, which eliminates eyewitness' confusion.
- Eyewitness is able to see suspect's face from different perspective and is able to use one that corresponds best to his experience, therefore being able to perceive features that he remembers.

2 Materials and Methods

2.1 Facial Area Localization

To develop and test suggested approach, a number of recordings was taken in different environments. Apart from recordings specifically taken for the purpose

of facial reconstruction, VMASS video sequences dataset [1] was used to test proposed approach with real-life data. Since proper reconstruction requires taking into account camera parameters in order to be able to reconstruct each pixel of an image into a three-dimensional point, a calibration of each camera is required. In case of video surveillance systems, calibration itself is an issue that has to be recognized, due to two factors - working infrastructure and cameras' specificity. One cannot assume particulars places, in which suspect will be present, and therefore there is no possibility of establishing infrastructure that will be properly calibrated. Investigators have to use existing, working infrastructure. Since most of such systems are not designed for processing of video recordings, one cannot expect that there will be any markers that will faciliate calibration, therefore calibration procedure has to employ data available on the video at the time of the recording without any previous preparation.

While some of cameras might be stable enough that their movement is negligible, one cannot assume that this is the case. Tremors related with heavy machinery operating, cars on a street, or even human steps in case of loosely installed camera might have significant influence on calibration process. Also, cameras might change their perspective based on program or human intervention. For example, change in tilt will not only result in small tremors induced by tilting mechanism, but will also change the observed volume which will result in a need for new calibration.

To deal with such issues, as first step of suggested approach, we employ a navigation algorithm previously described in [2]. It is used to calibrate cameras based on features present in the background image of the recording. By using calibration procedure described in [3], intrinsic parameters of cameras are obtained in order to be used later to adjust projection of feature points to their correspondences on the captured images.

Data used in further face detection will be in first stage applied for tracking of multiple objects in views from many video cameras. The first problem present in such situation will be identification of particular object/person and application of obtained information across all views. Objects will be therefore identified in single view, however the problem of reidentification in another camera's view is still present. Here, extremely important will be solution for problems of change in illumination, possibility of people moving in one view and the place between two views of cameras where object can be not visible.

One of the current methods that is suited for the problem at hand was presented during Multi-Camera Object Tracking (MCT) Challenge Zurich, 12th September 2014 in conjunction with ECCV 2014. The method formed in [4] extends the basic tracking idea. The general concepts of the method relies on searching for affinities between tracks basing on features and spatial temporal context. Algorithm consists of four main parts: online sample collection, discriminative appearance learning, relative appearance context learning and track association.

Online sample collection - basing on tracks computed by multi-target tracking by on-line learned discriminative appearance models and spatial context the presented method obtain trajectories describing each object. It may be assumed

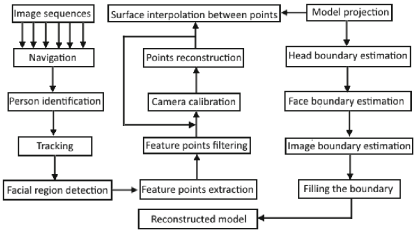

Fig. 1. Scheme of proposed approach

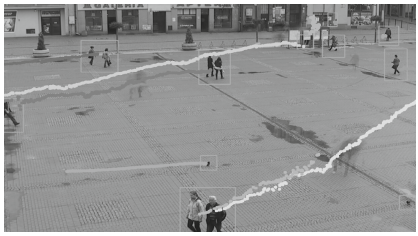

Fig. 2. Sample captured frame with tracking

that one object may occur only ones in one frame. In that part the presented method collects tracks and assigns them as positive or negative depending on possibility of fitting object inside one camera and between many.

Discriminative appearance learning - that part of algorithm relies on creation of strong feature model that will be able to distinguish very similar objects. In order to calculate the cost of affinity, this method applies the following pre-processing steps. Firstly color normalization has to be introduced which will compensate differences in colors across many cameras. On such processed image algorithm applies feature descriptors as a histograms of RGB colors or HOG [5], in which affinities are computed basing on correlation coefficients. Features are then used in order to discriminate objects across many cameras.

Relative appearance context learning - method introduces scheme for dealing with many objects moving in one group which is usually difficult to discriminate. The solution lies in identification of groups having similar velocities and distance.

Track association - last part of algorithm which connects short tracks in long trajectories related to particular object. Problem is formulated as construction of cost matrix solved by Hungarian algorithm.

