4D Unconstrained Real-time Face Recognition Using a Commodity Depthh Camera

Schimbinschi, Florin; Wiering, Marco; Mohan, R.E.; Sheba, J.K.

Published in:
7th IEEE Conference on Industrial Electronics and Applications

Publication date:
2012

4D unconstrained real-time face recognition using a commodity depth camera

Florin Schimbinschi*, Marco Wiering*, Rajesh Elara MOHAN†, Jaichandar Kulandaidaasan SHEBA‡
*University of Groningen
Nijenborgh 9, 9747 AG, The Netherlands
Email: aiflorin@ai.rug.nl, m.a.wiering@rug.nl
†Singapore University of Technology and Design
20 Dover Road, Singapore 138682
Email: rajeshelara@sutd.edu.sg
‡Singapore Polytechnic
500 Dover Road, Singapore 139651
Email: jai@sp.edu.sg

***Abstract*—Robust unconstrained real-time face recognition still remains a challenge today. The recent addition to the market of lightweight commodity depth sensors brings new possibilities for human-machine interaction and therefore face recognition.**

This article accompanies the reader through a succinct survey of the current literature on face recognition in general and 3D face recognition using depth sensors in particular. Consequent to the assessment of experiments performed using implementations of the most established algorithms, it can be concluded that the majority are biased towards qualitative performance and are lacking in speed.

A novel method which uses noisy data from such a commodity sensor to build dynamic internal representations of faces is proposed. Distances to a surface normal to the face are measured in real-time and used as input to a specific type of recurrent neural network, namely long short-term memory. This enables the prediction of facial structure in linear time and also increases robustness towards partial occlusions.

I. Introduction

A. General Aspects

1) Human computer interaction: The purpose of Artificial Intelligence (AI) is to create technology to augment our lives, since all man's inventions are in essence extensions of the mind and body. However, most modern human-computer interaction platforms are intrusive and do not encourage naturally flowing movement.

Instead of the user unconsciously and naturally training the AI, it is often the case that the machine is training the user by explicitly requesting him to perform certain actions or to keep certain body postures, etc., in order to eliminate variance and make the problem less difficult.

2) Typical challenges: Face recognition is one of the most studied topics in computer vision and one of the most successful applications of image analysis, pattern recognition and machine learning. Although there are many successful applications already, face recognition is still a challenge when compared to pattern recognition problems such as optical character recognition (OCR) [1]. This is due to the variance in face images, such as viewpoint, illumination, expression, occlusion, makeup and even aging.

3) Commodity sensors: The recent introduction on the market of the '2.5D' Kinect™ sensor, opens a new era for human-computer interaction and implicitly for face recognition. An interesting aspect of the Kinect sensor is that it can be used to compare 3D with 2D face recognition algorithms properly, on the same dataset, (and the combination of both), since it captures texture information (RGB images) in parallel with depth data, which are always aligned and synchronized.

4) Current commercial systems: Kinect Identity [2], a key component of Microsoft's Kinect™ for the XBox 360®, is a good example of an intrusive system. Face recognition, clothing color tracking, and height estimation are used within a multi-modal framework to achieve the goal of recognizing and tracking player identity.

While the additional information is quite useful and makes the problem less difficult, this implies that the recognition can not be performed accurately if most of the player's body is not in full view, thus the platform it can not deal with partial occlusions. Since the player has to comply with instructions regarding to the posture and distance to the sensor, it is evident that the Kinect Identity can not perform well in unconstrained, cluttered environments and is quite intrusive.

Regarding the face recognition component of the framework, it is not mentioned [2] what kind of technology is used for the discrimination of players: 2D, 3D or a combination of both.

B. Operational Goals: Time, Space, Features

1) Off-line: The Kinect Identity example is part of a process where face recognition is performed instantaneously, typically referred to as 'real-time' or 'on-line'. However, this is not always the case.

If one has to wait for the result, which usually happens when querying a very large database, then it is referred to as 'off-line'. This flavor of face recognition is usually depicted

in motion-pictures: the user inputs a compact disc with a face image in the computer, then the software starts searching for the identity. The processing takes an undetermined amount of time.

