Introducing FoxFaces: A 3-in-1 Head Dataset

Amel Aissaoui[1], Afifa Dahmane[1], Jean Martinet[2] and Ioan Marius Bilasco[2]
[1]*Computer Science Department, University of Science and Technologies Houari Boumediene -USTHB, Algiers, Algeria*
[2]*Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, IRCICA, F-59000 Lille, France*

Keywords: Biometric Dataset, Face Analysis and Recognition, Expression, Pose, Multimodal.

Abstract: We introduce a new test collection named FoxFaces, dedicated to researchers in face recognition and analysis. The creation of this dataset was motivated by a lack encountered in the existing 3D/4D datasets. FoxFaces contains 3 face datasets obtained with several devices. Faces are captured with different changes in pose, expression and illumination. The presented collection is unique in two aspects: the acquisition is performed using three little constrained devices offering 2D, depth and stereo information on faces. In addition, it contains both still images and videos allowing static and dynamic face analysis. Hence, our dataset can be an interesting resource for the evaluation of 2D, 3D and bimodal algorithms on face recognition under adverse conditions as well as facial expression recognition and pose estimation algorithms in static and dynamic domains (images and videos). Stereo, color, and range images and videos of 64 adult human subjects are acquired. Acquisitions are accompanied with information about the subjects identity, gender, facial expression, approximate pose orientation and the coordinates of some manually located facial fiducial points.

1 INTRODUCTION

The capability of a system to precisely assign an identity to a person in unconstrained, uncooperative, "in the wild" settings is still an open problem. Hard biometric techniques such as iris and fingerprint-based are widely used but require cooperative users (finger pressed strong against a reader, iris in front of a dedicated camera). Soft biometric techniques such as face-based are less intrusive, and yet generally working well if the illumination conditions are satisfactory and the user is cooperative (facing the camera with neutral expression).

Most of the proposed solutions are generally based on the 2D images (2D face recognition) and various color-, shape- and texture-features are used and require at least one frontal pose of the face. However, the 2D face recognition is very sensitive to changes in pose, illumination and facial expression. The 3D face recognition was explored in order to face problems encountered in 2D recognition. Indeed, the 3D shape of the face is an important discriminative information allowing better recognition accuracy. The 3D recognition is less sensitive to illumination and pose changes which makes it a potential solution to face recognition problem, as it was demonstrated in the FRGC evaluation (Phillips et al., 2005). However, 3D methods require expensive devices (3D scanners), long acquisition time and human cooperation. With the large availability of depth cameras, researchers tend to use "less constrained" depth cameras (e.g. Kinect), allowing rapid depth acquisition and requiring less human cooperation. Indeed, these advantages compensate the low quality of the depth information comparing to 3D scanner data.

Most of the solutions proposed in the field of face recognition are based on supervised or semi-supervised training. Datasets which through the years tried to get closer to "in the wild" settings are used for training and for measuring the effectiveness of proposed algorithms. A number of datasets that address the field of person recognition and encompass also depth information are available. For instance, the BU-3DFE dataset (Yin et al., 2006) presents filtered 3D models acquired by 3D scanners of persons expressing various expressions. However, the quality of models used as training data may hardly be achieved in a "wild" environment using regular depth camera. In order to provide researchers with adequate training datasets for handling 2D and depth information especially, but not only, in the field of face recognition, we have designed the FoxFaces dataset. The FoxFaces dataset narrows the gap between lab and wild conditions, by using off-the-shelf

Aissaoui, A., Dahmane, A., Martinet, J. and Bilasco, I.
Introducing FoxFaces: A 3-in-1 Head Dataset.
DOI: 10.5220/0005714405330537
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 533-537
ISBN: 978-989-758-175-5

camera allowing less constraining 3D acquisition in order to approximate more realistic conditions. We consider also, various expressions, poses and illuminations settings. The database is freely available for research purposes and can be requested at URL http://www.lifl.fr/FOX/index.php?page=datasets.

The paper is structured as follows. In Section 2, we introduce major datasets in the field of person recognition covering head pose, gender and facial expressions variations and using 2D, depth, stereo, or 3D data, emphasizing the need for constructing a novel dataset narrowing the gap between lab conditions and wild settings. The methodology employed to design and collect the dataset is described in details in Section 3. Some experiments, in a face recognition context, are presented in Section 4 before summarizing the contributions and the new challenges related to the proposed dataset in Section 5.

2 RELATED DATASETS

Many public datasets are available for face related researches. Existing resources can be 2D or 3D. 3D data can be divided into two major categories: data offering a complete 3D face models using 3D scanners and recently available, the data offering a view-based 3D face models using depth sensors (e.g. Microsoft Kinect) which is our context of study.

