Face Recognition Based on Sequence of Images

Jacek Komorowski and Przemyslaw Rokita

Abstract. This paper presents a face recognition method based on a sequence of images. Face shape is reconstructed from images using a combination of structure-from-motion and multi-view stereo methods. The reconstructed 3D face model is compared against models held in a gallery. The novel element in the presented approach is the fact, that the reconstruction is based only on input images and doesn't require a generic, deformable face model. Experimental verification of the proposed method is also included.

1   Introduction

Three dimensional face recognition is an active and growing field of research [1] [2]. Using spatial information allows to mitigate some of the problems faced by methods based solely on visual information. 3D face recognition methods are less dependent on face pose and lighting variations. One of the barriers to a mass deployment of this technology is a difficulty with a face shape acqusition. Active vision techniques, such as laser scanning, are not appropriate for practical usage. Laser scanners are rather large, expensive and may be damaging to human eyes. Alternative, passive techniques, such as stereovision, multi-view stereo or structure-form-motion, are not very well suited for human face shape reconstruction. These methods are based on finding corresponding points on multiple images, that is points which are projections of the same scene point. Human skin has a relatively homogeneous texture which makes an automatic matching a difficult task.

Majority of methods which use passive vision techniques for face shape reconstruction, either uses complex image acqusion setup (e.g. set of 5 cameras [11]) or utilises a generic, deformable face models (e.g. [3]). Complex camera setups

Jacek Komorowski · Przemyslaw Rokita
Military University of Technology, Faculty of Cybernetics, Warsaw, Poland
e-mail: jac99@o2.pl,p.rokita@ii.pw.edu.pl

complicate practical deployment. Model-based approach is criticized [4], that it doesn't allow to model subtle details important for accurate face recognition, as reconstruction result are limited by a model parameter space.

The method presented in this paper uses a sequence of images from a single camera. Therefore it's easy to use as there's no need for a complicated equipment. Additionally it's based solely on input images and doesn't require a generic face model. Multi-view stereo algorithms can be used to reconstruct a 3D object model from a set or sequence of images taken from multiple viewpoint. Over the last years a significant progress was made in this area and a number of high-quality algorithms were developed. Best methods reviewed in [10] can deal with very demanding scenarios, where input images depict objects with little texture, containing few points which can be automatically matched across multiple images. For very demanding DinoRing[1] test set, containing images of a plaster dinosaur taken from multiple viewpoints, the best algorithms surveyed in [10] were able to reconstruct over 90% of the object surface with error below 0.4 mm. Unfortunately multi-view stereo algorithms assume that all images are fully calibrated, that is both intrinsic (camera focal length, distortion coefficients) and extrinsic (camera pose) parameters for each image are known. Such algorithms cannot be used when a sequence contains images of an object moving freely in front of the camera. Intrinsic camera parameters are fixed, and can be estimated with a prior calibration. But extrinsic parameters are different for each image and cannot be easily estimated. To use some high-quality multiview-stereo algorithm for face shape reconstruction from a sequence of images, extrinsic parameters for each image in the sequence must be estimated.

2   Details of the Method

This section describes details of our face recognition method. The method is based on a sequence of images from a monocular camera. It's assumed that a person sits in front of the camera and is asked to rotate his head left and right. An exemplary input sequence is depicted on Fig. 1.

Fig. 1 Exemplary input sequence (5 from 80 images)

---

[1] http://vision.middlebury.edu/mview/data
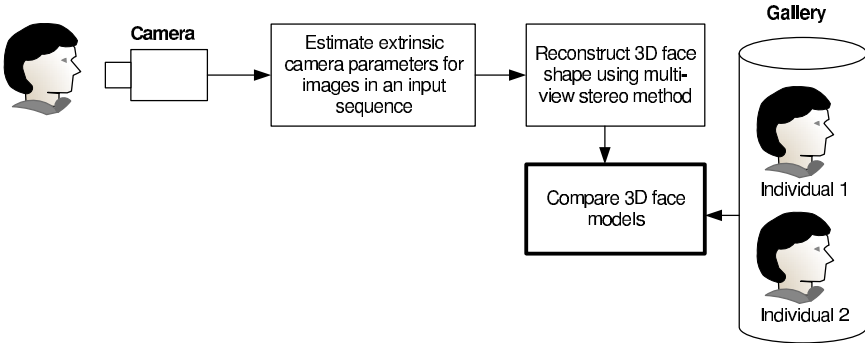
Main steps of the presented method are depicted on Fig. 2.