The main functional difference for off-line face recognition not only refers to the retrieval speed of the identity but also to the speed it takes to register a new person. Most off-line operating algorithms [3] require that the whole (new) dataset be presented to the AI each time a new person registers. This 'forgetting' behavior is often a problem with many machine learning algorithms.

2) Real-time (on-line): Real-time face recognition on the other hand, is more complex, and can be divided into two sub-categories. The first would be access control [4], where only the answer if a person's identity is acknowledged or not is required.

The quantized encrypted identification information can be gathered beforehand and stored on an external device such as a magnetic card and compared with the algorithm's output from processing the live camera captured image of the person's face. This leads to a one on one matching process which is less computationally expensive than a one to many comparison, nevertheless the algorithm still has to proceed in real-time.

Furthermore, in this circumstance, most of the variance can be explicitly controlled (illumination, pose, etc) since the user has to stop and present the ID. Instructing the user where to look or how to stand considerably makes the problem less challenging since it eliminates much of the variance in the input data.

This type of 'enrolling' can be observed in most of the 3D sensor face recognition literature experiments and facial expression recognition [5], [6], [7], [8], [9] algorithms, where the user has to perform certain head movements and is commonly confined to sit in the same position for some time until enough data is gathered from the sensor.

3) Freedom of movement: By unconstrained face recognition it is understood that both the identification and registration processes are not intrusive and are as transparent as possible to the user. This implies that there is no direct interaction for the purpose of face recognition. Nevertheless, the user could be asked for his name, or a confirmation at any given time, to link the stored internal model with an identity.

The ideal situation would be to gather 'no-name' faces into clusters, then gradually associate these with identities; in other words, to perform unsupervised, or in actual practice, semi-supervised learning.

The most challenging face recognition task is unconstrained biometric authentication [10] which implies identifying or registering a user in 'real-time' without the need of calibration instructions. This means that the algorithm usually has to deal with incomplete information, a situation which is typical in unpredictable environments.

4) Integrating shape over time: 4D face recognition refers here to the use of depth data - 3D point clouds - and their position in time (3D XYZ + 1D time) to compute the shape of faces and predict future changes due to facial expressions.

3D face recognition can be achieved by integrating the noisy point cloud data to obtain detailed models of the face. A detailed description of this type of model aggregation is presented in [11]. 3D registration has also been successfully performed by 'in-hand' object modeling [12]. However, it requires users to manually align the reconstructed object to displayed scans for reinitialization, which applied for face recognition would violate the constraint of non intrusiveness. Moreover, the final model is integrated using off-line processing, which implies that the user would not be identifiable in the immediate future from exposure to the system and would most probably cause an unpleasant delay in the natural interaction with the AI.

There are 4D photo-geometric [13] algorithms that combine the depth data with gray-scale intensity information to obtain a higher fidelity in the internal representations.

C. Theoretical Background

1) The stages of the process: The process of face recognition normally involves three main stages: detection, feature extraction (dimensionality reduction) and classification (storage), where localization and normalization [14], [15], [16] (detection and alignment) [17] are the preprocessing steps before face recognition [10] (feature extraction and recognition).

2) 2D face recognition: Thus far most of the face recognition research has been focused on 2D data from typical digital images [18], [19]. The typical combination of algorithms is the use of Eigenfaces or Fisherfaces with principal component analysis (PCA) [20] or linear discriminant analysis (LDA) [21]. Although the results are satisfactory, they lack in robustness, since they all suffer from the common 2D drawbacks due to high variation in images. Video-based face recognition [22], [23], [24], [25] research takes things further by adding time as a checking factor to eliminate some of the uncertainty. Frame-based face recognition methods using temporal voting schemes is also a common approach. This usually happens with additional various constraints which bring improvements but also add complexity.

3) Pseudo 3D: 3D face recognition is another kind of approach, which is primarily of two sub-types. The most common is in fact pseudo-3D [26], since 3D models which contain shape [27] as well as texture are inferred from 2D images [28], [29]. These are very successful in current commercial systems. Face.com Inc. develops the API that runs on Facebook ™. However, it should be noted that it can only operate on specialized computing clusters since the algorithm is very computationally expensive [30], [31], [32], [33].