The result of tracking algorithm which will be visible in the form of bounding box will be used in detection of faces. Tracking method gives reliability that faces in bounding boxes in different camera views will be related to particular object. One of the possible solution was described in [6] which detects parts of objects. Problem of object detection is solved by mixtures of multiscale deformable part models. System relies on methods for discriminative training of classifiers that make use of latent information. It also relies on efficient methods for matching deformable models to images.

2.2 Reconstruction Using Feature Points

Once facial region is obtained, facial feature points can be extracted from the images. Few issues have to be considered in selection of correct approach. First, since goal of the method is to reconstruct anthropometrically correct facial model on the basis of many frames from different cameras, the amount of extracted points is important, the higher the amount of points that the method can extract, the better. Second, surveillance cameras in most cases are installed above the

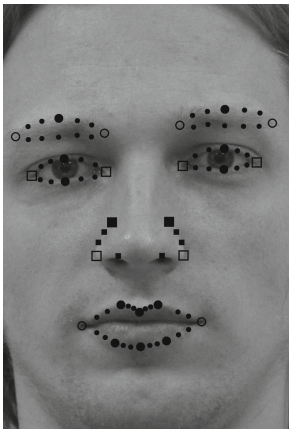

Fig. 3. Structural set (*rectangles*) and expressive set (*circles*) of contour points (*small figures*) and fiducial points (*big figures*); both global (*empty figures*) and local (*filled figures*).

height of human face, which is a first source of rotation from frontal pose. Second source is that one cannot expect suspect to look in the direction of camera, so the face will mostly be recorded with horizontal rotation as well. The method used for extraction of feature points has to be possibly robust, so that rotations of facial pose will not render most of frames useless. Third, the method has to be able to handle some occlusions, for example - with one eye occluded it has to be able to correctly establish the position of other feature points. This allows to use frames which do not have every point found, which is a commonplace in case of surveillance cameras - those points will be reconstructed using different frames. Fourth issue, is that the method should be able to use data from different frames. In many cases either occlusions or rotations will prove difficult to find feature points, an information from previous frame can be used to decide where to look for a given feature point, and therefore minimize the possibility of finding different feature instead. This aspect is also important due to facial expressions which can significantly influence the process of extracting feature points.

A method that proved to meet requirements of approach presented in this paper and therefore was selected is based on multi-state hierarchical shape model as described in [7]. This method is able to extract 26 fiducial points and many contour points between them (authors use 56 contour points, and though it can be easily changed, it proved to be a reasonable number for our approach). Since in proposed approach, the data from different frames is accumulated, there is a possibility, that some of those frames might contain different facial expressions. Using all feature points in further steps would therefore introduce errors related to differences between facial images. To mitigate the problem, all facial feature and feature contour points are divided into two sets: structural and expressive. Structural set contains points which are only slightly affected by facial expressions, those will be used in primary reconstruction. Expressive set contains points

which are greatly affected by expressions, those might be filtered out so that same facial expression will be reconstructed using different frames. Membership of each point is presented in Fig. 3. Further part of proposed approach was inspired by facial reconstruction algorithm presented in [8], although there are notable changes in described method, due to the differences in data between controlled environment (as in [8]) and environment that is under surveillance. Using automatic feature detection rather than manual one (as in mentioned article) significantly decreases the amount of time needed for proper reconstruction. Although the mentioned approach proposes using *Downhill Simplex* algorithm [9] to minimize residual error value and thus find appropriate camera's intrinsic parameters, it is not suitable for presented application. This is due to the following:

- Since feature points are selected in an automatic way, their position at this point is not necessarily as precise as in case of manual selection and some of them might be inappropriate, which would heavily influence the residual error value.
- Only small portion of recorded image represents suspect's face, it is therefore not a good representation of entire image and thus camera's properties.
- Intrinsic parameters can be calculated on the basis of entire recorded image, which is much more precise.

Fig. 4. Detected feature points

With known intrinsic parameters, the distance between found feature points that belong to the structural set and projection of their correspondences on generic model, is iteratively minimized using POSIT [10]. This step might be considered as calibration, and therefore be confusing since calibration was already performed. The difference is, that in this step, the orientation and position of camera around the face is found. For each image in the filtered set, different position and orientation of camera is found, even if in fact it is the same camera. This is due to the fact, that the position of face in three-dimensional space changed, which is modeled as different position of camera in relation to face.

Once orientation and position of the camera regarding to the face is obtained, the position of feature points is reconstructed by adaptive symmetry as in [8],

which changes generic feature points' positions to ones more related to the shape of reconstructed face. Since their position is still not perfect, calibration/reconstruction process is repeated until convergence. Once reconstruction of feature points from structural set is complete and the position and orientation data is calibrated, the feature points belonging to expressive set are reconstructed. There are two aspect that need to be considered. First, there is a chance, that automatic feature points selection will yield wrong results. Second, some of expressive points might be under influence of facial expressions (which is why those are not considered in calibration), which will reduce the effectiveness of reconstruction. Therefore, not every image in the filtered set has to be used to reconstruct every point. For each point in expressive set for each image, we calculate sum of distances to every point in reconstructed structural set. Only those expressive points in an image which have a sum of distances different by not more than standard deviation of all the sums of this expressive point in all images are considered a valid source of reconstruction. Having reconstructed all feature points from both structural and expressive sets, radial basis function in the form of $\sigma(r) = r$ is used to interpolate the changes in feature points and modify the surface between them.

2.3 Correcting Reconstructed Mesh Using Boundaries

Using radial basis function will properly reconstruct most of anthropometrically important features. However, due to the fact that it is not guaranteed that all feature points will be recognized, as well as the fact that recognized feature points might be far from each other, it is necessary to correct the estimation. Otherwise, it would be possible that interpolation from neighbouring feature points would result in creation of unwanted features in minimas of sum of feature points' influence. The method proposed by [8] is used here as well. The edges of the mesh are projected onto the image using calibration data creating the *projected boundary*. The desired boundary of the model should coincide with the boundary of face/head on the image, so the *image boundary* has to be found.

In case in which the calibration is correct, *projected boundary* is similar to the real *image boundary*, which makes it suitable for use as an initial clue in finding of *image boundary*. In other cases, due to following logical operations, the method will not yield any boundary, and therefore will not negatively affect the overall reconstruction. To obtain *image boundary*, the method presented in [11] is used with some modifications.

Both head boundary and face boundary is extracted using mentioned method. Since the projected model consist of entire head rather than only part of it, the logical disjunction of both boundaries is used. Next, the conjunction of obtained boundaries and the *projected boundary* broadened by half of intercanthal width as obtained from found feature points. The selection of intercanthal width is based on few reasons; the stability of this distance around different populations (see [12]), the fact that inner eye corners are feature points which are obtained with most accuracy, and the fact that in tests it proved to be usable without any further scaling. The conjunction removes estimates of boundaries

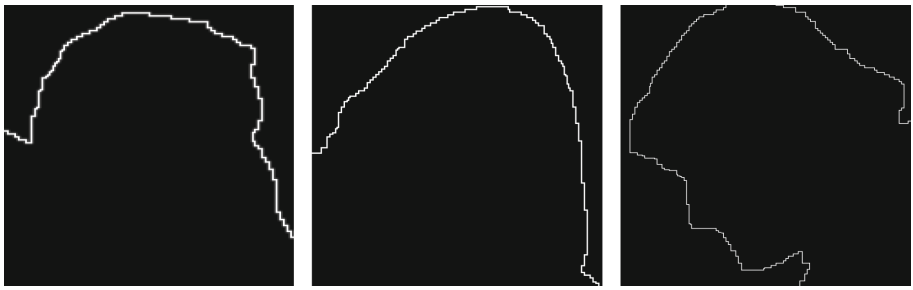

Fig. 5. Detected head boundaries

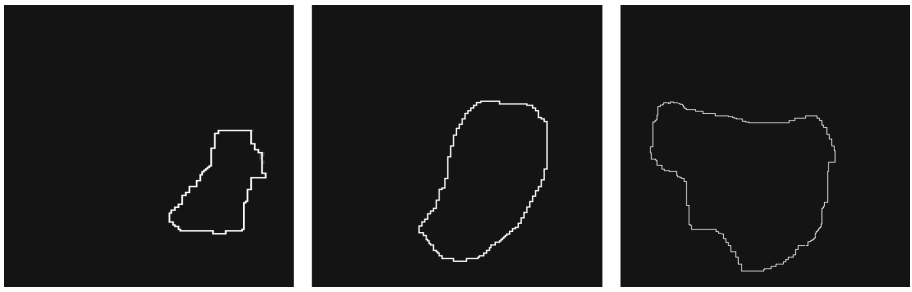

Fig. 6. Detected face boundaries

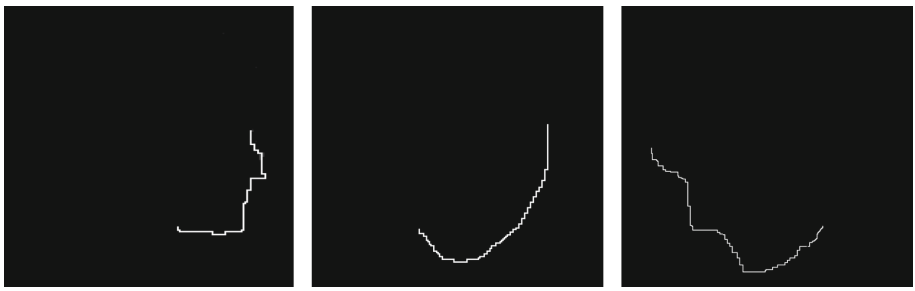

Fig. 7. Fragments of boundaries used for reconstruction

that are based on different features than those presented in projected model, e.g. line of hair above the forehead, hair on the side of face, shoulders and neck. Since the disjunction of head and face boundary was used, for each branch inside of conjunction, the decision is made to use the one with smallest average distance to *projected boundary*, thus obtaining *image boundary*. With both boundaries found, the aforementioned method [8] is used; maxima of distances between boundaries are found and used to find RBF interpolation that will transform mesh so that *projected boundary* coincides with *image boundary* thus reconstructing parts of 3D mesh that are not reconstructed well using found feature points.

Fig. 8. Reconstructed model

3 Conclusion

The approach presented in this paper is a composite solution for reconstruction of facial mesh from video-based surveillance systems' data. Presented solution produces satisfactory results in case of videos, where face is fully visible from different angles, however it still requires further work in case of low quality data (as in Figs. 4, 5, 6 and 7), from which the presented reconstructed model (Fig. 8) was built. Further work will be focused mostly on obtaining additional feature points through boundaries of facial features. Still though, results are promising enough to consider this approach worth further work so that it could be used in assisting eyewitnesses in suspect identification.

Acknowledgments. This work has been supported by the National Centre for Research and Development (project UOD-DEM-1-183/001 "Intelligent video analysis system for behavior and event recognition in surveillance networks")

References

1. Kulbacki, M., Segen, J., Wereszczyński, K., Gudyś, A.: VMASS: massive dataset of multi-camera video for learning, classification and recognition of human actions. In: Nguyen, N.T., Attachoo, B., Trawiński, B., Somboonviwat, K. (eds.) ACIIDS 2014, Part II. LNCS, vol. 8398, pp. 565–574. Springer, Heidelberg (2014)
2. Gudyś, A., Wereszczyński, K., Segen, J., Kulbacki, M., Drabik, A.: Camera calibration and navigation in networks of rotating cameras. In: Nguyen, N.T., Trawiński, B., Kosala, R. (eds.) ACIIDS 2015. LNCS, vol. 9012, pp. 237–247. Springer, Heidelberg (2015)
3. Hartley, R.I.: Self-calibration from multiple views with a rotating camera. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 800, pp. 471–478. Springer, Heidelberg (1994)

4. Cai, Y., Medioni, G.: Exploring context information for inter-camera multiple target tracking. In: Applications of Computer Vision (WACV) 2014, pp. 761–768. IEEE (2014)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition, pp. 886–893. IEEE (2005)
6. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
7. Tong, Y., Wang, Y., Zhu, Z., Qiang, J.: Robust facial feature tracking under varying face pose and facial expression. Pattern Recogn. **40**(11), 3195–3208 (2007)
8. Roussel, R., Gagalowicz, A.: Realistic face reconstruction from uncalibrated images. In: VMV 2004, pp. 141–149. Aka GmbH (2004)
9. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**(4), 308–313 (1965)
10. DeMenthon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. In: Sandini, G. (ed.) ECCV 1992. LNCS, vol. 588, pp. 335–345. Springer, Heidelberg (1992)
11. Shih, F.Y., Chuang, C.-F.: Automatic extraction of head and face boundaries and facial features. Inf. Sci. **158**, 117–130 (2004)
12. Farkas, L.G., Katic, M.J., Forrest, C.R.: International anthropometric study of facial morphology in various ethnic groups/races. J. Craniofac. Surg. **16**(4), 615–646 (2005)