Fig. 2 Recognition system concept

Step 1

Extrinsic camera parameters (rotation matrix **R** and translation vector **T**) are estimated for each image in the sequence. This is done using a method developed by authors and described in [7] and [8]. The method is designed to work well with demanding scenarios, where input images contain little texture. It doesn't use a generic, deformable face model and is based solely on input data. Results of this step are depicted on Fig. 3.

Step 2

Once camera extrinsic parameters are estimated, any multi-view stereo algorithm can be used to reconstruct a 3D face shape. In our implementation a patch-based multi-view stereo method PMVS [5][2] was used. An input to the PMVS algorithm is a sequence of images and estimated camera extrinsic parameters. The output is a cloud of oriented points (see Fig. 4).

Step 3

Face model reconstructed from an image sequence is compared with models in a gallery. Distance between point clouds is used as a similarity measure between two face models. Distance between two point clouds is defined as an average Euclidean distance between each point from the first model to the closest point in the second model. Two face models usually do not fully overlap. Due to differences in input

---

[2] http://grail.cs.washington.edu/software/pmvs/
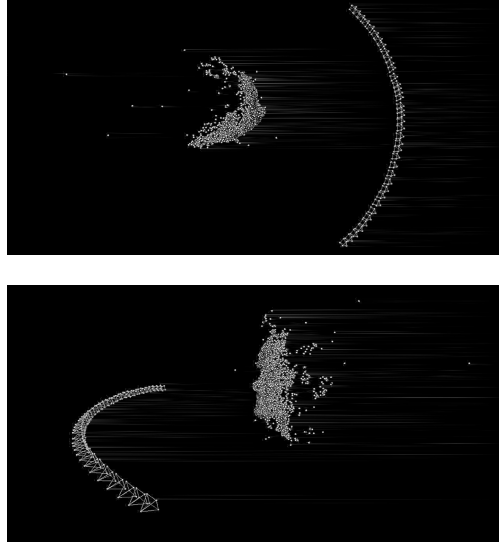
Fig. 3 Estimated camera poses (pyramids) and a sparse face model (point cloud) based on a sequence of images from Fig. 1
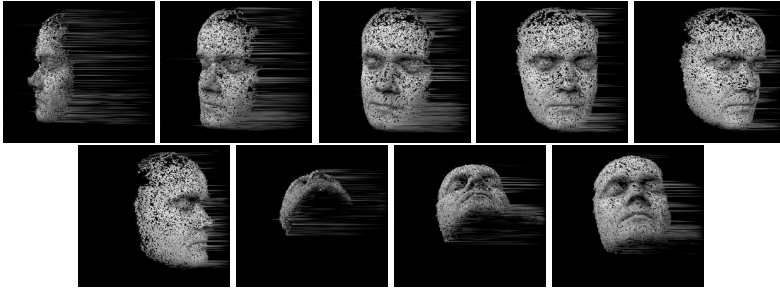
Fig. 4 Face reconstruction results based on a sequence from Fig. 1

sequences[3] one cloud may contain regions from a reconstructed object surface, not presented in the second model. To deal with this problem a relatively simple heuristic is used. A median distance between each point from the first model and the closest point in the second model is calculated, and points further away than some small multiple of the median are discarded. Formal definition of the distance metric used to compare 2 point clouds is as follows:

Let $\mathscr{C}_1 \subset \mathbb{R}^3$ and $\mathscr{C}_2 \subset \mathbb{R}^3$ denote two clouds consisting of points in 3D Cartesian space. $d\left(p, \mathscr{C}\right)$ denotes a distance of a point $d \in \mathbb{R}^3$ from the cloud $\mathscr{C} \subset \mathbb{R}^3$, defined as:

---

[3] E.g. different maximum face rotation angle.