4) 3D using depth data: Conversely, in true 3D face recognition [8], [16], [34], the input data contains actual depth information, the precision is higher, however the equipment is often bulky and usually too expensive. Three dimensional face recognition has the potential to achieve better accuracy than it's 2D counterpart by measuring the geometry of rigid features on the face. This avoids most the problems of traditional 2D face recognition algorithms as change in lighting, different facial expressions, make-up and head orientation.

5) Hybrid algorithms: There have also been proposed such 'hybrid' algorithms that combine results obtained independently from the 2D texture and 3D depth data, [35]. In [7] 2D face detection is used to align facial key-points to a 3D morphable model (3DMM) [36], [9] obtained from a commodity range sensor.

Other approaches blend together 2D and 3D data before any further processing is made [13], where 4D HOG descriptors are used to compute rank-1 (top match) identification rate on a comprehensive time-of-flight dataset (26 subjects, 364 facial images). However, they note that this method is still susceptible to variations in illumination, occlusions and facial makeup. This is to be expected since 2D data is used. It should also be noted that a high resolution range sensor was used in this case, which implies that the data was considerably less noisy than the latest commodity depth sensors.

II. 3D REGISTRATION ALGORITHMS

A. Overview

3D registration [37] refers to the goal of alignment of two sets of 3D points, according to their similarities. This section contains a discussion about the current approaches to the various processes involved in face recognition using depth sensors. Even though most of the current systems report good qualitative performance, it is often the case that their use in real-time applications is not realistic since the whole process either requires intense processing of the sensor data or imposes constraints on the user.

Face recognition using commodity depth cameras has not been thoroughly investigated yet. The current surveys [38], [18] do not cover the extra challenges involved with the low resolution, noisy sensor data. A thorough qualitative and quantitative evaluation of the possibilities of face recognition with a commodity depth sensor is required, in light of the latest 3D paradigms.

Apparently, there has been more successful work done around expression recognition [5], where PCA and 3DMM are used to capture a finite set of expressions (also called action units: AU) based on face movement. However, instead of trying to identify unique features in each individual's face, the goal in expression recognition is to 'filter out' unique features of faces in order to better align the 3DMM with the face and record face movement.

Similar techniques have also been used for virtual face model animation [39], [40] where instead of giving a semantic meaning to the expression, virtual 'puppets' are animated, sometimes even in real-time. Almost all of the facial expression recognition research is focused on capturing the similar inter-related features and not the intra-related differences, as required in face recognition.

Regardless if the task is face expression recognition, animation or face recognition, most algorithms require the active involvement of the user for the initial calibration, and have only been tested in laboratory conditions. It can be observed from the papers in this section that, when the qualitative results are satisfactory good, the speed is 'strangely' not reported and vice-versa. Furthermore, quantitative experiments are often not performed.

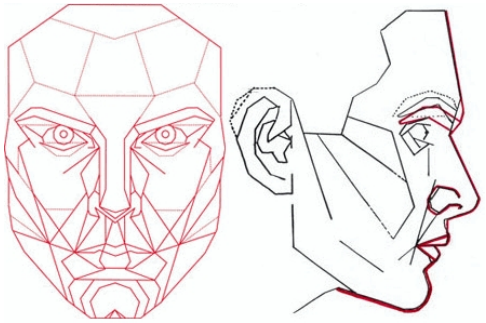

Fig. 1. A general 3D face template can be aligned in real-time to the sensor data using the iterative closest point algorithm on motion segmented and down-sampled areas for tracking the face in linear time. The template will consist of only the upper part of the face, excluding the jaw area where there is too much variance in shape.

B. ICP

A general registration algorithm, it is often the choice for determining the location of an object. A generic 3D average face template (Fig. 1) can be used to perform euclidean 3D alignment of the target to the sensor data, usually by means of the non-rigid iterative closest point (ICP) [37], [41], [42] matching algorithm. While this approach is quite fast, it may lack in precision and it is not robust towards sudden moves of the head. Another problem with ICP is convergence, since it can sometimes be stuck in local minima. However, it is often more important how an algorithm is used, than its reported performance.