The use of 3D data in conjunction with 2D data has been broadly spread in face analysis research. Datasets offering 3D information on faces were therefore proposed in order to evaluate algorithms in this field. Table 1 lists some well-known bimodal 2D-3D datasets and their specifications. We included the proposed dataset in the bottom of the table in order to show its contributions. 3D FRGC (Phillips et al., 2005) and Texas (Gupta et al., 2010) datasets are face recognition-oriented with some variations in facial expressions and illumination. Bosphorus dataset (Savran et al., 2008) is more general and can be used in different face analysis tasks. It contains a large variation of head poses and face expressions. BU-3DFE dataset (Yin et al., 2006) is expression recognition-oriented. 7 expression shape models are collected and a corresponding facial texture image captured from two views is also given. BU-4DFE dataset (Yin et al., 2008) is an extended version of the BU-3DFE dataset offering temporal information by capturing dynamic 3D data.

The above datasets are made via expensive equipments (3D scanners) offering a high quality data (complete 3D face models). Hence, they require specific acquisition conditions: sufficient time for scanning, cooperation of the person to be identified installed in front of the scanner until the end of scanning.

Recently, in order to extend the scope of 3D, the research interest focuses increasingly on the use of less restrictive 3D equipments (e.g. Kinect). In Table 1, we include 2 recent datasets obtained via Microsoft Kinect sensor. BIWI database (Fanelli et al., 2013) is compound of head movements 3D sequences of 20 subjects under ambiant lighting and neutral expression. Eurecom dataset (Min et al., 2014) contains still 2D and 3D images of 52 subjects showing few changes in expression and head pose orientation. Very few datasets acquired with depth sensors are available and none, to the best of our knowledge, encompasses all variations. In this paper, we present a new face dataset which encompasses different data modalities. This dataset presents 2D, 3D and stereo images of the face, in order to allow testing and comparing face analysis methods in low constrained context. The dataset offers both static and dynamic data. 3D dynamic data, also called 4D data, allows the extension of studies to a time-varying 3D faces. Additionally to these aspects, various face expressions and head poses are taken from 64 subjects. Face analysis can be performed under varying pose, expression and illumination. In the remainder, we introduce the dataset and highlight its usefulness for bimodal face recognition.

3 THE FoxFaces DATASET

3.1 Acquisition Devices

The dataset has been built using an acquisition system composed with 3 sensors:

- **Infrared Sensor:** we used Microsoft Kinect, that contains a color camera, an infrared light, and an infrared CMOS sensor (QVGA 320x240, 16 bits) able to generate a depth map of the scene by estimating the amount of reflected infrared light. The farther an object, the less light it reflects.
- **3D Time-of-Flight Sensor:** we have used Mesa Imaging SR4000 sensor, that flashes the scene with infrared light pulses,
- **Stereo Camera:** we used Point Grey Bumblebee XB3, a multi-baseline sensor equipped with three 1.3 megapixel cameras. The large baseline offers a higher precision in higher distances from camera, and the small baseline enhances the matching of small intervals.

Table 1: List of some public 2D-3D face datasets (N = number of subjects).

Face dataset (N)	Devices	Expressions	Illumination	Pose	Dyn./Static
3D FRGC (466) (Phillips et al., 2005)	3D scanner	various	various	frontal	static
Texas (118) (Gupta et al., 2010)	3D scanner	neutral/smliling	3 illuminations	frontal	static
Bosphorus (105) (Savran et al., 2008)	3D scanner	35 expressions	homogeneous	13 poses	static
BU-3DFE (100) (Yin et al., 2006)	3D scanner	6 expressions	homogeneous	2 sides	static
BU-4DFE (101)(Yin et al., 2008)	3D scanner	6 expressions	homogeneous	frontal	dynamic
BIWI (20) (Fanelli et al., 2013)	Kinect	neutral	homogeneous	free movements	dynamic
Eurecom (52) (Min et al., 2014)	Kinect	3 expressions	2 illuminations	3 yaw	static
FoxFace (64)	Kinect ToF camera Stereo	7 expressions	3 illuminations	30 poses	static+dynamic

Figure 1: Top row: example of infrared sensor images (Kinect): (a) color image (b) depth image. Middle row: example of TOF sensor images (SR4000): (a) infrared image (b) depth image (c) confidence matrix (bright pixels show high confidence. Bottom row: example of image triple acquired with the stereo camera (Bumblebee XB3).

Three collections have been acquired with the acquisition system: one stereo collection (FoxStereo), and two bimodal 2D-3D collections (FoxKinect and FoxTof). The 3D data for all collections are given as depth maps.