$$d(p, \mathscr{C}) = \min_{p' \in \mathscr{C}} \left\| p' - p \right\| , \tag{1}$$

where $\left\| p' - p \right\|$ is an Euclidean distance between points p i p'. Distance between point cloud $\mathscr{C}_1$ and $\mathscr{C}_2$ with a threshold k is defined as:

$$d_k(\mathscr{C}_1, \mathscr{C}_2) = \frac{1}{|\mathscr{C}_1 \setminus \mathscr{O}_k|} \sum_{p \in \mathscr{C}_1 \setminus \mathscr{O}_k} d(p, \mathscr{C}_2) , \tag{2}$$

where $\mathscr{O}_k$ is a set of points in a cloud $\mathscr{C}_1$ not having close neighbours in a cloud $\mathscr{C}_2$, defined as:

$$\mathscr{O}_k = \{ p \in \mathscr{C}_1 \mid d(p, \mathscr{C}_2) > km \} , \tag{3}$$

where m is a median distance between each point from the first cloud and the closest point from the second cloud. In the implementation of the presented method threshold $k = 4$ was chosen.

Two face models being compared may have a different scale and orientation. Scale difference is caused by the fact, that extrinsic parameters can be estimated for a sequence of images only up to an unknown scale factor. Thus a metric reconstruction is also possible up to a scale factor. Orientation may be different because reconstructed head pose is aligned with the head pose on the first image. In order to calculate a distance between two point clouds, they must be aligned first. We use a variant of the popular ICP[4] [9] algorithm, which can find a rigid body transformation aligning two point clouds.

Let C_s denotes a source point cloud and C_d a destination point cloud. Our modified version of ICP method has the following steps:

1. Compute centroids of a source and destination cloud

   a. $\bar{c}_s = \left(\sum_{p \in \mathscr{C}_s} p \right) / |\mathscr{C}_s|$
   b. $\bar{c}_d = \left(\sum_{q \in \mathscr{C}_d} q \right) / |\mathscr{C}_d|$

2. Scale a source point cloud to match a destination cloud scale using a formula from [6]:

   a. Compute scaling factor: $scale = \sqrt{\frac{\sum_{q \in \mathscr{C}_d} \|q_i - \bar{c}_d\|^2}{\sum_{p \in \mathscr{C}_s} \|p_i - \bar{c}_s\|^2}}$
   b. Multiply coordinates of points in $\mathscr{C}_s$ by $scale$

3. Align centroid of a source point cloud with a centroid of a destination cloud

   a. Translate all point in C_s by a vector $\bar{c}_d$ - $\bar{c}_s$.

4. Choose a random sample $\mathscr{S} = \{p_i\}$ of s points from a source cloud C_s
5. Match each point from a sample S with the closest point in a destination cloud C_d. Let $\mathscr{M} = \{(p_i, q_i)\}$ denotes a set of corresponding point.

---

[4] *ang.* Iterative Closest Point.

6. Remove outliers from $\mathscr{M}$, that is remove pairs (p_i, q_i) for which $|d_i - q_i| > km$, where m is a median distance between pairs of corresponding points in $\mathscr{M}$, and k is a small integer[5].

7. Find a rigid body transformation (rotation matrix $\mathbf{R}$ and translation vector $\mathbf{T}$) minimizing error metric $E(\mathbf{R}, \mathbf{T})$ and apply the transformation on a source point cloud $\mathscr{C}_s$

8. If number of iterations $< N$, go to point 4 else terminate the algorithm

Algorithm parametrization and error metric E were chosen experimentally to achieve good convergence and a reasonable running time. Sample size M is set to 500 (out of app. 40'000 points in clouds) and number of iterations $N = 15$, as it was verified that larger values increase running time but do not improve convergence. As an error metric E, a point-to-plane error metric is chosen as it gives much faster convergence than a classic point-to-point error metric. Point-to-plane error metric is given by the formula [9]:

$$E_{\text{point}-\text{to}-\text{plane}}(\mathbf{R}, \mathbf{T}) = \sum_i ((\mathbf{R}p_i + \mathbf{T} - q_i) \cdot n_i)^2 \; , \tag{4}$$

where n_i is a normal to the destination cloud surface at point q_i.

3 Experiments

This section presents results of an experimental verification of accuracy of the face recognition method presented in this paper. Test database built by authors contains 81 image sequences of 27 individuals, 3 sequences per one person. Images were acquired with Point Grey Chameleon camera[6] with 800x600 pixels resolution. In each sequence a persons sitting in front of a camera is asked to rotate his head right and left. Exemplary sequences are depicted on Fig. 5. The database was split into 2 parts: 27 image sequences (1 per each individual) were used to build a gallery, 54 sequences (2 per each individual) were used to build a test set.

Error metrics

Face recognition system can be used to perform 2 tasks: verification and identification. Verification is a task where the biometric system attempts to confirm an individual's claimed identity. 2 error metrics are used to assess accuracy of an identity verification task: FAR[7] and FRR[8]. FAR is defined as a ratio of a number of attempts when an identity was falsely positively verified to a number of all attempts. FRR is

---

[5] In implementation $k = 4$ was chosen.

[6] http://www.ptgrey.com/products/chameleon/
    chameleon_usb_camera.asp

[7] False Acceptance Ratio.

[8] False Rejection Ratio.

Fig. 5 Exemplery sequences from a test database

defined as a ratio of a number of attempts when an identity was falsely negatively verified to a number of all attempts.

Identification is a task where biometric system searches a gallery for a reference matching submitted biometric sample, and if found, returns a corresponding identity. Accuracy of identification tasks is measured with a CMC[9] curve. CMC is a function of a recognition rate as a number of best n-maches considered. For a given n, recognition rate is a ratio of attempts when a chosen individual from a test set was among n closest matches in the gallery to number of all attempts. Clearly, when n equals to the number of individuals in the gallery, recognition rate is equal to one.
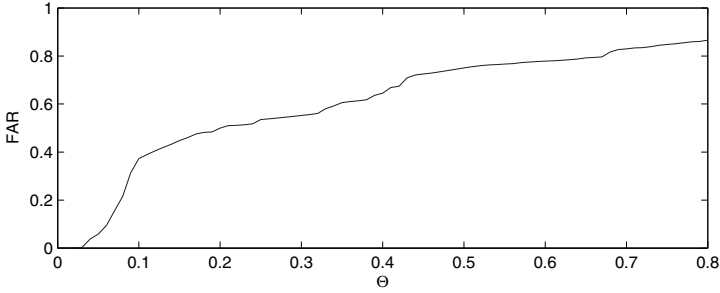
Experiment 1

In this experiment accuracy of identity verification scenario was tested. Each sequence from a test set was used to reconstruct a 3D face model which was matched against each face model in the gallery. If the distance between face model from a test set and a face model from a gallery was below a threshold Θ the identity was positively verified. Otherwise identity was negatively verified.

Both FAR and FRR are dependent on threshold Θ. When it's increased, more distant faces are identified as belonging to the same individual thus leading to FAR
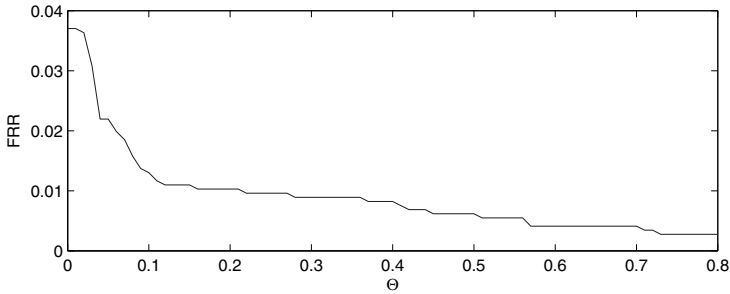
---

[9] Cummulative Match Characteristics.

increase and FRR decrease. Fig. 6 depicts values of FAR as a function of a threshold Θ. Fig. 7 shows values of FRR as a function of a threshold Θ.

Fig. 6 FAR as a function of a threshold Θ

Fig. 7 FRR as a function of a threshold Θ

The trade-off between FAR and FRR rates is expressed using ROC[10] curve and is shown on Fig. 8. ERR[11], that is a rate at which FAR = FRR is equal to 0.025 and is a rather low value. It means that in 2.5% of attempts identity was falsely positively verified and in 2.5% of attempts identify was falsely negatively verified.

Experiment 2

In this experiment identification in a closed-set scenario was tested, as each individual from a test set was present in the gallery. Each sequence from a test set was used to reconstruct a 3D face model which was matched against each face model in the gallery. Models with the closest distance were declared as a match.