C. Faceprints

An alternative to ICP is the use of spherical intersection profiles (faceprints) [43] for both face tracking and recognition. In [8] the dataset was comprised from faceprints taken from ten subjects positioned approximately one meter from the camera, at various angles. They report a comparison rate of almost 150,000 per second, which is appropriate for real-time systems. However, it should be noted that the raw data is preprocessed to isolate the '3D points comprising the subjects head', in other words, the whole data is segmented according to a precomputed distance between the user and the sensor. It unfortunately lacks in qualitative and quantitative results and imposes too many constraints. Furthermore, experiments related to robustness relative to the variation in face expression have not been reported.

An alignment accuracy comparison between faceprints (here called IRAD) and ICP is considered in [6]. The nose tip is selected as the main discrimination criteria in a radial basis function (RBF) [44], [45] model of the facial surface. Even though the faceprint method seems to be more accurate, the time needed to compute the alignment is not reported. This algorithm, not only uses the same procedure as described above, but also changes the actual depth information that it captured. The algorithm successfully normalizes the pose

of 3D faces at rates of 99%. However, the experiment is performed in the same constrained conditions as the previously described paragraph and processing speed is not reported.

D. Local Feature Histograms

The main motivation for histogram matching is low computational cost. As described in [46], a template can be used to assign similar shape patches to the same histogram cells, which eliminates the need to solve correspondence problems by computing a comparison measure. This results in a recognition rate of 89% with only 128 histogram cells. However, the experiment is performed with on a static database of objects which was altered synthetically.

For live tracking of a subject's face one would need to compute histogram features for each frame of the target point cloud data for matching. This would not result in fast recognition rates. Nevertheless, with combined techniques inspired from video based face recognition, this technique might improve significantly if the feature re-computation task can be avoided. Face recognition is attempted with a relative success of 78% in [47] where they use multi layer perceptrons (MLP) [48] and the same type of features. However, the experiments isolate the face and lack a proper quantitative measure, since only three subjects are used.

E. Model aggregation

One strategy for face recognition would be to aggregate face data over time in order to build detailed models of the face. Poisson reconstruction [49] is an already established method which stands at the core for many such aggregation algorithms. While is very useful for 3D model capturing [11] and animation [40], it would be quite cumbersome and computationally expensive to be used for real-time face recognition, since the models contain too much information than it is needed to actually differentiate between faces.

A strategy has been described in [12] which is used to aggregate object models using ICP. However, the camera as well as the reconstructed object were static and the processing was done off-line. In the case of face recognition the intention however is to eliminate these constraints and allow the user to move freely in the environment.

Since the head and therefore the face would be moving in 3D space, temporal filtering would also increase the accuracy and precision. An example is the work done in [11] which is a robust algorithm for geometry and motion reconstruction of dynamic shapes. This is a good example of how temporal filtering adapts to the speed of motion. The robust template tracking based on an adaptive deformation model is the key component in the described algorithm. This method of detail synthesis exploits the accurate registration to aggregate and propagates geometric detail into occluded regions.

F. Combined 2D

Active appearance models (AAM) [35] that combine 2D with 3D information have also been evaluated, but they suffer from the same drawbacks of 2D texture data, such as illumination invariance and usually fail to track the object if it performs sudden moves, which often cause blurring.

Combining 3D deformable models with active shape models (ASM) [9] has also been considered. However, ASMs are unstable in the presence of outliers in the training set. Another similar approach [50] uses similar features to track faces in real time. Again, being a 2D method, it suffers from the same drawbacks as all 2D methods.

As reported in a survey [38], appearance based algorithms may suffer from insufficient generalization ability due to lighting and texture variations, while feature based algorithms may perform poorly due to the lack of semantic features or the occlusion of profile poses, etc.

A problem with measuring textures in faces is that they are not necessarily connected to the underlying shape [42]. For example the position of the eyebrows relative to the underlying bone structure varies strongly between subjects.

Even though two dimensional data is cumbersome to work with, most of the algorithms used for 2D processing can be successfully re-applied on 3D depth data.

