Multiview Active Shape Models with SIFT Descriptors for the 300-W Face Landmark Challenge

Stephen Milborrow
University of Cape Town
milbo@sonic.net

Tom E. Bishop
Anthropics Technology Ltd.
t.e.bishop@gmail.com

Fred Nicolls
University of Cape Town
fred.nicolls@uct.ac.za

Abstract

The 2013 iBUG 300-W Challenge [25] tests the ability of participants' software to locate landmarks in unseen images of faces. We approach the challenge with Active Shape Models (ASMs [6]) that incorporate a modified version of SIFT descriptors [18]. We make use of multiple ASMs, searching for landmarks with the ASM that best matches the face's estimated yaw. This simple approach is shown to be effective on a variety of datasets.

1. The Challenge

The 2013 iBUG 300-W Challenge [25] tests the ability of participants' software to automatically locate 68 facial landmarks in images of faces. Two hidden test sets are used to evaluate the software. These consist of 300 indoor and 300 outdoor faces with reference landmarks created semi-automatically [24]. Participants submit their software to the challenge organizers who run the software on the test sets and report the results.

Accuracy of fit is measured as the mean distance between the automatically located points and the reference landmarks, divided by the distance between the reference outer eye corners. This fitness measure is reported for all 68 landmarks (what we call the *ec68* measure in this paper) and also for the 51 landmarks interior to the face (the *ec51* measure).

To prevent variation caused by use of different face detectors, face positions are estimated with a standard iBUG face detector. This face detector is not made available to the participants, although for training purposes the detector rectangles are supplied for several thousand example faces. The organizers also supply reference landmarks for these faces.

2. Our Approach

Active Shape Models (ASMs [6]) have long been the starting point for more sophisticated methods of facial landmarking. Nowadays ASMs are typically used with 2D gradient descriptors [22] instead of the 1D gradient profiles of the original method. Such 2D ASMs have performed surprisingly well against more elaborate techniques. For example, 2D gradient ASMs outperformed other automatic landmarkers in three of the four tests in an independent 2013 study [2]. (The study was limited to methods for which software could be freely downloaded.)

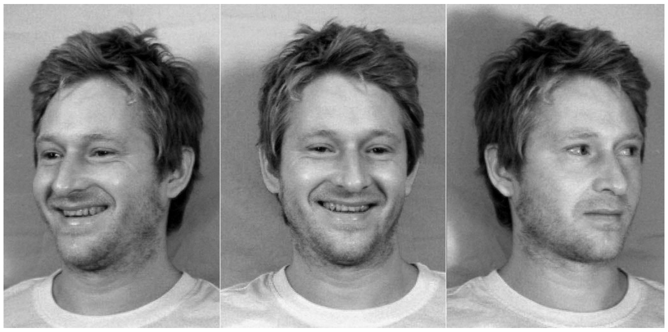

Figure 1. The three canonical yaws used in our model: left three-quarter, frontal, and right three-quarter. (The images are from [21].)

Additionally, the conceptual simplicity of ASMs is beneficial in commercial applications where ease of maintenance and integration is important. The execution speed of ASMs (typically less than one- or two-hundred milliseconds a face) is acceptable for many applications.

Given these considerations we felt it worthwhile to pursue an ASM-based approach for the 300-W Challenge. We have previously used ASMs with SIFT descriptors [18] for template matching [23]. After some engineering of the SIFT implementation, this approach on frontal faces gave significantly better fits and faster search times than 2D gradients. However, performance remained inadequate on non-frontal faces. Therefore for the 300-W Challenge we adopt a simple strategy of three submodels. These are SIFT-based ASMs optimized for frontal views, left three-quarter views, and right three-quarter views respectively (Figure 1). Using the position of the face supplied by the iBUG face detector, we first estimate the face's yaw and in-plane rotation, rotate the image so the face is approximately upright, and then search for landmarks with the submodel that best matches the face's estimated yaw (Figure 2).

As we shall see, this straightforward strategy works well on neutral or smiling faces (wedding photographs) but not so well for faces that are highly expressive (excited spectators at a sporting event). Since there are no submodels for large yaws, good results cannot be expected on side views. However, the use of eye-corner distances for normalization is a hint that there are no such views in the 300-W test sets. Very large pitch variation is also a problem, although our experience has been that ASMs are far more sensitive to yaw than pitch.

In this paper, after a review of related work we discuss how the face's pose is estimated. We then describe the modified ASMs we use as submodels. Finally results are presented and future directions are discussed. The focus is on engineering details because our approach here is essentially to tailor existing methods to the 300-W Challenge.