---

[10] Receiver Operating Characteristic.
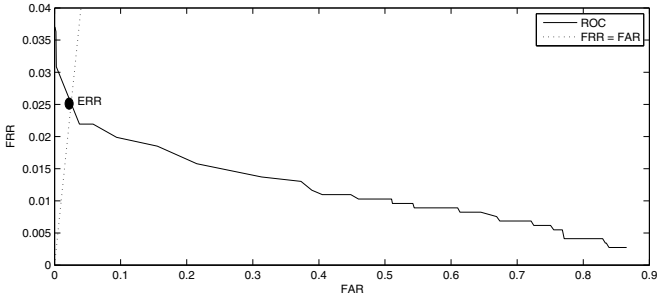
[11] Equal Error Rate.

Fig. 8 ROC curve and ERR point

Fig. 9 shows resultant CMC curve. When finding a single, best match in the gallery ($n = 1$) for each individual from a test set, the method achieved almost 75% accuracy. If considering 5 best matches in the gallery ($n = 5$), over 90% accuracy was achieved.
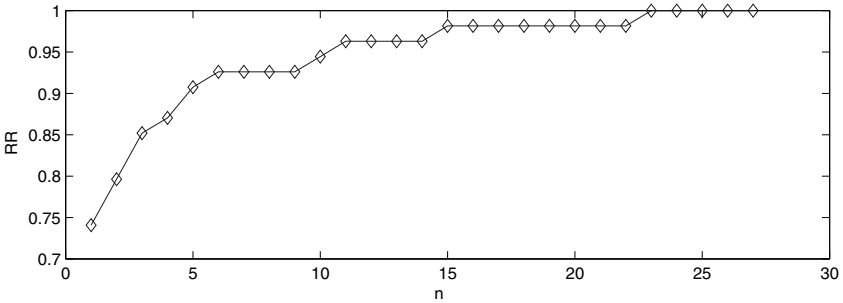

Fig. 9 CMC curve

4 Conclusions and Future Work

The presented method allows to achieve a reasonably good face recognition accuracy. Although the results should be taken with care, as they were obtained using a relative small test database. To ensure validity of the proposed approach, experiments using much larger test database should be done. For face recognition a relatively simple approach is used, based on direct comparison of two point clouds. It's worth to investigate more advanced approaches, e.g. based on comparison of a local characteristics such as nose profile, or a relative position of eyes, nose and lips. It must be noted that recognition is based only on spatial information and 2D information (texture) is not used. Combining 2 modalities (shape and texture) may allow to achieve better recognition rates.

Face reconstruction method presented in this paper consists of two separate and distinct steps. Sparse point cloud build during the process of estimating extrinsic parameters is discarded, and only extrinsic paratmers are passed to the second step (multi view stereo reconstruction). Potentially 3D points from a sparse point cloud created in the first step can be used to initialise multi-view stereo reconstruction process.

References

1. Bowyer, K., Chang, K., Flynn, P.: A survey of 3d and multi-modal 3d+2d face recognition. Computer Vision and Image Understanding 101(1), 1–15 (2006)
2. Chang, K., Bowyer, K., Flynn, P.: An evaluation of multi-modal 2d+3d face biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 619–624 (2005)
3. Cheng, C., Lai, S.: An integrated approach to 3d face model reconstruction from video. In: IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (2001)
4. Fidaleo, D., Medioni, G.: Model-assisted 3d face reconstruction from video. In: Proceedings of the 3rd International Conference on Analysis and Modeling of Faces and Gestures (2007)
5. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(8), 1362–1376 (2010)
6. Horn, B.: Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America 4(4), 629–642 (1987)
7. Komorowski, J., Rokita, P.: Camera pose estimation from sequence of calibrated images. Advances in Intelligent Systems and Computing, Image Processing and Communication Challenges 4 184, 101–110 (2012)
8. Komorowski, J., Rokita, P.: Extrinsic camera calibration method and its performance evaluation. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2012. LNCS, vol. 7594, pp. 129–138. Springer, Heidelberg (2012)
9. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings of the Third International Conference on 3D Digital Imaging and Modeling, pp. 145–152 (2001)
10. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2006)
11. Spreeuwers, L.: Multi-view passive acquisition device for 3d face recognition. Proceedings of the Special Interest Group on Biometrics and Electronic Signatures 137, 13–24 (2008)