III. A different technique

The above presented methods focus exaggeratedly on the qualitative aspects of the sub-processes of face recognition. Within each stage, there is an emphasis to achieve greater accuracy and the goal seldom considers speed or the freedom of movement of the user. While accuracy is ultimately important, this does not imply that the whole ensemble of algorithms needs to be precise. By focusing on a relatively loose face alignment, which is less computationally expensive, the focus can move towards the goal of absorbing the fluctuations in pose variance within the face model. Of course, this implies that the variations in pose alignment are not substantial. The trade-off for accuracy to speed is too biased, in the sense that for a minute amount of increase in accuracy, the computing complexity increases far more. Since these speed versus accuracy experiments have not been performed yet, it is an intention of the authors to perform such a study in the near future.

A. Face tracking

1) Localization: The permanent location of the head and more importantly the face is required at all times since, for the purpose of unconstrained real-time face recognition, it is necessary to update the internal models with new data whenever the system reports confidence levels that are below a predefined threshold.

By using ICP, the precise location of the face (along with pan, tilt, roll) is known. There is no need to produce changes in the raw data captured with the sensor in order to align the face data as described in [6].

A variant of the standard non-rigid ICP, EM-ICP [51] that uses Expectation Maximization proves to be a perfect solution for the purpose of multiscale non-rigid object tracking. The reported experiments on real data reveal an improvement of the performances of EM-ICP in terms of robustness (a factor of 3 to 4) and speed (a factor 10 to 20) with a similar accuracy.

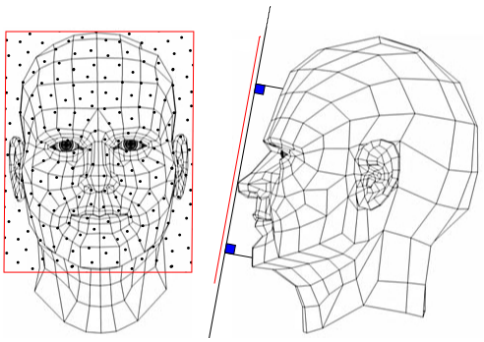

Fig. 2. Once the face is found, the normal surface plane through the nose can be useed to project all the distances to the face depth data.

The use of faceprints for the purpose of face tracking has been proven to be fast and accurate, however restrictions in pose and distance make is limited to be used only when the user is relatively close to the camera [8].

A weighted combination of motion segmentation [52] and head tracking [53], [54] can be used to further speed up the tracking process by designating areas with low probability of containing a head / face. Since the user will also be standing still this is not an entirely safe approach. However, once the head has been found, the direction, speed and acceleration of the motion of the head can be analyzed, and used to predict the head-pose and the resultant position for the next frame.

Since for the purpose of face tracking speed is important and there is no information extracted, the sensor data can be down-sampled before the alignment of the template with the sensor data will take place. Since ICP measures euclidean distance between points, having less data to compare will significantly increase speed. This comes as a trade-off to precision. Nevertheless an equilibrium can always be found.

Using ICP, once the rough position of the face is matched to the template, one can deform the face template, which can now be considered as a 3DMM, to minimize the distance error from the mask template to the face. This, in turn, reduces the oscillation of the template on the face and allows precise future location of the face in 3D space. Since for the recognition process, the distance form the face to the mask is not measured, this has no negative implications on the discriminative process.

2) Normalization: The notion of normalization in the case of face recognition refers to the alignment of the face data in a canonical position. This is necessary since the data captured from the sensor needs to be in the same frame of reference. During normalization, the face size is also estimated according to the distance from the sensor and the pose (angle) of the face in 3D space, in what is usually called a 3D front view mug shot pose.

The normalization approaches described in [6] focuses mainly on the qualitative aspect of alignment and aggregates data from multiple poses. It also relies on explicitly modifying the sensor data to align it to an actual frontal mug shot pose. The process described not only alters the data but is also computationally expensive. While this might be a good strategy for off-line processing, it is not appropriate for real-time applications.

Instead of changing and aligning the input data, once the face has been localized using ICP it is only necessary to align the normal plane to the face. The plane size is proportional to the size of the face and has its center through the nose, as shown in (Fig. 2). Later, during the feature extraction procedure distances from the face to the surface of the plane can be projected.

Since data will not be altered, this strategy will allow faster computation cycles and will focus on the quantitative aspects of acquiring depth data, since this information can be captured in real-time. Even if the alignment of the plane with the face is not completely precise, with enough captured frames, the average distances can be computed.

B. Feature extraction

The advantage with aligned descriptors, is the fact that no information is discarded in the process of information extraction, therefore the discriminative capabilities are much higher. However, the alignment process can be computationally expensive, which this is especially true if landmarks are used for normalization.

Nonrigid ICP, although not as precise and more computationally expensive if used on the raw input data, is capable of identifying the position of the face in 3D space even if the subject is moving freely in the environment. However, as described in the tracking section, head tracking, Kalman filters and down-sampling the data can considerably improve the performance. Furthermore, possible combinations with simple histogram based trackers could improve accuracy.

Using ICP, the position of the face is known since the aligned ICP upper face mask template reveals the 3D coordinates of the face in real-time. Using these coordinates, the location of the nose can be pinpointed in order to trace a 2D surface plane normal with the centroid tangent to the nose.

Once the surface normal to the face is available, the distances from the plane to the face can be captured in real time. Since the size of the face would vary according to the distance from the sensor, the resolution of the data would also change. Therefore, the number of distances to the surface normal would vary with the number of point cloud samples available. Therefore a normalization step has to be performed here as well in order to maintain the number of distance measurements constant.

The whole purpose of aligning a surface normal to the face is to avoid having to process the raw data since that would make the process slower. One solution would therefore be to always down sample the input data according to a maximum distance that a face can be distinguished by the face tracking algorithm. However, the lack of resolution might cause a consistent decrease in variation of the distances to surface measured. This obviously has to be empirically verified.

Another solution to the problem of data resolution would be to aggregate models of the face in short time intervals as frames are captured from the sensor. This would also cause an initial

delay since the frame 'buffer' would need to be populated with enough frames. The method described in [11] could be used to create such temporary fixed resolution (face only) vertexes to enrich the incoming data in case the face is too far away and the resolution is not high enough.

Median filtering can be used as a preprocessing step in order to reduce noise. Nevertheless, the distances to the surface normals will always oscillate since the commodity sensors will always produce such data. Median filtering, however, might also cause loss of detail which is not desired. Therefore the proper level of filtering has to be empirically determined. As long as the distances to the surface normal don't deviate too much from a realistic value, which can be computed using the area around the face template as an average distance surface-space, the flow of incoming values should be stable. (Fig. 3).

Most paradigms integrate data in time to build a static model such as the previously described faceprints [8]; the problems with such representations is the lack of capability to deal with occlusions and varying facial expressions. A solution to this lack of robustness is to transform the task into a time series prediction problem by capturing the distance oscillations in real-time according to the changes in face expression or occlusions.

C. Classification

The human brain is a complex recurrent neural network (RNN) [55] - a network of neurons with feedback connections. Due to its properties it can learn many behaviors and sequence processing tasks that are not learnable by traditional machine learning methods. A special kind of RNN is the long short-term memory (LSTM) [56] recurrent neural network.

This type of network has been previously used in such tasks as handwriting recognition, speech recognition or stock market prediction. The advantage of using such an algorithm is that it can overcome the fundamental problems of traditional RNNs and efficiently learn to solve many tasks such as the recognition of temporal order of noisy input streams through its robust storage capability of high precision real numbers across extended time intervals.

In order not to be computationally intensive, the data will not be constantly fed into the network. By learning partial sequences of data, the network can be trained to capture new data when the recognition rate falls below a certain threshold. If the user moves around, each new pose or angle should trigger another learning cycle.

In this manner, the prediction of the missing 'signals' from the face is possible in case there are partial occlusions of the face. Since this type of network can lean the areas of constant shape and those that change often, such as the lower area of the face near the jaw, it could also be used to map sequences of signals to facial expressions.

IV. Conclusion

Previously performed experiments in 3D face recognition using depth sensors often lack proper quantitative measurements or an assessment of the effect of partial occlusions or facial expressions. Therefore, it is reasonable to state that a proper investigation towards real-time unconstrained face recognition using this type of sensors has not been performed yet.

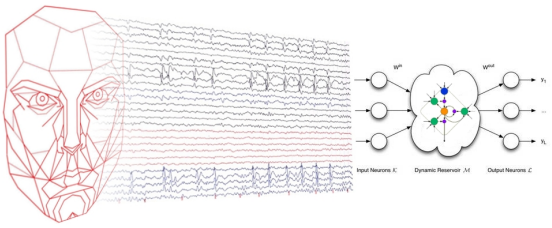

Fig. 3. The distance oscillations from the normal plane surface to the face are recorded and used as input for a LSTM RNN.

A novel method for unconstrained real-time face recognition using commodity depth cameras has been proposed. By capturing distances from the face to a surface normal tangent to the nose, a dynamic model can be captured in real-time. The internal representation can deal with partial occlusions of the face and changes due to various facial expressions by predicting movement of the face and fitting in missing data with previously learned models.

Our previously performed experiments on 2D face recognition using auto-encoders [57] have revealed that there are underlying universal facial features which can be used to discriminate successfully even between previously unseen subjects.

However, this method of encoding features is only partially robust towards facial occlusions and can deal only with low variation in head pose. This can be compensated by first classifying the type of pose and consequently training kernel auto-encoders per each pose category.

The same type of nonlinear PCA can be used as a kernel on data extracted from depth sensors and used to discriminate between 3D facial features.

Initial tracking results are promising and point towards the fact that EM-ICP [51] is reliable and can be used in real-time provided the data is down-sampled to a level that still retains the most important features of face geometry.

LSTMs have the proven ability to perform recognition of temporal order of noisy input streams and are ideal for this kind of data processing.

References

[1] S. Mori, C. Suen, and K. Yamamoto, "Historical review of ocr research and development," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1029–1058, 1992.

[2] T. Leyvand, C. Meekhof, Y. Wei, J. Sun, and B. Guo, "Kinect identity: Technology and experience," *Computer*, vol. 44, no. 4, pp. 94–96, 2011.

[3] H. Moon and P. Phillips, "Computational and performance aspects of pca-based face-recognition algorithms," *PERCEPTION-LONDON-*, vol. 30, no. 3, pp. 303–322, 2001.

[4] S. McKenna and S. Gong, "Non-intrusive person authentication for access control by visual tracking and face recognition," in *Audio-and Video-based Biometric Person Authentication*. Springer, 1997, pp. 177–183.

[5] M. Breidt, H. Biilthoff, C. Curio *et al.*, "Robust semantic analysis by synthesis of 3d facial motion," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 713–719.

[6] N. Pears, T. Heseltine, and M. Romero, "From 3d point clouds to pose-normalised depth maps," *International Journal of Computer Vision*, vol. 89, no. 2, pp. 152–176, 2010.

[7] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang, "3d deformable face tracking with a commodity depth camera," *Computer Vision–ECCV 2010*, pp. 229–242, 2010.

[8] S. Meers and K. Ward, "Face recognition using a time-of-flight camera," in *Computer Graphics, Imaging and Visualization, 2009. CGIV'09. Sixth International Conference on*. IEEE, 2009, pp. 377–382.

[9] C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein, and D. Metaxas, "The best of both worlds: Combining 3d deformable models with active shape models," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–7.

[10] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 1, pp. 4–20, 2004.

[11] H. Li, B. Adams, L. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 5, p. 175, 2009.

[12] T. Weise, B. Leibe, and L. Van Gool, "Accurate and robust registration for in-hand modeling," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[13] S. Bauer, J. Wasza, K. Muller, and J. Hornegger, "4d photogeometric face recognition with time-of-flight sensors," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*. IEEE, 2011, pp. 196–203.

[14] W. Chen, M. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 2, pp. 458–466, 2006.

[15] S. Du and R. Ward, "Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 9, pp. 1165–1175, 2010.

[16] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[17] A. Nefian and M. Hayes III, "Maximum likelihood training of the embedded hmm for face detection and recognition," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 1. IEEE, 2000, pp. 33–36.

[18] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[19] M. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[20] J. Yang, D. Zhang, A. Frangi, and J. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 1, pp. 131–137, 2004.

[21] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional data-with application to face recognition," *Pattern recognition*, vol. 34, no. 10, p. 2067, 2001.

[22] D. Gorodnichy *et al.*, "Video-based framework for face recognition in video," in *Second Workshop on Face Processing in Video (FPiV'05) in Proceedingsof Second Canadian Conference on Computer and Robot Vision (CRV'05)*, 2005.

[23] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Computer Vision and Image Understanding*, vol. 91, no. 1, pp. 214–245, 2003.

[24] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–313.

[25] G. Edwards, C. Taylor, and T. Cootes, "Improving identification performance by integrating evidence from sequences," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1. IEEE, 1999.

[26] I. Kemelmacher-Shlizerman and R. Basri, "3d face reconstruction from a single image using a single reference face shape," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 394–405, 2011.

[27] B. Horn and M. Brooks, "The variational approach to shape from shading," *Computer Vision, Graphics, and Image Processing*, vol. 33, no. 2, pp. 174–208, 1986.

[28] J. Huang, B. Heisele, and V. Blanz, "Component-based face recognition with 3d morphable models," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 1055–1055.

[29] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063–1074, 2003.

[30] Y. Taigman and L. Wolf, "Leveraging billions of faces to overcome performance barriers in unconstrained face recognition," *Arxiv preprint arXiv:1108.1122*, 2011.

[31] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 1978–1990, 2011.

[32] "Face.com is operated by face.com inc. they provide a cloud computing platform for face detection and identification. they report recognition of 91.3% +/- 0.3, achieved on the lfw (labelled faces in the wild) test set."

[33] G. Huang, M. Mattar, T. Berg, E. Learned-Miller *et al.*, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[34] A. Bronstein, M. Bronstein, and R. Kimmel, "Expression-invariant 3d face recognition," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 62–70.

[35] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2d+ 3d active appearance models," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE Computer Society; 1999, 2004.

[36] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194.

[37] P. Besl and N. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[38] K. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3d and multi-modal 3d+2d face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[39] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/off: Live facial puppetry," in *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 2009, pp. 7–16.

[40] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," *ACM Transactions on Graphics*, vol. 30, no. 4, 2011.

[41] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek, "The trimmed iterative closest point algorithm," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3. IEEE, 2002, pp. 545–548.

[42] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[43] S. Meers and K. Ward, "Head-pose tracking with a time-of-flight camera," *Faculty of Informatics-Papers*, p. 720, 2008.

[44] J. Park and I. Sandberg, "Universal approximation using radial-basis-function networks," *Neural computation*, vol. 3, no. 2, pp. 246–257, 1991.

[45] J. Carr, R. Beatson, J. Cherrie, T. Mitchell, W. Fright, B. McCallum, and T. Evans, "Reconstruction and representation of 3d objects with radial basis functions," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 67–76.

[46] G. Hetzel, B. Leibe, P. Levi, and B. Schiele, "3d object recognition from range images using local feature histograms," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2. IEEE, 2001, pp. II–394.

[47] H. Ding, F. Moutarde, A. Shaiek *et al.*, "3d object recognition and facial identification using time-averaged single-views from time-of-flight 3d depth-camera," 2010.

[48] M. Hagan, H. Demuth, M. Beale, and B. University of Colorado, *Neural network design*. PWS Pub, 1996.

[49] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*. Eurographics Association, 2006, pp. 61–70.

[50] W. Zhang, Q. Wang, and X. Tang, "Real time feature based 3-d deformable face tracking," in *Proceedings of the 10th European Conference on Computer Vision: Part II*. Springer-Verlag, 2008, pp. 720–732.

[51] S. Granger and X. Pennec, "Multi-scale em-icp: A fast and robust approach for surface registration," *Computer VisionECCV 2002*, pp. 69–73, 2006.

[52] S. Smith and J. Brady, "Asset-2: Real-time motion segmentation and shape tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 814–820, 1995.

[53] M. Bohme, M. Haker, T. Martinetz, and E. Barth, "Head tracking with combined face and nose detection," in *Signals, Circuits and Systems, 2009. ISSCS 2009. International Symposium on*. IEEE, 2009, pp. 1–4.

[54] S. Gokturk and C. Tomasi, "3d head tracking based on recognition and interpolation using a time-of-flight depth sensor," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. Ieee, 2004, pp. II–211.

[55] G. Dorffner, "Neural networks for time series processing," in *Neural Network World*. Citeseer, 1996.

[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[57] M. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.