3.2 Methodology

The data acquisition has been carried out indoor, in our office rooms, with 64 subjects (46 males, 18 females) aged 22-59. Note that among the 46 males persons, two are twin brothers. Subjects are located 1 meter away from the cameras[1]. Three parameters (lighting conditions, face expression, and head pose) are varied throughout the data acquisition. For each subject, 40 images are recorded, corresponding to:

- 3 lighting conditions: ambient, frontal, side;
- 7 face expressions: joy, sadness, hanger, disgust, fear, surprise, *neutral*;
- 30 head poses resulting from a combination of 9 positions in *yaw* (from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$, using $\frac{\pi}{8}$ steps), with 3 *pitch* directions (downwards, frontal, upwards), plus 2 *roll* positions (left and right).

To obtain the head poses, we have used markers which correspond to the different horizontal and vertical rotation of the head. Users are asked to look straight ahead to the markers by moving their entire head and not to direct their eyes.

In total, the collection contains 2560 images. Figure 2 shows an example of all possible variations for a subject. In addition to static images (2D and depth), a video sequence is also recorded for each subject, containing all variations, for the three datasets FoxStereo, FoxKinect et FoxTof.

Figure 2: Example of all possible variations for a subject: (a) 3 lighting conditions (b) 7 face expressions (c) 30 head poses.

3.3 Annotation

A manual annotation has been performed for 4 main face interest points (eyes, nose tip, mouth center) in order to enable a precise face localization. Figure 3 shows annotated point on a face from the dataset. All images are given a label that encodes the following information: subject id, image id, pose, lighting condition, face expression, image type (2D, depth).

[1]It is the minimal distance for the Kinect.

Figure 3: Annotated interest points on a face.

4 EXPERIMENTS AND DISCUSSION

We present in this section some experiments conducted on an example of application (in a face recognition context) using two data partitions (FoxKinect and FoxStereo).

Experiments are based on a bimodal (2D and depth) face recognition framework (Aissaoui and Martinet, 2015). The bimodal framework considered here is strongly inspired by LBP applied on both depth and color channels. A variation of LBP characterizing depth images (called DLBP) was proposed in (Aissaoui et al., 2014). We use this descriptor for constructing a depth-related feature vector. This depth-related feature vector is combined with texture-related LBP feature vector extracted from 2D images in order to characterize, in a bimodal context, training and testing faces. Before extracting depth and color features, several preliminary transformations are applied. Faces are extracted from gray-level and depth images using the annotations present in the dataset. They are then normalized to 100x100 pixels. Depth images presenting lot of missing data are dropped at this stage. Denoising of depth images is also performed as described in (Aissaoui et al., 2014). We notice that head pose variation is not considered in these experiments.

In order to quantify the challenges brought by the new dataset, we compare in Fig 4, the face recognition rates obtained by applying some descriptors (LBP, DLBP and 3DLBP (Huang et al., 2006)) on 3D data from our dataset to those obtained on other well-known 3D datasets (FRGC, TEXAS and Bosphorus) (Aissaoui et al., 2014). Descriptor's parameters given the best rates are chosen. The recognition is performed using 1NN method with 10-cross validation.

The obtained results show that with this framework, comparing to other state of the art datasets, the new dataset is more challenging.

Figure 5 shows results obtained by the bimodal framework (DLBP for depth images and LBP for grey-level images) applied on Foxkinect and FoxStereo data. Recognition is performed using SVM classification scheme with an RBF kernel. The well-used SVM is used in order to provide overall comparisons metrics between various approaches on the present dataset. Experiments include mono-modal (2D and 3D) and bimodal methods employing various fusion strategies(early, late and bi-level fusion).

Early fusion does not seems to increase the performances, probably due to noise brought by each feature, but also, to the different nature of the information characterized by features (depth and texture). Decision fusion appears more efficient as the used classifier seem to manage better noise and variations when only one feature-type is considered. The bi-level fusion, which combines 2D, 3D and early fusion decisions, proves to be better than the three other strategies used separately.

5 CONCLUSION

In this paper, we described a new collection for evaluating algorithms on face recognition. The constructed collection aims to offer a rich evaluation resource for researchers in face identification using different modalities. It contains static (images) and dynamic (videos) acquisitions, using 2D, 3D and stereo modalities. The used equipments offer a depth acquisition with less constraints in terms of cost and time, comparing to the 3D scanners. The available data in this collection allows the evaluation of algorithms in a large range of research fields in particular:

2D, 3D and Bimodal Face Recognition: color images can be used in the conventional 2D face recognition. Depth images can be used to evaluate algorithms of 3D face recognition. They can also be combined with 2D images to be used in a bimodal recognition context. The recognition can be evaluated across changes in pose, expression and illumination.

