The MUCT Landmarked Face Database

Stephen Milborrow
University of Cape Town
www.milbo.users.sonic.net

John Morkel
University of Cape Town
www.eleceng.uct.ac.za/people

Fred Nicolls
University of Cape Town
www.eleceng.uct.ac.za/people

***Abstract*—We present the MUCT database consisting of 3755 images of human faces with 76 manual landmarks. Compared to existing publicly available 2D manually landmarked databases, the MUCT database provides more diversity of lighting, age, and ethnicity. As an example application of the database we show that an Active Shape Model trained on the MUCT data is a little more flexible than one trained on the XM2VTS data. The database is freely available for academic use.**

I. Introduction

Many databases of face images are publicly available. For certain applications, not only the images themselves but also the coordinates of the facial features are necessary. With the current state of the art, these coordinates, or *landmarks* must be located manually, that is, by a human clicking on the screen. This paper introduces the MUCT database of 3755 faces with 76 manual landmarks. Our main motivation for creating the database was to provide more variety than the existing publicly available landmarked databases — variety in terms of lighting, age, and ethnicity.

In this paper we first describe the MUCT images and landmarks. We then give an overview of similar databases. Finally we compare Active Shape Models (ASMs) [1] trained on the MUCT data to those trained on the XM2VTS data [2].

II. The MUCT Database

Figure 1 shows some images from the MUCT database. (MUCT stands for "Milborrow / University of Cape Town"). The subjects in the database were sampled from people around the Leslie Social Sciences Building on the University Of Cape Town campus in December 2008. This diverse population included students, parents attending graduation ceremonies, high school teachers attending a conference, and employees of the university such as cleaners and security personnel. A wide range of subjects was photographed, with approximately equal numbers of males and females, and a cross section of ages and races. To recruit subjects, one of the researchers approached people asking if they would volunteer to be photographed, with the promise of a bar of chocolate as an inducement.

Subjects who wore makeup, glasses, or headdresses retained those for the photographs. Subjects were not asked to display any particular facial expression; in practice this meant that most were photographed with a neutral expression or a smile. All subjects were 18 or more years of age.

Each subject was photographed with five webcams arranged as shown in Figures 3 and 4, yielding the views shown in Figure 5. An attempt was made to trigger all five cameras

Fig. 1. Some images from the MUCT database.

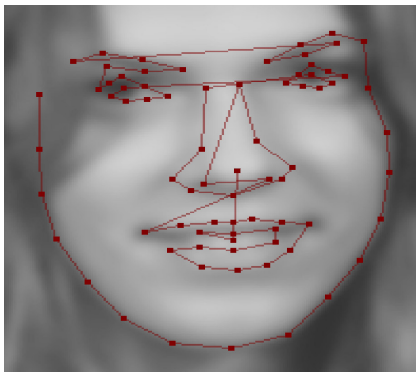

Fig. 2. The 76 MUCT landmarks. These landmarks are the 68 points defined by the popular FGnet [3] markup of the XM2VTS database [2], plus four extra points for each eye.

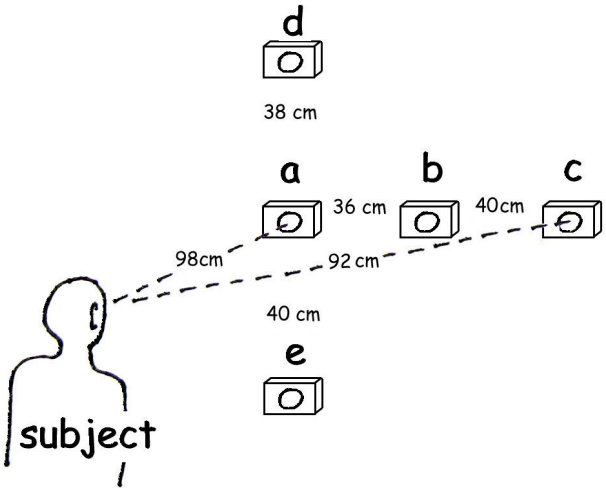

Fig. 3. The five cameras and their relationship to the subject's face. Camera `a` is directly in front of the subject's face, up to variations in height of seated subjects.

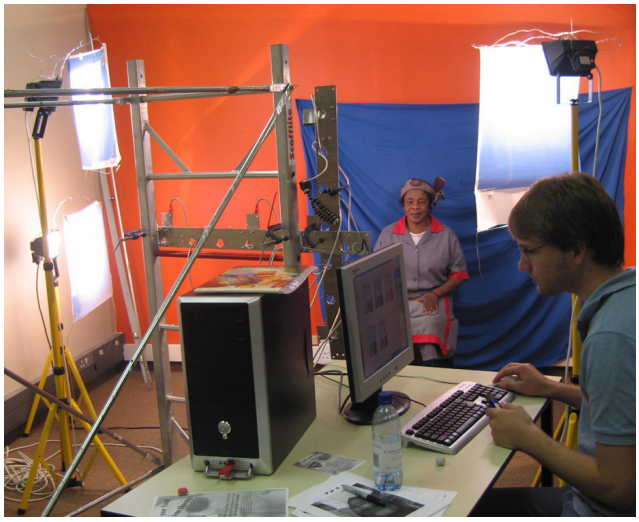

Fig. 4. The overall setup, looking towards the subject past the frame holding the cameras.

simultaneously, making the database also useful for applications requiring multiple simultaneous views of the subject — although software delays meant that there was some difference in triggering times. Each subject was seated and faced camera `a`, however differences between the seated height of subjects and their posture introduced some variation in their orientation relative to the cameras (the subjects were seated to minimize variation in height). Note that no cameras were located to the left of the subject, since those views can be approximated by mirroring the images from the cameras on the right.

Ten different lighting setups were used, and each subject was photographed with two or three of these lighting sets. Not every subject was shot with every lighting setup, to achieve diversity without too many images. Table I gives details, and Figure 6 shows one subject shot under three different lighting sets. Standard neon office lighting was augmented by halogen

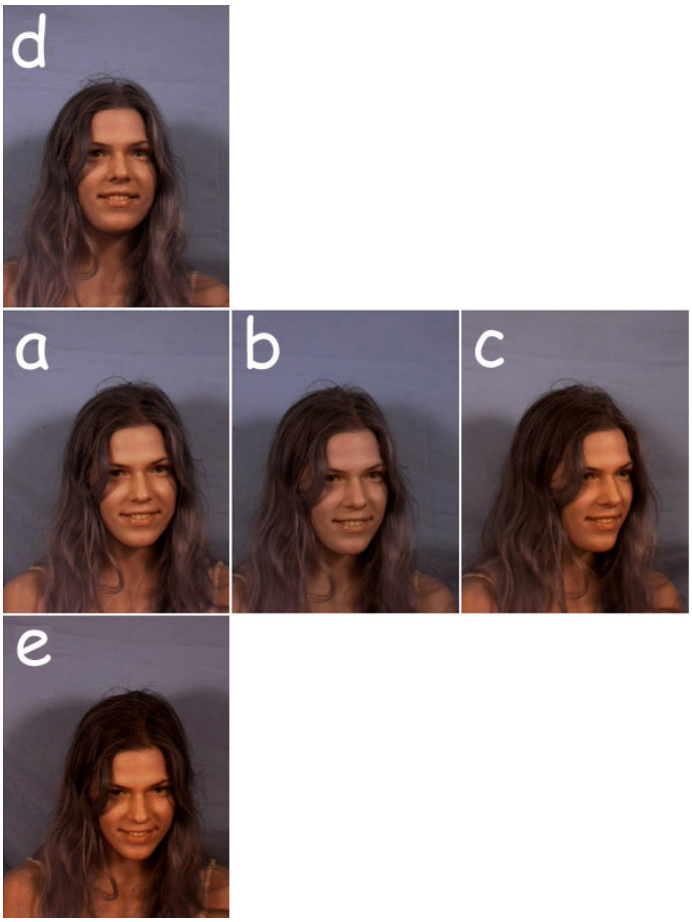

Fig. 5. The five camera views of one subject. This subject was shot under different lighting sets; only one set is shown here.

Fig. 6. A subject shot under three different lighting sets. Five views of this subject were shot for each of these lighting sets; only the frontal view (camera `a`) is shown here.

lamps (such as those found in hardware stores) with diffusion screens (Figure 4).

The cameras were Unibrain Fire-i [4] webcams with Sony ICX098BQ CCD sensors. The image format was RGB (3 x 8 bit) with a resolution of 640 x 480 pixels. Camera settings were fixed for all images as follows (please refer to the Unibrain camera documentation [4] for details on these settings): Exposure manual 350, WhiteBalance manual 85/50, Brightness manual 180, Gain 70, Shutter 6, Gamma 0, Saturation 120, FrameRate 3.75, and Sharpness 50. These

TABLE I
LIGHTING SETS IN THE MUCT DATABASE. THE "NBR OF IMAGES" COLUMN IS THE NUMBER OF SUBJECTS × THE NUMBER OF LIGHTING SETS (2 OR 3) × THE NUMBER OF VIEWS (5).

Lighting set	Subject ID	Nbr of subjects	Nbr of images
q r s	000 - 090	91	1365
t u v	200 - 307	108	1620
w x	400 - 451	52	520
y z	600 - 624	25	250
		Total 276	Total 3755

settings were chosen to produce a reasonably wide dynamic range with no or very little saturation, with the color balance on a white card uniformly distributed across the three color channels (in retrospect, we should have set the color balance for better skin tones). For reference, photos were also made of a GretagMacbeth color chart [5] in the same environment. Although automatic exposure and white balance is more typical for webcams, we used fixed settings for uniformity across all images. However, it should be mentioned that we found considerable difference between the five individual Unibrain cameras (for example, the color balance of the image was different across cameras when photographing the same scene with the same camera settings).

III. THE LANDMARKS

Figure 2 shows the positions of the MUCT landmarks. The definition of these landmarks is the same as the 68 XM2VTS [2] points, plus 4 extra points around each eye. The position of landmarks obscured by hair or glasses was estimated by the human landmarker. Landmarks that were obscured behind the nose or side of the face in a three-quarter view were marked as such with a special value in the database (this only affects images taken with cameras `b` and `c`). All landmarks were carefully checked by a third party.

How reliable are manual landmarks? To test the validity of the assumption that manual landmarks are a suitable ground truth, we manually re-landmarked the left eye pupil and nose tip on 300 BioID faces [6]. The objective was to see how these differed from the original BioID landmarks. (The position of the pupil can usually be estimated reliably by hand. Estimating the position of the tip of the nose is much more subjective. The mean accuracy of other landmarks in the interior of the face can be expected to be somewhere between these two extremes.) We measured the euclidean distance between each of these remarked landmarks and its original position in the BioID FGnet data [3]. We then divided this distance by the inter-eye distance to prevent arbitrary dependence on face size. The mean inter-eye distance is 61 pixels for these 300 images, thus a distance of 1/61 = 0.016 corresponds to one pixel. Figure 7 shows the distribution of these normalized distances. The figure shows that one can expect median uncertainties in manual landmark positions of roughly one or two pixels for faces like the BioID faces.

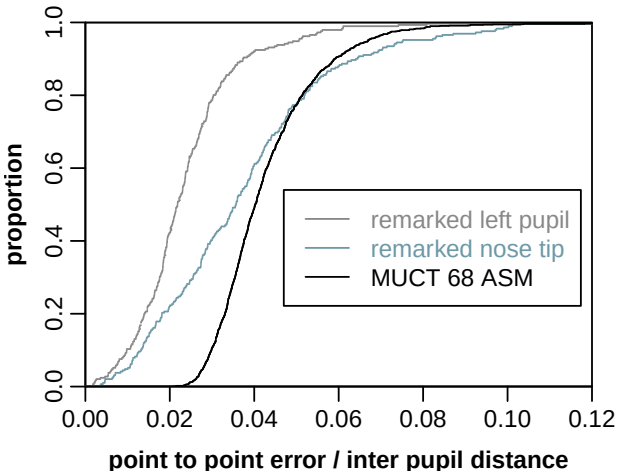

Fig. 7. Density of the discrepancy between original BioID landmarks and manually remarked points. Measured on 300 BioID faces as the euclidean distance between the original and remarked coordinates, divided by the inter-pupil distance. The MUCT-trained ASM result from the left of Figure 10 is shown for reference.

IV. EXISTING LANDMARKED FACE DATABASES

This section is a review of publicly available manually landmarked 2D face databases. We believe that this review is comprehensive at the current time (Sep 2010), although it is possible that our web searches missed a database. Note that we are interested here only in databases with complete sets of landmarks, and thus ignore databases such as the well-known FERET [10] and CMU [11] sets, which have landmarks only for the eye pupils. We also ignore 3D databases such as the Bosphorus set [12].

Examples from the databases are shown in Figure 8. The key statistics of the databases are summarized in Table II. In alphabetical order, the databases are:

- **The AR database** (Purdue University [13])
  The manually landmarked subset of the AR database consists of 508 768x576 color images of 126 subjects with 22 landmarks. The faces were shot with a neutral expression, or a smile, or an angry expression (usually very mild), or with the left light on. The pose is nominally frontal but with considerable variation. The AR, BioID, and XM2VTS images were manually landmarked by the Face and Gesture Recognition Working Group (FGnet) [3].
- **The BioId database** (BioID AG [6])
  The BioId dataset consists of 1521 384x286 images of 23 subjects with 20 landmarks. The pose is nominally frontal but with considerable variation. There is more variety of face sizes than in the other databases, although the mean face size is smaller (see Table II). There is a wider variety of expressions (such as smiles) than the XM2VTS data. The background is an office interior with stable lighting.
- **The IMM database** (Technical University of Denmark [14])
  The IMM database consists of 240 640x480 images of 40

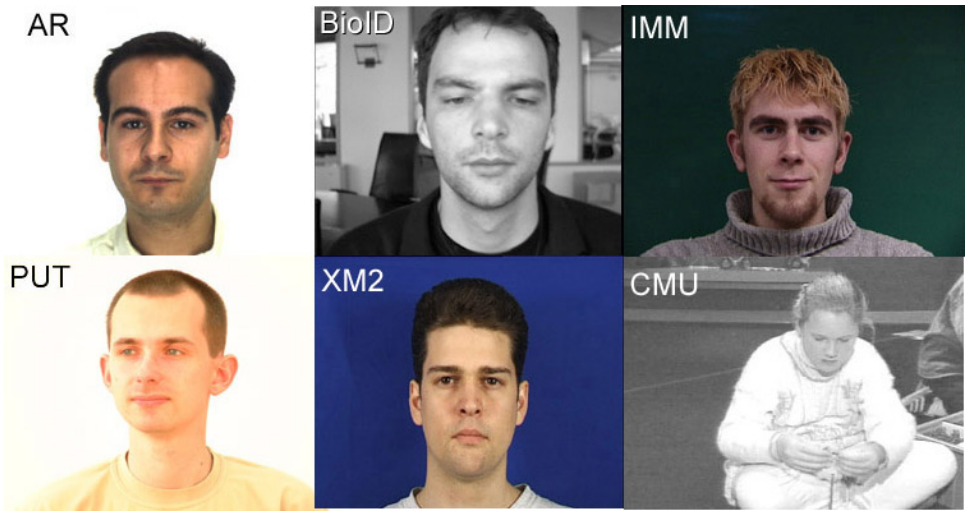

Fig. 8. Examples from other databases

TABLE II
STATISTICS OF PUBLICLY AVAILABLE LANDMARKED FACE DATABASES

	MUCT	AR	BioId	IMM	PUT	XM2VTS
Nbr of landmarked images[a]	3755	508	1521	240	2193	2360
Nbr of landmarks	76	22	20	58	199	68
Nbr of subjects	276	126	27	40	200	295
Image size (pixels)	480x640	768x576	384x286	648x480	2048x1536	720x576
Mean inter-pupil dist (pixels)	88	112	54	?	280	101
Coeff of var inter-pupil dist[b]	0.08	0.07	0.21	?	0.10	0.06
Image color	color	color	mono	color/mono	color	color
Nbr of lighting sets	10	2[c]	1	2	1?	1?
Background[d]	blue	white	office	green	various	blue
Female percent	51	41	40	20	11	47
Wearing glasses percent	18	31	32	0	0	36
Viola Jones fail percent [7][e]	1.8	1.2	4.3	0.0	0.2	0.7
Rowley fail percent [8]	3.8	4.9	12.0	0.8	0.5	3.6
Rowley eye fail percent[f]	11.8	9.6	24.7	4.1	1.2	10.7

[a] Only landmarked images are counted (some of these databases include images that are not landmarked).
[b] The coefficient of variation is defined as the standard deviation divided by the mean.
[c] The manually landmarked subset of the AR faces uses only two of the four AR lighting sets.
[d] The background for all databases is flat, except for the BioID faces where the background is a more-or-less stable office scene.
[e] Percentage of faces not found by the Viola Jones detector. We used the OpenCV [9] implementation.
[f] Percentage of the faces for which the Rowley detector did not find both eyes (including images where the face was not found at all). The fail rate of the face and eye detectors can be used as rough measure of the "difficulty" of the images. The fail rates for real world data (such a typical personal photo collection) will typically be much higher than the figures in the table.

subjects with 58 landmarks. Each subject was shot with a fixed set of 6 different poses and lighting conditions.

- **The PUT database** (Poznan University of Technology [15])
  The landmarked subset of the PUT database consists of 2193 2048x1536 color images of 200 subjects with 199 landmarks (the original PUT data had 194 landmarks; five extra landmarks were added by the authors to allow use of the me17 measure described in Section V). Each subject appears in 22 face orientations, all under the same lighting. This database is distinguished by the high resolution of the images. Nearly all subjects are white males in their early twenties with a neutral expression, and none of the subjects wear glasses.
- **The XM2VTS database** (University of Surrey [2])
  The manually landmarked subset of the XM2VTS database consists of 2360 720x576 color images of 295 subjects with 68 landmarks. The pose is nominally frontal but with considerable variation. The lighting is uniform, with a flat background.

Of these databases, the XM2VTS data is arguably the best for training ASMs and similar models because it contains a

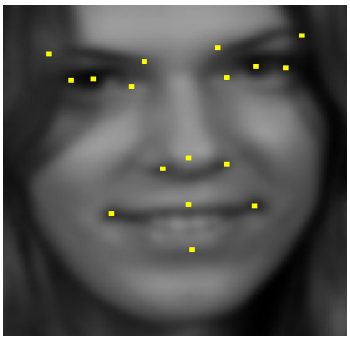

Fig. 9. The me17 landmarks

fairly large number of faces of different types with a fairly large number of landmarks. However, the variety of XM2VTS images is rather limited: the lighting is uniform, most of the subjects are white, the facial expressions are mostly neutral, and the poses are mostly frontal. The MUCT database was developed to address those issues.

The BioID data is useful as a test set because of its variety of face orientations, expressions, and face sizes — although it should be remembered that none of these landmarked databases exhibit the variety of real world data (for example, compare the images from landmarked databases in Figure 8 to the CMU image in the same figure).

V. Active Shape Model Results

In order to characterize the MUCT data, we trained one Active Shape Model (ASM) [1] on the MUCT data and another on the XM2VTS data. To expose the differences between the databases independently of the size of the databases, we used the same number of landmarks and images in both training sets. More precisely, for the XM2VTS training set we used all 68 landmarks and 2360 images; for the MUCT training set we used a matching subset of 68 landmarks, and 2360 images randomly selected from the total of 3755 in the MUCT set.

We used MUCT images from all lighting sets and cameras `a`, `d`, and `e` (ignoring cameras `b` and `c` to avoid complications raised by obscured landmarks). Other subsets would also be reasonable. We then ran both ASMs on the BioID faces. We followed the training protocol described in Milborrow and Nicolls [16], [17]. Note especially that we did not use the BioID test set during model training or tuning.

Following Cristinacce [18], we measured results using the me17 distance. The me17 distance is calculated by taking the mean of the euclidean distances between each of 17 internal face points located by the ASM search (Figure 9) and the corresponding manually landmarked point. This mean is then normalized by dividing by the distance between the manually landmarked eye pupils. The measure, like any, is to some extent arbitrary. It ignores, for example, points on the face perimeter (intentionally, because the points on the perimeter have a high variability across human landmarkers).

The left side of Figure 10 shows the results of this test. We see that, on the BioID data, the MUCT-trained ASM outperforms the XM2VTS-trained ASM.

The MUCT data has a wider range of mouth shapes than the XM2VTS data, and we would expect it to be better for training ASMs that must deal with a range of mouth shapes. The right side of Figure 10 shows that to be so.

We also compared the MUCT-trained ASM and XM2VTS-trained ASM on the PUT images. The PUT faces can be considered to be "easy" — they are high quality uniformly lit faces without glasses, and most of the subjects are young white males, a group well represented in the XM2VTS training data. The left side of Figure 11 shows that the two sets of results on the PUT data are comparable.

We mention that the MUCT results are a little better if all 76 landmarks are used for training (right side of Figure 11).

Summarizing, a MUCT-trained ASM is better able to deal with a wider variety of faces than an XM2VTS-trained ASM, and gives comparable results on "easy" faces. Admittedly that is not a completely water-tight conclusion, because we measured results only on the BioID and PUT data, used only the me17 measure, and ignored statistical uncertainty. (Statistical uncertainty is difficult to estimate for such tests, however the above results were very similar when we repeated the tests independently on each half of the BioID data.)

We emphasize that our use of ASMs here is for illustrative purposes and should not be taken to imply that the MUCT data is suitable only for training ASMs.

VI. Conclusion

Although we used an Active Shape Model in the example above, the MUCT data should be suitable for training and evaluating a wide assortment of models. The data and software to reproduce the results in this paper may be downloaded from www.milbo.org/muct. We hope that other researchers will find the database useful.

Acknowledgments

A thanks goes out to the people who allowed their faces to be used in the database. We also thank Oliver Walker, Liz Walker-Watts, and Gill Andrew for the tedious work of defining over 280 thousand landmarks. We thank David Root of the UCT Research Ethics Committee for his help. And finally we thank Duke Metcalf for making available the room used for photographing the subjects.

References

[1] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, *Active Shape Models — their Training and Application*. Comput. Vis. Image Underst., 1995.

[2] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, *XM2VTS: The Extended M2VTS Database*. AVPBA, 1999.

[3] FGnet, *Face and Gesture Recognition Working Group*. www-prima.imag.fr/FGnet, 2004.

[4] Unibrain Inc., *Fire-i Digital Camera*. www.unibrain.com, 2008.

[5] GretagMacbeth, *Color Chart*. Pantone, 2007, www.pantone.com.

[6] O. Jesorsky, K. Kirchberg, and R. Frischholz, *Robust Face Detection using the Hausdorff Distance*. AVPBA, 2001.

[7] P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features*. CVPR, 2001.

[8] H. A. Rowley, S. Baluja, and T. Kanade, *Neural Network-Based Face Detection*. PAMI, 1998.

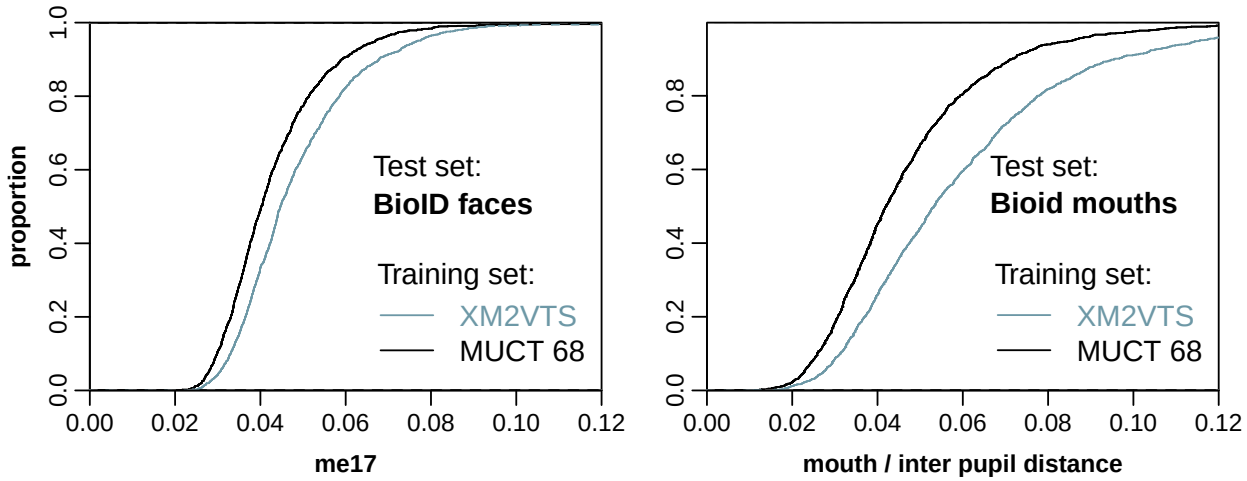

Fig. 10. Comparison of an ASM trained on the MUCT data to an ASM trained on the XM2VTS data. In each training set the same numbers of images and landmarks were used (2360 and 68 respectively). **Left** The results on the me17 points. **Right** The results on just the mouths.

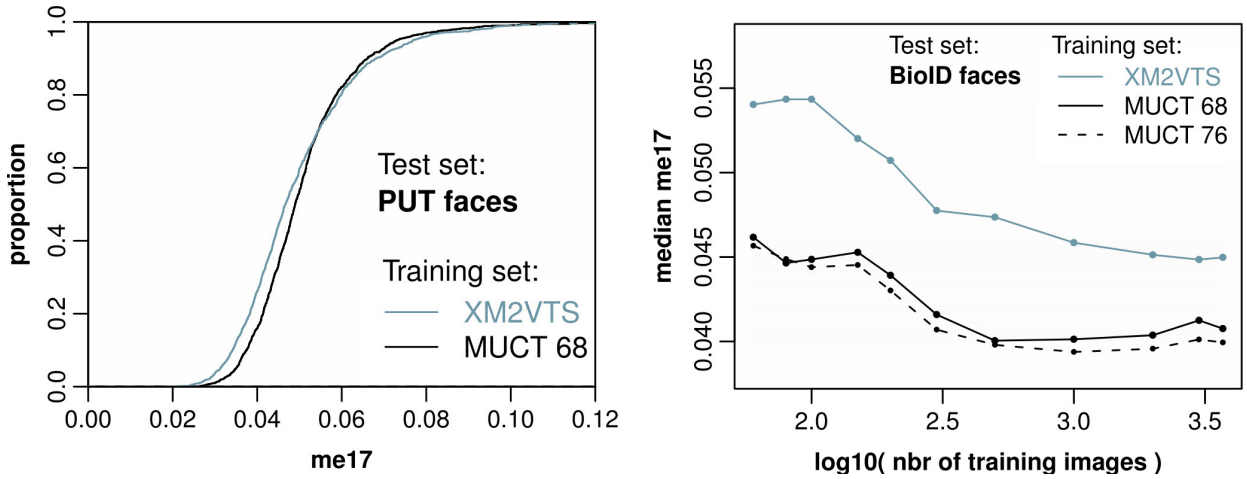

Fig. 11. **Left** Performance on the PUT data. We see that the performance of the two ASMs is comparable on this set of "easy" faces. **Right** Performance on the BioID data for different training sets and sizes. (Both the original and mirrored images were used for training, which made bigger training sets possible.) We see that the 76 point MUCT data gives slightly better results than the 68 point data.

[9] Intel Research, *Open Source Computer Vision Library*. Intel, 2008.

[10] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, *The FERET database and evaluation procedure for face-recognition algorithms*. Image and Vision Computing J. 16(5): 295-306, 1998.

[11] Baluja, Kanade, Poggio, Rowley, and Sung, *CMU Frontal Face Images*. Carnegie Mellon University, Robotics Institute, 1995.

[12] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, *Bosphorus Database for 3D Face Analysis*. COST Workshop on Biometrics and Identity Management (BIOID), 2008.

[13] A. Martinez and R. Benavente, *The AR Face Database*. CVC Tech. Report 24, 1998.

[14] M. M. Nordstrøm, M. Larsen, J. Sierakowski, and M. B. Stegmann, *The IMM Face Database - An Annotated Dataset of 240 Face Images*. Technical Report, Technical University of Denmark, 2004.

[15] A. Kasinski, A. Florek, and A. Schmidt, *The PUT Face Database*. Image Processing and Communications, 13(3-4), 59-64, 2008.

[16] S. Milborrow, *Locating Facial Features with Active Shape Models*. Master's thesis. University of Cape Town, 2007.

[17] S. Milborrow and F. Nicolls, *Locating Facial Features with an Extended Active Shape Model*. ECCV, 2008.

[18] D. Cristinacce and T. Cootes, *Feature Detection and Tracking with Constrained Local Models*. BMVC, 2006.