3. Related Work

At the core of the approach in this paper are our "HAT" ASMs. As described later, and in more detail in [23], these are ASMs that use modified SIFT descriptors for template matching. SIFT descriptors have been used in other face landmarkers. For example, Zhou *et al.* [27] use SIFT descriptors with Mahalanobis distances. They report improved eye and mouth positions on the FGRCv2.0 database. Another example is Li and Ito [15], who use SIFT descriptors with GentleBoost.

The idea of using different submodels for different poses is not new. Cootes *et al.* [7] is an early example. They demonstrate that a small number of models can represent a wide range of poses. Their model, AAM based, can also synthesize views of the face from new angles. The approach presented in the current paper rigidly separates pose detection and landmarking; a limitation is that if the pose is misestimated the wrong ASM is applied with no chance for correction. In contrast it seems that most examples in the literature unify pose and landmark detection. A recent example is Zhu and Ramanan [28]. As in our method, Belhumeur *et al.* [1] use SIFT descriptors and multiple models, but they use a large number of global models, jointly optimizing pose and landmark detection using a probabilistic model with SVMs and RANSAC. They report excellent best-case but not so good fourth-quartile fits, and slow speeds. Kanaujia and Metaxas [12] also use SIFT descriptors and multiple models. They determine pose clusters automatically, unlike our method which simply assumes that three models suffice for the poses of interest. They align the descriptors to the shape border; we do not align the descriptors but rotate the face upright before the search begins. Our approach is more basic than any of the methods mentioned.

An ongoing issue has been the difficulty of objectively comparing algorithms in different papers. Different authors use different landmark definitions, metrics, and test sets. When graphing results, some authors discard faces for which the face detector failed; others include such faces, with a fit of infinity. Some authors do not report search times. It is impossible from reading papers like those given above to determine which algorithms are best for a given application. Given these issues, and the fact that the ranking of models in nonpartisan studies such as Çeliktutan *et al.* [2] differs considerably from that claimed in papers, we welcome projects like the 300-W Challenge.

A popular de facto test set has been the BioID data [11], but its faces are almost all frontal and were photographed in an office setting, not "in the wild". As the state of the art has progressed the need for a more diverse test set of landmarked faces has arisen. Complicating the issue are copyright concerns preventing redistribution of faces obtained from the web. For many of us a test set is most informative if it is representative of the distribution of faces in commercial applications, such as those for Photoshop-style retouching or remodeling, or for virtual makeup or eyewear generation. In this context a test set should not have a too large proportion of faces with extreme expressions (although smiles are important) or of low resolution (eye-mouth distance less than 100 pixels). The test set should contain a sufficient number of faces to minimize over-fitting (optimizing for the test set) and to bound sampling variation, especially in the worst 10% of fits (so we would say at least a thousand faces, more is better). No commonly accepted standard test set has yet emerged.

4. Pose Detection

We now move on to a description of our model. Before the ASM search begins we must choose the appropriate submodel based on the estimated yaw of the face. We use the position of the face determined by the iBUG face detector, but to estimate yaw apply three further face detectors. These are optimized for frontal, left three-quarter, or right three-quarter faces respectively. Each of these detectors uses a truncated cascade of just 12 weak classifier stages, rather than the typical 20 or more stages. This greatly increases the number of detections per face, although also increasing false detections.

Our detectors are Viola Jones detectors [26, 17] with Local Binary Patterns [16] built with the OpenCV tools on the AFLW dataset [13]. The AFLW data was used for training because apart from its wide variety of faces it also provides the estimated pose of each face, making it possible to select subsets with the range of yaws appropriate for training each detector. Note that we do not need the AFLW manual landmarks.

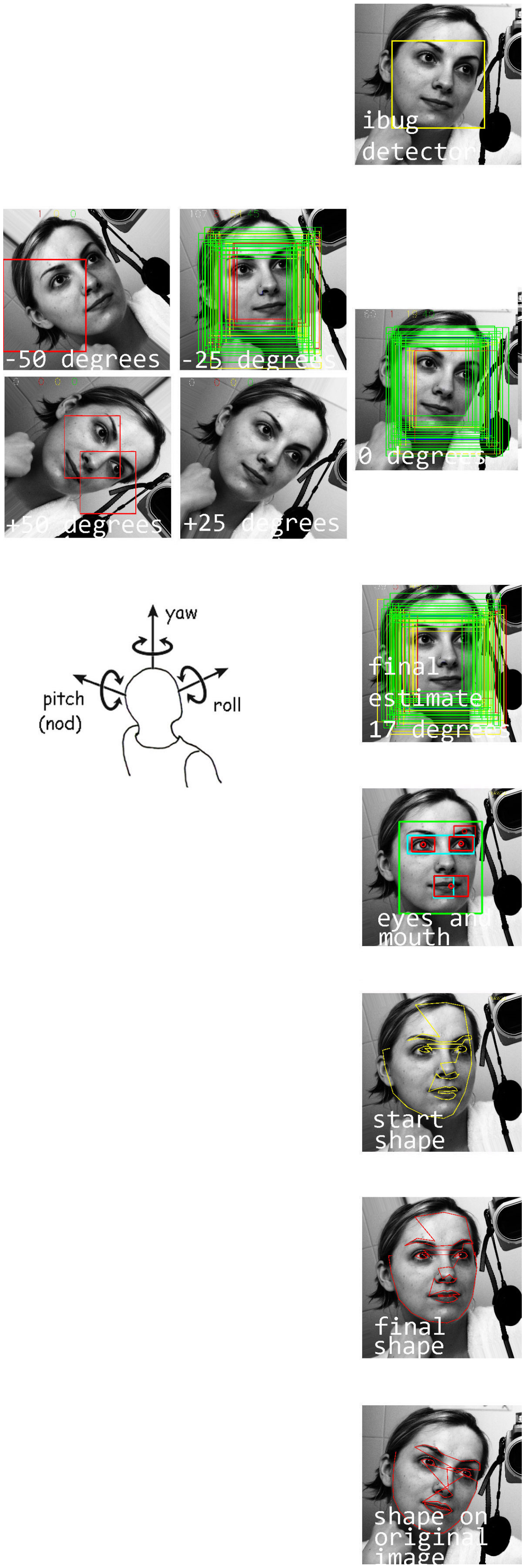

Figure 2. **Overview Of Our Method.**

(a) The original image with the iBUG face detector rectangle.

(b) **Estimate Rotation** The image is rotated by -50, -25, 0, 25, and 50 degrees. The left three-quarter, frontal, and right three-quarter face detectors are applied to each rotated image in the area around the iBUG rectangle.

In this example, the -25 degrees image has the largest number of detects. The estimated rotation (17 degrees) is a thus a weighted sum of -50, -25, and 0 degrees. The weights are the number of detects in the -50, -25, and 0 degree images. The false detects on the highly rotated faces are essentially ignored.

(c) **Estimate Yaw** After estimating the in-plane rotation of the face, the image is rotated so the face is upright. We work with this rotated image until the ASM search is complete.

The three face detectors are re-applied and a MARS model estimates the face's yaw using the relative number of detections by each detector on the upright face. In this example the estimated yaw is 20 degrees.

(d) **Estimate Eye And Mouth Positions** Eye and mouth detectors are applied. (The eye and mouth positions will be used to help position the start shape.) The false detection on the right eyebrow is ignored because its centroid is outside the legal zone (cyan).

(e) **Start Shape** The start shape for the right three-quarter ASM is aligned to the face, using the face detector rectangle and the eye and mouth positions if available. We use the right three-quarter model because the yaw was estimated to be 20 degrees in step (c) above.

(f) **Landmark Search** The face is scaled to an eye-mouth distance of 100 pixels and the right three-quarter HAT ASM is applied (Section 5.3). The image to the left shows the final shape after the ASM search.

(g) **Back to Image Frame** The final shape is derotated and scaled back to the original image. The shape is also converted to the sixty-eight iBUG points.

4.1. Rotation

In-plane rotation is used as a surrogate for the face's roll. To estimate the rotation, we first generate images rotated by -50, -25, 0, 25, and 50 degrees (Figure 2b). On these rotated images we apply the three face detectors independently in the area around the iBUG face rectangle. Face rectangles that are too far from the median rectangle in each image are discarded as false positives. One of the rotated images will have the largest number of detections. The estimated rotation of the face is taken to be the weighted sum of this image's rotation angle and the rotations to each side of it, with the weights being the number of detections in the three rotated images.

4.2. Yaw

The face's yaw is estimated from the number of detections by each detector on the upright face. We make this estimation with a Multivariate Adaptive Regression Spline model (MARS [9]). The MARS model was trained by repeating the process in Section 4.1 on 2000 AFLW images, for each face regressing the ground truth yaw on the counts of the three detectors on the upright face. Intuitively, MARS is appropriate for this regression because the yaw can be estimated as a weighted sum of the counts but with some adjustment for non-linearity (Figure 3).

We also trained an SVM to estimate the yaw from the histogram-equalized patch in the face detector rectangle. This technique did not give as good estimates of yaw and we do not use it.

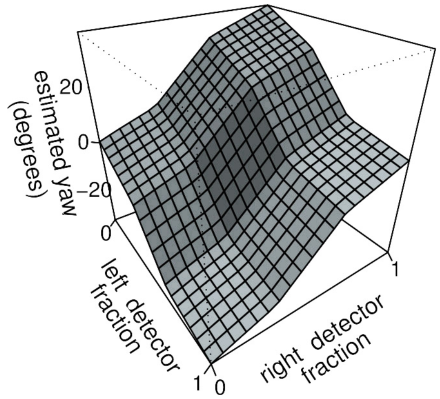

Figure 3. MARS model to estimate the face's yaw from the number of detections by the left three-quarter and right three-quarter face detectors (as a fraction of the total number of detections by all three detectors). This model gives a regression R^2 of 0.79 on the training data. For a linear model the above surface would be a plane with an R^2 of 0.76.

5. The ASMs

Per-face yaw data is needed to train our pose-specific models. We thus cannot (easily) train our ASMs on the supplied 300-W training data. Instead, the models were trained on the MUCT data [21]. The AFLW data [13] was also considered for training, but not used because it has an insufficient number of landmarks for our purposes.

5.1. Conversion to iBUG Landmark Definitions

After the ASM search we must convert the MUCT landmarks to the 68 iBUG definitions (which follow the Multi PIE definitions [10]). This is done either by copying points (*e.g.* eye corners) or by interpolating between nearby points (*e.g.* jaw landmarks). Interpolation constants were generated with the iBUG re-marked Helen training data. (We also used the same data for tuning other aspects of our model.)

Because the landmarks are defined differently this conversion introduces inaccuracies, a significant disadvantage. Even the points we copy are not defined identically. Our hope is that in general the conversion noise is small compared to other inaccuracies, but we do not expect our best case fits to match those of models trained directly on the 300-W data.

5.2. Training Details

The frontal model was trained on the 6008 frontal and near frontal images of the MUCT `abde` subset (using mirroring to bring the number of faces from 3004 to 6008). This model is available as open source software from `http://www.milbo.users.sonic.net/stasm`. The images include faces with some pitch variation and natural minor tilting of subjects faces, thereby training the model to cope with such variation. Explicitly rotated faces are unnecessary because we rotate the face upright before starting the search.

The three-quarter models were trained on the much smaller set of 1502 yawed faces of the MUCT `bc` subset. We artificially increased the size of this set by applying x- and y-stretching, changing the intensity gradient across the face, and applying small rotations to the descriptor patches around the landmarks.

5.3. HAT ASMs

First we give a brief overview of ASMs in general [5] then describe our variant of the ASM.

An ASM starts the search for landmarks from the mean training face shape aligned to the position and size of the image face determined by a face detector. It then repeats the following two steps until convergence:

(i) Suggest a tentative shape by adjusting the locations of shape points by template matching around each point. In the classical ASM this is done by matching 1D gradient profiles along a search line orthogonal to the shape boundary.

(ii) Conform the tentative shape to a global shape model. This pools the results of the matchers to form a stronger overall classifier. Each individual template matcher examines only a small portion of the face and cannot be completely reliable.

The entire search is repeated at each level in an image pyramid, typically four levels from coarse to fine resolution.

Our variant of the ASM modifies step (i) above, leaving step (ii) unchanged. Specifically, for template matching at landmarks we replace the 1D gradient profiles of the classical ASM with "Histogram Array Transform" (HAT) descriptors. We search for the best position of a landmark by searching in a 2D area around the current position of the landmark. This is done by taking a 15 × 15 patch around each search point, forming a HAT descriptor from the patch, and matching the descriptor against the model descriptor. The new position of the landmark is taken to be the point in the search area that gives the best match.

Like SIFT descriptors [18], HAT descriptors are grids of image orientation histograms, with orientations weighted by the gradient magnitude and smoothed out over nearby histogram bins. Before starting the ASM search for points, we prescale the face to a fixed eye-mouth distance of 100 pixels and rotate the face so it is upright. Therefore the extreme scale invariance of SIFT is not required, nor is the SIFT descriptor's automatic orientation of the patch to the local average gradient direction. By not orienting the patches we gain both fit and speed advantages. Some rotational variance will still remain in the upright face (not every face is the same, and the eye detectors sometime give false positives on the eyebrows or fail to find eyes, causing mispositioning of the ASM start shape), and so we must also rely on the intrinsic invariance properties of the descriptors.

Full details of our HAT ASM may be found in [23], where we show that HAT ASMs give significantly better fits and faster search times than 2D gradient ASMs.

6. Results

We start our presentation of results with Figure 4, which shows the fit for faces at different yaws with different submodels. The combined model (black curve) achieves the best fit across all yaws by automatically selecting the ASM appropriate to the yaw, as described in Section 4. However, misestimates of yaw are not uncommon, as can be seen from misplaced colored dots. The effect of these misestimates is mitigated by the partial robustness of the models to poses out of their yaw range, but do cause the poorer performance of the black curve against the red and green curves on the edges of the graph.

Note also from the figure that fits are generally worse on non-frontal faces, even with submodels. The smaller training set used for our three-quarter models may be playing a part here (Section 5.2). More importantly, yawed faces have lower inter-eye pixel distances, so any fitness measure that uses inter-eye distances for normalization upweights perceived error on yawed faces.

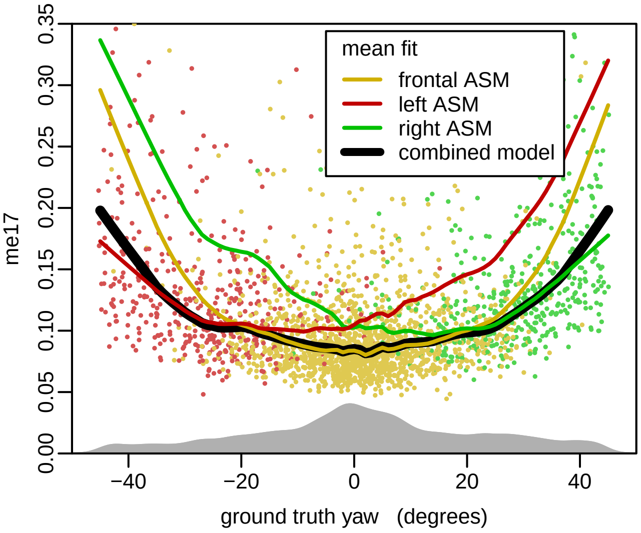

Figure 4. Fit versus ground truth yaw using different submodels applied individually, then combined. The dots show the fit on each face using the combined model, with their color indicating the ASM submodel that was selected using the estimated yaw. The shaded curve on the bottom shows the density of faces in the test set. Fit is measured here using the me17 measure [8] on 3000 faces iid from the AFLW set [13] with ground truth yaws ranging from -45 to 45 degrees.

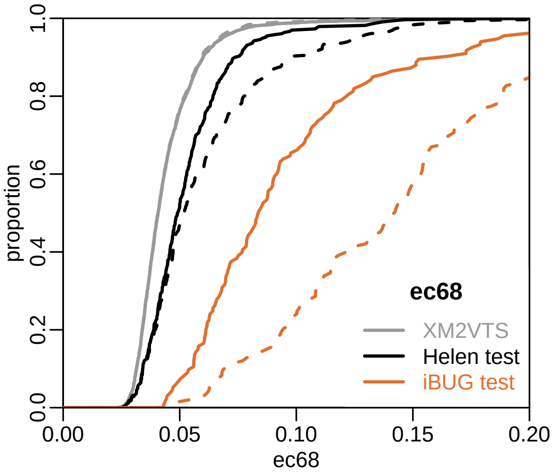

Figure 5. Combined model (solid lines) versus frontal-only model (dashed lines). The combined model gives better fits on datasets with non-frontal faces. (The solid curves are the identical to the curves of the same color in Figure 9. Please see the description of that figure for details on the fitness measure and datasets.)

Figure 5 shows the cumulative distribution of fits on various datasets using the combined model (solid lines) and the frontal-only model (dashed lines). For the XM2VTS set, with nearly all frontal faces, the frontal-only model suf-

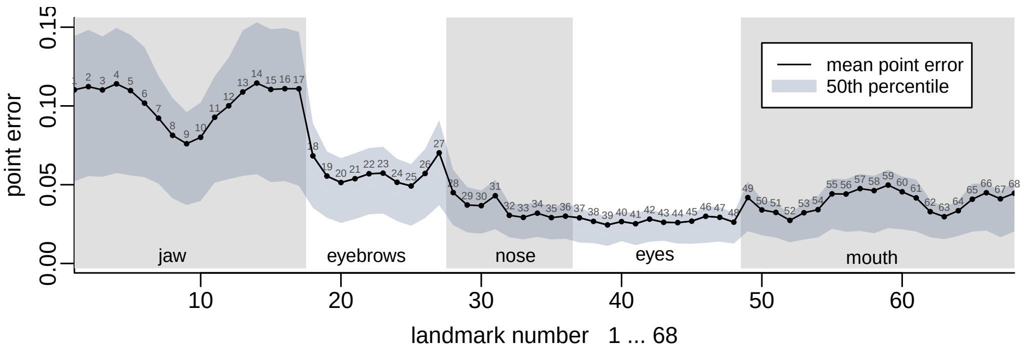

Figure 6. Per landmark error of our method. Measured on the 2000 faces of the re-marked iBUG Helen training set [25]. Error here is the distance between the automatic and reference landmarks, divided by the reference eye corner distance.

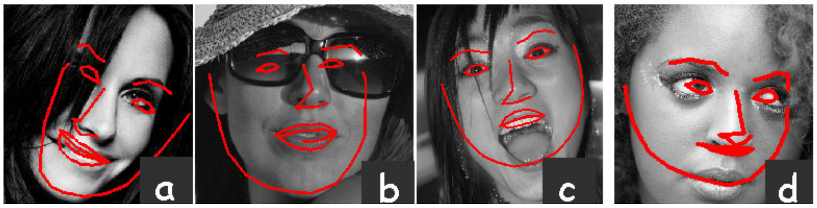

Figure 7. Pathologies
(a) yaw was incorrectly estimated as being leftwards, so the wrong submodel was used
(b) eye locations confused by sunglasses
(c) mouth open, confused bottom lip for chin
(d) confused base of nose for mouth, and mouth for chin.

fices; for sets with three-quarter views the combined model gives better fits. The slight degradation of fits of the combined model on the XM2VTS set (barely perceptible in the graph) is caused by misestimates of yaw.

Figure 6 shows the mean fit error per landmark of the combined model. The figure shows that positioning of points interior to the face is better than that of the exterior points. Of the interior points, the eyebrows fare the worst, especially the outer eyebrows (landmarks 18 and 27). Error is especially pronounced in the upper jaw. Although our model's conversion from MUCT to iBUG points contributes to inaccuracies here (Section 5.1), we suspect that the curve would have approximately the same shape for any model because it reflects the level of inherent ambiguity in the landmark positions.

Figure 7 shows representative examples of bad results with our method. The yaw was misestimated in Figure 7a because the hair across the side of the face was mistaken as background by the face detectors. Shadows across the face can similarly mislead the face detectors. Sunglasses as in Figure 7b are nearly always a problem. Perhaps a sunglass detector could be used to indicate that the local texture in the region of the eyes is not to be trusted, and that the position of the eyes should be determined by the overall shape model or simply marked as not available. Likewise an open mouth detector could resolve cases like Figure 7c. However in general we prefer to avoid such accretions to the algorithm.

Figure 8 shows the results of running our model on the 300-W hidden test sets. Consistent with Figure 6, results are better on the 51 interior points than across all 68 points.

Figure 9 shows the results of running our model on the combined hidden test set, and on the datasets supplied on the 300-W web page. Not surprisingly we do best on the clean frontal faces of the XM2VTS set. The Helen test set is perhaps closest to the test set described in the last paragraph of Section 3 (although too small). The Helen training set is not shown because it was used when training the models (Section 5.1), invalidating its use as a test set.

Figure 10 shows the linear relationship between the median fit and diversity of the faces. It plots for each dataset the median ec68 versus the standard deviation of the ratio

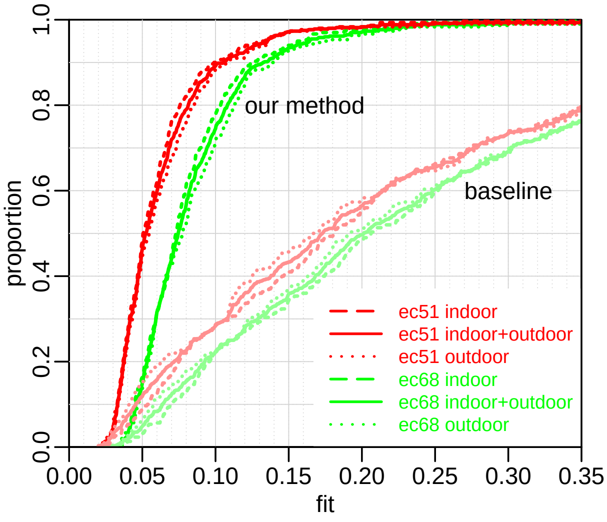

Figure 8. Fits for our method on the 300-W hidden test sets as reported by the organizers of the challenge. Also shown are the results for the baseline method (which is an iBUG implementation of the method in [19] using the edge-structure features in [4]). The solid red and green curves are identical to the solid red and green curves in Figure 9.

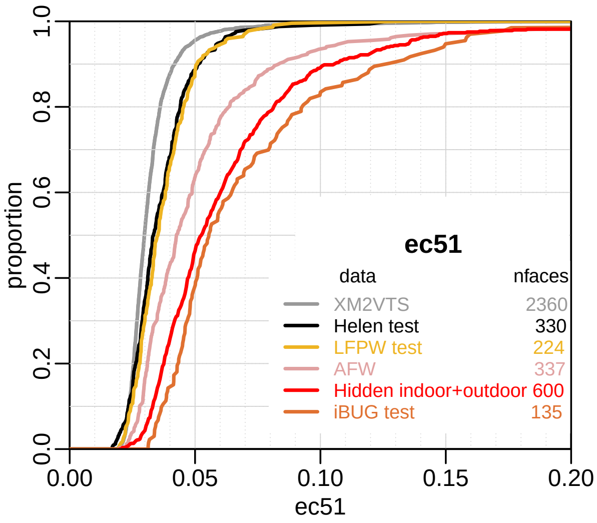

Figure 9. Fits for our method on the hidden test set and on the datasets provided on the 300-W web page [25].
Note that the reference landmarks in this and related figures are the iBUG re-marked landmarks [25], not the landmark data from the original databases (XM2VTS [20], Helen [14], LFPW [1], and AFW [28]).
Left The mean distance between the 51 automatically located points internal to the face and the reference points, divided by the distance between the reference eye outer corners.
Right The same measure for all 68 landmarks.

of the eye-mouth distance to the eye-corner distance. (The eye-mouth distance is measured from the mean of the eye outer corner coordinates to the bottom of the bottom lip, point 58). This “aspect ratio” will vary as the pose varies and as the mouth opens. Its standard deviation across a set of faces is a convenient measure of the diversity of that set. Its mean value for neutral frontal adult faces is coincidentally close to 1 (its mean value over the XM2VTS faces for example is 0.98).

Figure 11 shows search times. In the current unoptimized implementation, detecting the pose takes about half the total time. The mean eye-mouth distance in pixels is shown after the set name on the horizontal axis. We see that pose detection takes longer on bigger faces, because the Viola Jones detectors are slower.

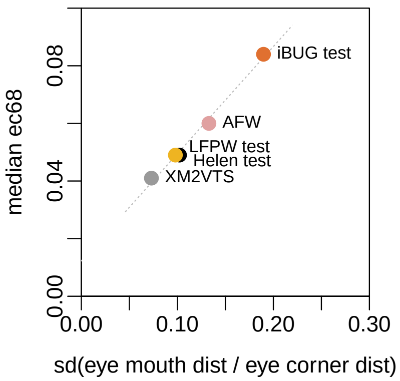

Figure 10. Median fit versus standard deviation of the face “aspect ratio”, a measure of diversity. Diversity causes worse median fits. The dots correspond to the points on the median line (proportion 0.5) in the right figure of Figure 9.

7. Future Directions

Increasing the number of basic poses above the three in our current multiview model is an obvious future direction.

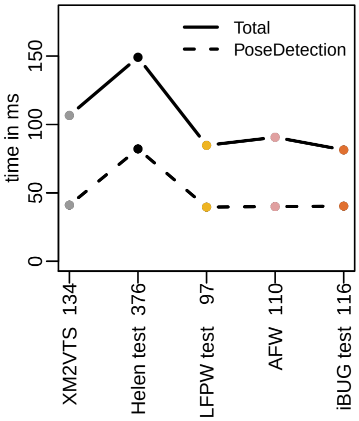

Figure 11. Search times.
Total is for all steps except reading the image from disk and once-off model initialization. Face detection by the iBUG face detector is not included.
PoseDetection is for steps b and c in Figure 2.
Times were measured on a 3.4 GHz i7 with the same datasets as Figure 9.

Certainly models for side views are necessary to handle the full gamut of faces. Models for pitched views may also help, but one quickly reaches the point of diminishing or even negative returns, a problem compounded by combinatorial explosion of poses and misestimates by our crude pose detector.

Ad hoc additions to the model to handle some of the worst cases in Figure 7 may be of some benefit, but would affect only a small proportion of faces in realistic test sets and may have unwanted side effects in the bulk of the faces.

Performance would probably be improved with a better training set (we trained on the MUCT faces which were not photographed in the wild and have a limited number of three-quarter views).

A disadvantage of ASMs is the lack of wider context around a point during template matching, especially at fine resolutions. The descriptors of Belhumeuer *et al*. [1] may help in this regard, and would fit easily into our scheme. They form their descriptors by concatenating two SIFT descriptors at different image resolutions — the lower resolution descriptor provides context for the higher resolution descriptor. Of approaches that also make use of the context around the landmark (but without impractical search times), another avenue could be regression based models like the Random Forest model of Cootes *et al*. [3].