3D Face Reconstruction: the stereo pairs can be used in order to estimate depth maps of the face using stereo-based reconstruction algorithms. 3D face model reconstruction algorithms based on combining depth maps captured from different point of views can also be applied on this dataset.

Facial Expressions Recognition: Different expression for each identity are acquired in the collection. Facial expression recognition methods, in 2D, 3D or 2D-3D, can be evaluated using our collection.

Head Pose Estimation: the collection is rich in term of changes in pose. Faces are captured under 30 different pose. Hence, the collection can be very useful for 2D, 3D and bimodal head pose estimation.

Face Detection: the annotation information can be used for evaluating face and fiducial points (eyes,

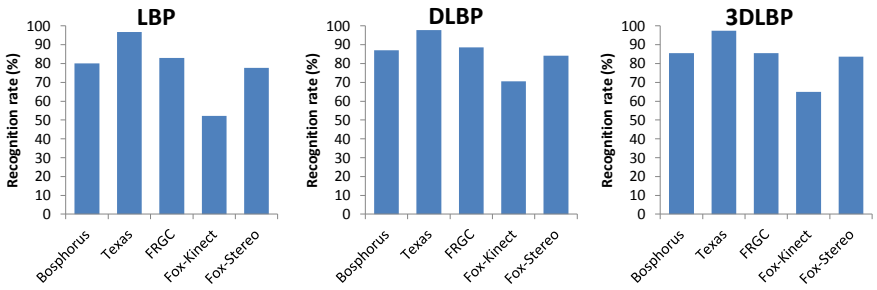

Figure 4: Comparing the new dataset to state-of-the-art datasets using 3 different descriptors and 1NN method.

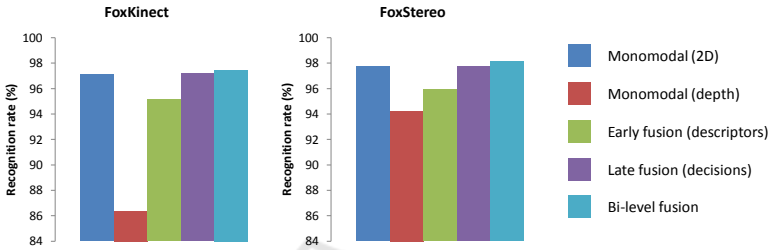

Figure 5: Performances of monomodal methods (2D, depth/3D) and bimodal methods with various fusion strategies on FoxKinect and FoxStereo.

nose and mouth) detection methods in both 2D and 3D modalities.

Dynamic Face Analysis: A latest interest in face analysis consists in using time dimension for identification and expressions recognition. Videos available in our collection (2D and 3D) can be very useful for this areas of research.

Experiments conducted here give an example of using a partition of the dataset in a frontal face recognition context. However, our dataset is available for researchers in other related research fields.

REFERENCES

Aissaoui, A. and Martinet, J. (2015). Bimodal 2d-3d face recognition using a two-stage fusion strategy. *5th International Conference on Image Processing Theory, Tools and Applications (IPTA 2015)*, pages p. 279–284.

Aissaoui, A., Martinet, J., and Djeraba, C. (2014). DLBP: A novel descriptor for depth image based face recognition. In *Proceedings of the 21th IEEE international conference on Image processing*, pages 298–302.

Fanelli, G., Dantone, M., Gall, J., Fossati, A., and Van Gool, L. (2013). Random forests for real time 3d face analysis. *Int. J. Comput. Vision*, 101(3):437–458.

Gupta, S., Markey, M. K., Castleman, K. R., and Bovik, A. C. (2010). Texas 3D face recognition database. *IEEE SSIAI*, pages 97–100.

Huang, Y., Wang, Y., and Tan, T. (2006). Combining statistics of geometrical and correlative features for 3D face recognition. In *Proceedings of the British Machine Vision Conference*, pages 879–888.

Min, R., Kose, N., and Dugelay, J.-L. (2014). Kinectfacedb: A kinect database for face recognition. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(11):1534–1548.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. (2005). Overview of the face recognition grand challenge. In *Proceedings of CVPR'05*, pages 947–954.

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Biometrics and identity management. chapter Bosphorus Database for 3D Face Analysis, pages 47–56. Springer-Verlag, Berlin, Heidelberg.

Yin, L., Chen, X., Sun, Y., Worm, T., and Reale, M. (2008). A high-resolution 3d dynamic facial expression database. In *8th IEEE Intl. Conf. on Automatic Face and Gesture Recognition (FG 2008), Amsterdam, The Netherlands*, pages 1–6.

Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE.