Facial Landmark Localization in Depth Images using Supervised Ridge Descent

Necati Cihan Camgöz
Computer Engineering Department
Boğaziçi University
Istanbul, Turkey
cihan.camgoz@boun.edu.tr

Vitomir Štruc
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
vitomir.struc@fe.uni-lj.si

Berk Gokberk
Computer Engineering Department
MEF University
Istanbul, Turkey
berk.gokberk@mef.edu.tr

Lale Akarun, Ahmet Alp Kındıroğlu
Computer Engineering Department
Boğaziçi University
Istanbul, Turkey
akarun, alp.kindiroglu@boun.edu.tr

Abstract

Supervised Descent Method (SDM) has proven successful in many computer vision applications such as face alignment, tracking and camera calibration. Recent studies which used SDM, achieved state of the-art performance on facial landmark localization in depth images [4]. In this study, we propose to use ridge regression instead of least squares regression for learning the SDM, and to change feature sizes in each iteration, effectively turning the landmark search into a coarse to fine process. We apply the proposed method to facial landmark localization on the Bosphorus 3D Face Database; using frontal depth images with no occlusion. Experimental results confirm that both ridge regression and using adaptive feature sizes improve the localization accuracy considerably.

1. Introduction

Landmark localization is a crucial initial step for face processing applications. Such applications include but are not limited to biometrics [1], facial expression recognition [15], age estimation [8] and sign language recognition [3]. In biometrics applications, the localized landmarks are used to align faces before matching or to extract local features. On the other hand, in facial expression analysis and sign language recognition, the landmarks are tracked through time to extract features in the spatio-temporal domain. For all these different applications a better landmark localization results in a better performance of the overall system. Most of the systems use 2D images since 2D images are easy to acquire using commonly available 2D cameras. However, 2D face images are vulnerable to illumination and pose changes. The availability of inexpensive depth cameras has led to the widespread use of 3D face images, which overcome these difficulties. Therefore, the development of a reliable 3D facial landmark localization method has become essential.

Facial landmark localization methods generally utilize heuristic approaches as well as statistical methods. Heuristics rely on unique properties of the facial landmarks on the face: For example, the nose tip resides on the symmetry axis of the face and can be localized using the shape properties. Similarly, the corners of the eye and mouth can easily and successfully be localized by heuristics using shape properties. Such an example to these methods is [1] in which Alyüz et al. propose a heuristic method which uses curvature information, symmetry axis and shape index to locate the nose tip, nose and eye corners in 3D faces.

Statistical 3D landmark localization methods also exploit the features of facial landmarks such as local texture and shape. Unlike heuristic-based approaches which require a unique rule for each landmark, feature statistics are utilized in a uniform approach for all landmarks. Most recent statistical methods also use the shape information represented by the facial landmarks. Creusot et al. [5] propose a statistical facial landmark localization method utilizing shape information in addition to the local features of landmarks. Several candidates are identified on a local 3D mesh and the most probable candidate is identified through shape analysis. Another statistically motivated method using shape information proposed by Sukno et al. [14] localizes facial landmarks under occlusion and expression changes. In [14], the shape context of facial landmarks is used together with local feature analysis. Different subsets of candidate points

are evaluated, resulting in robustness against missing landmarks due to occlusions. A similar concept for estimating occluded 3D landmarks is also proposed in [2] where partial Gappy Principal Component Analysis is used to restore missing landmark coordinates. In another study Farrelli et al. [9] proposed a random forest based framework in which patches extracted from depth images cast votes to localize facial landmarks.

The Supervised Descent Method (SDM) [16] was proposed to solve nonlinear optimization problems by turning the problem into a least squares form and applying regression. In 2D domain, SDM has been proven to be successful for facial landmark localization. Recently Camgoz et al. [4] achieved state of the art performance on facial landmark localization in 3D depth images using SDM. They experimented with Scale-Invariant Feature Transform (SIFT) [11] and Histogram of Oriented Gradients (HOG) [6] features for localization and showed that both methods yield accurate localization results. Taking [4] as a baseline, we propose to use ridge regression for Supervised Descent Method (which we call *Supervised Ridge Descent*) instead of least squares regression for facial landmark localization in depth images. Additionally we propose to change feature sizes in each iteration in a coarse to fine fashion. In this way, we aim to capture more details in later iterations by focusing on smaller regions.

In Section 2, we briefly explain the Supervised Descent Method and give the details of ridge regression extension. In Section 3 we report the experimental results conducted on the Bosphorus 3D face database and compare the performance of the proposed method with the state of the art approaches. Finally, we evaluate the findings of this study and discuss future work in Section 4.

2. Proposed Method

2.1. Supervised Descent Method (SDM)

The Supervised Descent Method has achieved state of the art performances in several computer vision applications which previously relied heavily on nonlinear optimization methods [16, 17]. Xiong et al. [16] proposed to approach the non-linear optimization by learning the descent directions from a training set and then use these previously learned descent directions on new unseen test samples. SDM's best known application is facial landmark localization, also known as the IntraFace [7]. It has been used to achieve state of the art performances in face tracking and alignment.

Facial landmark localization using SDM starts with creating an average face shape which provides initial landmark locations for the facial images. At the beginning of the training, landmarks are placed in these initial locations (x_0). Then the shape increment (Δx) required to displace the landmarks from their current location (x_k) to its ground truth location (x_*) is calculated. This is written as a function of the features extracted from the current shape estimate (ϕ_k) as:

$$\Delta x_k = x_* - x_k = R_k \phi_k + b_k \tag{1}$$

To estimate the parameters of this function, R_k and b_k, the problem is written in least squares format as in Equation 2, where i and k represent the sample and iteration indices, respectively.

$$\underset{R_k, b_k}{\operatorname{argmin}} \sum_{x_k^i} \left\| \Delta x_k^i - R_k \phi_k^i - b_k \right\|^2 \tag{2}$$

By using the closed form solution of least squares regression, both R_k and b_k parameters are estimated. Then R_k and b_k are used to update the location of the landmark as:

$$x_{k+1} = x_k + R_k \phi_k + b_k \tag{3}$$

The training procedure continues until the landmarks converge to the actual positions. When a test sample comes, landmarks are placed in their initial positions (x_0) and their positions are updated using Equation 3.

2.2. Supervised Ridge Descent (SRD)

SDM was originally designed to use least square regression (LSR) to estimate its predictor parameters. While using LSR, one needs to take the inverse of the $X^T X$ matrix, X being the observations of predictors. However, the $X^T X$ matrix becomes singular when the observation size is large and/or the predictors are strongly correlated. To overcome the singularity issue Xiong et al. [16] proposed to use PCA to regularize their matrix before taking the inverse of it.

In this study we propose to use *ridge regression* (RR) instead of LSR, in which the matrix singularity issue is dealt by adding a $\Gamma^T \Gamma$ matrix to the $X^T X$ matrix, Γ being the regularization term which is proportional to the identity matrix. Although we lose precision by taking the inverse of $\Gamma^T \Gamma + X^T X$ instead of $X^T X$, we avoid over-fitting and large variances in the estimators.

Our formalization of ridge regression can be seen in Equation 4, in which β_k, λ_k, b_k represent the estimator, regularization term and offset parameter of the k^{th} iteration, respectively. The rest of the parameters Δx_k^i and ϕ_k^i represent the landmarks' distance from the ground truth and their features in these positions of the i^{th} sample, respectively. As in [4] and [17] we used HOG features as facial landmark descriptors. However, in each iteration, the size of the HOG features and the regularization term's value has been decreased to be able to descend more precisely to the ground truth.

$$\underset{\beta_k, b_k}{\operatorname{argmin}} \sum_{x_k^i} \left\| \Delta x_k^i - \phi_k^i \beta_k - b_k \right\|^2 + \left\| \lambda_k \beta_k \right\|^2 \quad (4)$$

To calculate the ridge regression estimator, β_k, for each iteration, we use Equation 5 in which I and λ_k represent the identity matrix with the same size as the observation matrix and the regularization term of the k^{th} iteration. Φ_k and ΔX_k are constructed by concatenating each training samples' HOG features and distances from the ground truth into two matrices, respectively. Note that both the feature matrix Φ_k and shape increment ΔX_k are normalized to zero mean before regression.

$$\beta_k = ((\Phi_k)^T \Phi_k + \lambda_k I)^{-1} (\Phi_k)^T \Delta X_k \quad (5)$$

After learning the ridge regression estimator, β_k, and calculating the offset b_k for each iteration, we use Equation 6 to localize facial landmarks starting from the initial points which are defined by the average landmark positions of the training samples.

$$x_{k+1}^i = x_k^i + \phi_k^i \beta_k + b_k \quad (6)$$

In Equation 6 ϕ_k^i, x_{k+1}^i and x_k^i represent the i^{th} sample's HOG features of the k^{th} iteration and the same sample's facial landmarks' locations of the $k + 1^{th}$ and k^{th} iterations, respectively.

3. Experimental Results

To evaluate the proposed method, we experimented on the commonly used Bosphorus 3D Face Database [12]. The Bosphorus database contains 4666 face samples belonging to 105 users. Each sample's 2D color image, 3D point cloud and manually annotated 24 facial landmark positions are provided by the database. The Bosphorus database contains a variety of pose and facial expression variations as well as occluded faces, making it a challenging database.

In our experiments, we worked on samples with frontal poses which had no occluding objects covering the face. 22 of the 24 facial landmarks were selected to be localized since the other two are ear dimples and are not visible in frontal images. Selected landmarks are eye, mouth, nose and eyebrow corners, middle points of lips, eyebrows, nose, chin, and the nose saddles, all of which can bee seen in Figure 1.

We compared our method with the state of the art 3D facial landmark localization methods methods working on depth images. A summary of these methods are given in Table 1. To be able to compare our method with the most successful Sukno et al. [14] and Camgoz et al. [4], whom are both using statistical facial landmark localization methods, we used the same experimental setup as theirs and reported

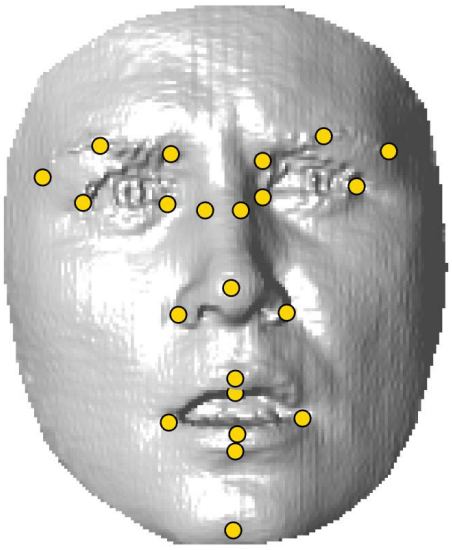

Figure 1: 22 landmarks used in our experiments

our results on 10 landmarks, which are common to all methods. We selected the frontal non-occluded face samples and divided them into two folds in which the users were exclusive to their groups. All the experiments have been done using two-fold cross validation and we iterated Supervised Ridge Descent (SRD) six times since it usually converges after the fourth iteration.

In our first experiments, our aim is to find the optimum λ_k values and HOG feature sizes. Our experiments yield optimum λ_k values to be $[300.00, 110.40, 40.60, 14.94, 5.49, 2.02, 0.74]$ and HOG feature sizes to be $[0.20, 0.17, 0.14, 0.12, 0.09, 0.06] \times ImageSize$ for the iterations from one to six, respectively.

The SRD method has two main novelties when compared to the SDM: 1) the use of ridge regression and 2) the use of adaptive feature sizes from coarse to fine resolution. In order to evaluate the independent contributions of these novelties, we performed several experiments by incrementally adding ridge regression and adaptive features to the classical SDM. As shown in Table 2, using ridge regression instead of least squares regression improves the performance drastically (See columns SDM and SRD with Fixed Feature Size). Similarly, using adaptive features instead of fixed features increases the performance for both SDM and SRD approaches (See SDM vs. SDM with Adaptive Feature Size columns and SRD vs. SRD with Fixed Feature Size columns). By incorporating both ridge regression and adaptive features, our SRD approach attains the best overall results (See column SRD).

As observed from Table 2, our best performing landmarks are eye and mouth corners, which have strong ge-

	#LM	# Training Size	# Test Size	Features	Method Used
Alyüz et al. [1]	5	–	2902	Shape Index	Heuristics
Creusot et al. [5]	14	99	2803	Surface Descriptors	LDA and Adaboost
Sukno et al. [14]	14	1402 x 2	1402 x 2	ASPC [13]	Statistical Shape Models
Camgoz et al. [4]	10	1446 x 2	1446 x 2	SIFT [11] - HOG [6]	SDM [16]
SRD (Our method)	22	1420 x 2	1420 x 2	HOG [6]	SDM [16] - Ridge Regression [10]

Table 1: Summary of the proposed method and the state of the art methods (LM = Landmarks)

ometric characteristics. However, our method struggled to localize chins and nose saddles which are difficult to locate accurately even by manual annotation. These findings were also backed up as we visualized the best and worst performing facial samples which can be seen in Figure 3. It can be seen from Figure 3 that ground truth locations of nose saddles differ for each subject which is probably due to the subjective preferences of the manual annotators.

To see if these results are consistent with all the samples we created the cumulative error distribution, which can be seen in Figure 2. By analyzing the curves of chin and nose saddles, we can confirm that both of these landmarks are problematic landmarks and their error is distributed over the whole database. This may be caused by false annotation of the data, as previously mentioned these landmarks are more ambiguous than the others.

To compare our method with the the state of the art methods, we used a subset of 10 points that most methods reported results on. As it can be seen in Table 3 the proposed method achieves the state of the art performance on all landmarks except the nose tip. Considering the manual annotation error for the nose tip (2.96mm, see Table 3), our average automatic localization error (2.65mm) can still be considered as not too high.

4. Conclusion

Many applications rely on the analysis of facial data to analyze, recognize and understand humans and their behaviors. Many of these applications start with facial landmark localization to be able to either align faces or track these landmarks. Thus a successful facial landmark localization is essential to the success of various facial processing tasks.

In this study, we use ridge regression to train Supervised Descent Method instead of the least squares method. We also use decreasing feature sizes in each iteration, which become smaller as the system iterates, turning the localization into a coarse to fine approach. Our experiments show that both improvements increase the performance significantly.

SRD was trained using HOG features in a similar manner to SDM. We experimented on the Bosphorus 3D Face Database and compared our method with the state of the art methods, which work on 10 common facial landmarks of the Bosphorus database, namely, eye corners, nose tip, nose corners, mouth corners and chin. Except for the nose tip, our approach achieved the best performance on all landmarks. However, our nose tip error is close to human annotation done by [1], which may indicate that the annotation variance may be the reason of this behaviour.

To improve our system, we plan to use 3D descriptors instead of 2D descriptors. To generalize our system, cross database experiments should also be conducted. Furthermore, feature learning methods can be used to learn features instead of using descriptors such as HOG, or structured iteration strategies may be implemented.

Acknowledgments

This research was supported by SANTEZ 0341.STZ.2013-2 project and ICT COST Action IC1106.

References

[1] N. Alyuz, B. Gokberk, and L. Akarun. Regional registration for expression resistant 3-d face recognition. *Information Forensics and Security, IEEE Transactions on*, 5(3):425–440, 2010.

[2] N. Alyuz, B. Gokberk, L. Spreeuwers, R. Veldhuis, and L. Akarun. Robust 3d face recognition in the presence of realistic occlusions. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, pages 111–118, March 2012.

[3] İ. Ari, A. Uyar, and L. Akarun. Facial feature tracking and expression recognition for sign language. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*, pages 1–6. IEEE, 2008.

[4] N. Camgoz, B. Gokberk, and L. Akarun. Facial landmark localization in depth images using supervised descent method. In *23th IEEE Signal Processing and Communications Applications Conference (SIU), 2015*, pages 1997–2000, May 2015.

[5] C. Creusot, N. Pears, and J. Austin. A machine-learning approach to keypoint detection and landmarking on 3d meshes. *International Journal of Computer Vision*, 102(1-3):146–179, 2013.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[7] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn. Intraface. In *Automatic Face and Gesture Recog-*

Landmarks	SDM	SDM with AFS	SRD with FFS	**SRD (Our Method)**
Outer left eyebrow	5.01 ± 2.97	4.16 ± 2.41	4.39 ± 2.58	$\mathbf{4.13 \pm 2.36}$
Middle left eyebrow	5.17 ± 3.07	4.69 ± 2.67	4.68 ± 2.81	$\mathbf{4.37 \pm 2.56}$
Inner left eyebrow	4.02 ± 2.45	3.52 ± 1.92	3.48 ± 2.08	$\mathbf{3.13 \pm 1.74}$
Inner right eyebrow	3.86 ± 2.23	3.28 ± 1.75	3.34 ± 1.97	$\mathbf{2.99 \pm 1.66}$
Middle right eyebrow	4.68 ± 2.86	4.19 ± 2.39	4.11 ± 2.49	$\mathbf{3.88 \pm 2.25}$
Outer right eyebrow	5.02 ± 4.10	4.19 ± 3.43	4.23 ± 3.53	$\mathbf{4.02 \pm 3.33}$
Outer left eye corner	3.16 ± 2.00	2.81 ± 1.57	2.63 ± 1.68	$\mathbf{2.56 \pm 1.45}$
Inner left eye corner	2.28 ± 1.55	2.12 ± 1.23	1.93 ± 1.39	$\mathbf{1.90 \pm 1.14}$
Inner right eye corner	2.10 ± 1.46	2.03 ± 1.21	1.84 ± 1.34	$\mathbf{1.84 \pm 1.15}$
Outer right eye corner	3.04 ± 2.00	2.89 ± 1.81	2.57 ± 1.84	$\mathbf{2.51 \pm 1.63}$
Nose saddle left	7.61 ± 3.96	7.08 ± 3.77	7.16 ± 3.73	$\mathbf{6.78 \pm 3.59}$
Nose saddle right	7.77 ± 4.03	7.29 ± 3.81	7.32 ± 3.82	$\mathbf{6.92 \pm 3.66}$
Left nose peak	2.51 ± 1.99	2.21 ± 1.31	2.18 ± 1.81	$\mathbf{1.96 \pm 1.20}$
Nose tip	3.34 ± 2.41	2.96 ± 1.90	3.01 ± 2.27	$\mathbf{2.65 \pm 1.76}$
Right nose peak	2.56 ± 2.04	2.18 ± 1.23	2.18 ± 1.96	$\mathbf{1.99 \pm 1.26}$
Left mouth corner	4.37 ± 3.82	3.09 ± 1.97	3.41 ± 3.39	$\mathbf{2.92 \pm 2.13}$
Upper lip outer middle	3.66 ± 3.52	2.71 ± 1.95	2.99 ± 3.25	$\mathbf{2.46 \pm 2.04}$
Right mouth corner	4.50 ± 3.85	3.05 ± 1.92	3.54 ± 3.33	$\mathbf{2.91 \pm 2.07}$
Upper lip inner middle	3.62 ± 3.47	2.64 ± 1.90	2.84 ± 3.25	$\mathbf{2.39 \pm 1.96}$
Lower lip inner middle	4.65 ± 5.01	2.60 ± 2.09	3.56 ± 4.44	$\mathbf{2.39 \pm 2.28}$
Lower lip outer middle	5.49 ± 5.59	3.14 ± 2.35	4.30 ± 5.07	$\mathbf{2.90 \pm 2.65}$
Chin middle	6.45 ± 5.60	5.32 ± 3.60	5.65 ± 4.87	$\mathbf{5.08 \pm 3.45}$
Mean Error	4.31 ± 3.18	3.55 ± 2.19	3.70 ± 2.86	$\mathbf{3.30 \pm 2.15}$

Table 2: Landmarks' mean and standard deviation of errors. SDM = Supervised Descent Method, SRD = Supervised Ridge Descent, FFS = Fixed Feature Size, AFS = Adaptive Feature Size

	Inner Eye Corners	Outer Eye Corners	Nose Tip	Nose Corners	Mouth Corners	Chin
Manual Annotation [1]	2.51	–	2.96	1.75	–	–
Alyüz et al.[1]	3.70	–	3.05	3.10	–	–
Creusot et al. [5]	4.14 ± 2.63	6.27 ± 3.98	4.33 ± 2.62	4.16 ± 2.35	7.95 ± 5.44	15.38 ± 10.49
Sukno et al. [14]	2.85 ± 2.02	5.06 ± 3.67	$\mathbf{2.33 \pm 1.78}$	3.02 ± 1.91	6.08 ± 5.13	7.58 ± 6.72
Camgoz et al. [4] (SIFT)	2.26 ± 1.79	4.23 ± 2.94	2.72 ± 2.19	4.57 ± 3.62	3.14 ± 2.71	5.72 ± 4.31
Camgoz et al. [4] (HOG)	2.33 ± 1.92	4.11 ± 3.01	2.69 ± 2.20	4.49 ± 3.62	3.16 ± 2.70	5.87 ± 4.19
SRD (Our method)	$\mathbf{1.87 \pm 1.14}$	$\mathbf{2.54 \pm 1.54}$	2.65 ± 1.76	$\mathbf{1.97 \pm 1.23}$	$\mathbf{2.92 \pm 2.10}$	$\mathbf{5.08 \pm 3.45}$

Table 3: Mean and standard deviation of 10 common facial landmark localization errors on Bosphorus 3D face database

nition (FG), 2015 11th IEEE International Conference and Workshops on, pages 1–8. IEEE, 2015.

[8] H. Dibeklioglu, F. Alnajar, A. Ali Salah, and T. Gevers. Combining facial dynamics with appearance for age estimation. *Image Processing, IEEE Transactions on*, 24(6):1928–1943, June 2015.

[9] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013.

[10] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, volume 5372 of *Lecture Notes in Computer Science*, pages 47–56. Springer Berlin Heidelberg, 2008.

[13] F. M. Sukno, J. L. Waddington, and P. F. Whelan. Rotationally invariant 3d shape contexts using asymmetry patterns. In *GRAPP-International confernece on computer graphics theory and applications*, 2013.

[14] F. M. Sukno, J. L. Waddington, and P. F. Whelan. 3-d facial landmark localization with asymmetry patterns and shape re-

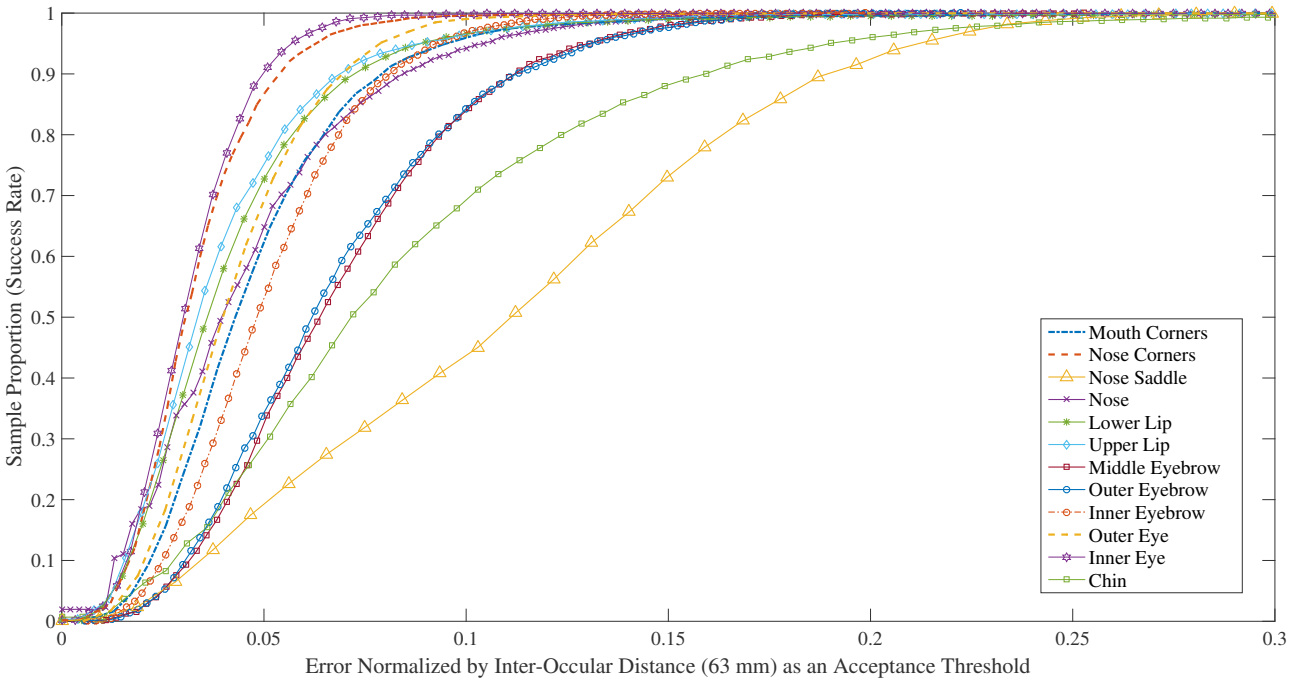

Figure 2: Cumulative error distribution of different landmarks

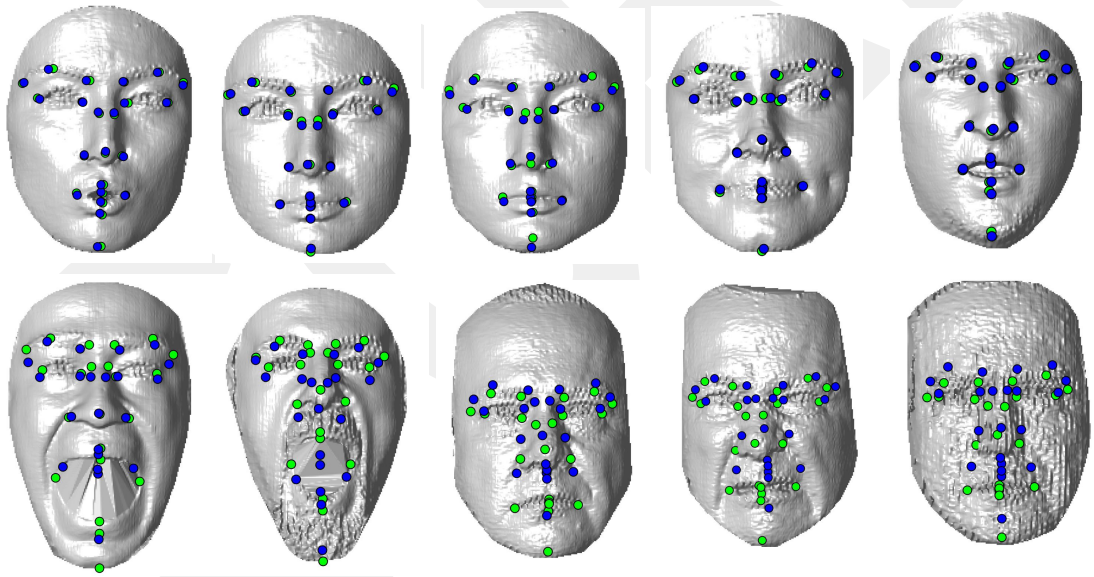

Figure 3: The first row shows the faces with the best landmark localization performance, while the second row shows samples with the worst performance. Green (Light) Dots = Ground Truth, Blue (Dark) Dots = Prediction (Best seen in color)

gression from incomplete local features. In *IEEE Transactions on Cybernetics*, 2014.

[15] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn. Fera 2015-second facial expression recognition and analysis challenge. *Proc. IEEE ICFG*, 2015.

[16] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[17] X. Xiong and F. De la Torre. Supervised descent method for solving nonlinear least squares problems in computer vision. *arXiv preprint arXiv:1405.0601*, 2014.