For efficiency the integral images required by the Viola Jones face detectors should be shared rather than redundantly regenerated afresh for each of the three detectors. Our current implementation applies the face detectors and template matchers sequentially, but straightforward parallelization on multicore processors could significantly reduce the times seen in Figure 11.

We thank the organizers of the 300-W Challenge for their patient and attentive responses to our queries.

References

[1] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. *Localizing Parts of Faces Using a Consensus of Exemplars*. CVPR, 2011. 2, 7, 8

[2] O. Çeliktutan, S. Ulukaya, and B. Sankur. *A Comparative Study of Face Landmarking Techniques*. EURASIP Journal on Image and Video Processing, 2013. http://jivp.eurasipjournals.com/content/2013/1/13/abstract. 1, 2

[3] T. Cootes, M. Ionita, C. Lindner, and P. Sauer. *Robust and Accurate Shape Model Fitting using Random Forest Regression Voting*. ECCV, 2012. 8

[4] T. F. Cootes and C. J. Taylor. *On Representing Edge Structure for Model Matching*. CVPR, 2001. 6

[5] T. F. Cootes and C. J. Taylor. *Technical Report: Statistical Models of Appearance for Computer Vision*. The University of Manchester School of Medicine, 2004. http://www.isbe.man.ac.uk/~bim/Models/app_models.pdf. 4

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

[7] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. *View-Based Active Appearance Models*. Image and Vision Computing, 2002. 2

[8] D. Cristinacce and T. Cootes. *Feature Detection and Tracking with Constrained Local Models*. BMVC, 2006. 5

[9] J. H. Friedman. *Multivariate Adaptive Regression Splines (with discussion)*. Annals of Statistics, 1991. http://www.salfordsystems.com/doc/MARS.pdf. 4

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. *Multi-PIE*. IVC, 2010. 4

[11] O. Jesorsky, K. Kirchberg, and R. Frischholz. *Robust Face Detection using the Hausdorff Distance*. AVBPA, 2001. 2

[12] A. Kanaujia and D. Metaxas. *Large Scale Learning of Active Shape Models*. ICIP, 2007. 2

[13] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. *Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization*. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011. 2, 4, 5

[14] V. Le, J. Brandt, Z. Lin, L. Boudev, and T. S. Huang. *Interactive Facial Feature Localization*. ECCV, 2012. 7

[15] Y. Li and W. Ito. Shape Parameter Optimization for AdaBoosted Active Shape Model. *ICCV*, 1:251–258, 2005. 2

[16] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li. *Learning Multi-scale Block Local Binary Patterns for Face Recognition*. International Conference on Biometrics, 2007. 2

[17] R. Lienhart and J. Maydt. *An Extended Set of Haar-Like Features for Rapid Object Detection*. IEEE ICP, 2002. 2

[18] D. G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. IJCV, 2004. 1, 5

[19] I. Matthews and S. Baker. *Active Appearance Models Revisited*. IJCV, 2003. 6

[20] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. *XM2VTS: The Extended M2VTS Database*. AVBPA, 1999. 7

[21] S. Milborrow, J. Morkel, and F. Nicolls. *The MUCT Landmarked Face Database*. Pattern Recognition Association of South Africa, 2010. 1, 4

[22] S. Milborrow and F. Nicolls. *Locating Facial Features with an Extended Active Shape Model*. ECCV, 2008. 1

[23] S. Milborrow and F. Nicolls. *Active Shape Models with SIFT Descriptors and MARS*. VISAPP, 2014. 1, 2, 5

[24] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. *A Semi-Automatic Methodology for Facial Landmark Annotation*. CVPR AMFG Workshop, 2013. 1

[25] G. Tzimiropoulos, S. Zafeiriouk, and M. Pantic. *300 Faces in-the-Wild Challenge (300-W)*. ICCV Workshop, 2013. http://ibug.doc.ic.ac.uk/resources. 1, 6, 7

[26] P. Viola and M. Jones. *Rapid Object Detection using a Boosted Cascade of Simple Features*. CVPR, 2001. 2

[27] D. Zhou, D. Petrovska-Delacrétaz, and B. Dorizzi. *Automatic Landmark Location with a Combined Active Shape Model*. BTAS, 2009. 2

[28] X. Zhu and D. Ramanan. *Face Detection, Pose Estimation and Landmark Localization in the Wild*. CVPR, 2012. 2, 7