

3D FACIAL GEOMETRIC FEATURES FOR CONSTRAINED LOCAL MODEL

Shiyang Cheng* Stefanos Zafeiriou* Akshay Asthana* Maja Pantic*[†]

* Imperial College London, Department of Computing, United Kingdom

[†] University of Twente, EEMCS, Netherlands

{shiyang.cheng11, s.zafeiriou, a.asthana, m.pantic}@imperial.ac.uk

ABSTRACT

We propose a 3D Constrained Local Model framework for deformable face alignment in depth image. Our framework exploits the intrinsic 3D geometric information in depth data by utilizing robust histogram-based 3D geometric features that are based on normal vectors. In addition, we demonstrate the fusion of intensity data and 3D features that further improves the facial landmark localization accuracy. The experiments are conducted on publicly available FRGC database. The results show that our 3D features based CLM completely outperforms the raw depth features based CLM in term of fitting accuracy and robustness, and the fusion of intensity and 3D depth feature further improves the performance. Another benefit is that the proposed 3D features in our framework do not require any pre-processing procedure on the data.

Index Terms— Constrained local model, deformable face alignment, 3D facial geometry, histogram-based 3D feature.

1. INTRODUCTION

Automatic deformable face alignment is a critical step in multiple tasks of computer vision, for example face recognition, facial performance transfer and facial expression recognition. In general, it refers to the problem of localizing the facial landmarks following a model-based approach [1, 2, 3, 4]. An accurate alignment of the facial features not only provides position and structure information for successive tasks, but also improves the appearance-based recognition algorithms that rely heavily on accurate registration.

A variety of deformable face models have been proposed in the past decade, lifting the performance of face alignment to a higher level. Constrained Local Model (CLM) [3] is one of the state-of-the-art methods, in that it models the local variations around each landmark point by training a local detector (known as *patch expert*). Regularized Landmark Mean-Shift (RLMS), proposed by Saragih et. al. [4], is considered to be the state-of-the-art CLM fitting method. It employs a non-parametric estimation of posterior probability and the resulting optimization is reminiscent of mean-shift algorithm.

Although varying degrees of success have been achieved by the CLM, a further improvement is hampered by the ex-

istence of extreme lightening conditions, which raises questions on the robustness of patch experts. Since the current methodologies mainly focus on color/grayscale image alignment, which possesses large variations in illumination, comparatively less attention is drawn on the use of intrinsic 3D geometric information in depth image. As the commercial 3D scanners (from Di3D¹ dynamic face capturing system, Konica Minolta 3D scanning devices to Microsoft Kinect) become more accessible, acquiring accurate and reliable 3D depth image is much easier. Therefore, the use of 3D data and potential fusion with intensity image for face alignment demand a more in-depth study.

In this paper, we present a robust CLM framework that utilizes the histogram-based 3D geometric feature extracted from the depth data. Our framework takes the advantage of robust response map generated by the histogram-based 3D features (e.g., Histogram of Oriented Normal Vectors [5] and Local Normal Binary Patterns (LNBP)s [6]) to perform generic face alignment in depth image. We conduct facial landmark localization experiments on Face Recognition Grand Challenge (FRGC) [7] without any pre-processing step, and show that by using 3D feature instead of the raw depth map, the fitting performance improves drastically. In addition, we find that a simple fusion of intensity data and 3D features further improves the performance by a considerable margin.

2. RELATED WORK

Model-based generic face alignment generally refers to the alignment of generic facial landmarks (e.g. eye corners, nose tip) either via a Holistic or Part-based Deformable Model. Holistic Model aims at modelling the holistic facial texture, while Part-based Model utilizes the local patch information of every landmark. Typical examples for the Holistic Model are Active Appearance Models (AAM) [2] and 3D Morphable Model (3DMM) [8]. The later category includes Active Shape Model (ASM) [1], Constrained Local Model (CLM) [9, 3, 4, 10] and Tree-based pictorial structure [11].

¹<http://www.di3d.com>

Although the current CLM framework has well been established and shown to achieve good results on color/grayscale data, the focus is seldom drawn on exploring the 3D facial geometry feature on depth data for the CLM. To the best of our knowledge, CLM-Z framework, proposed by Baltrušaitis et. al. [12], is the only CLM framework tailored for depth image to perform non-rigid face tracking as well as rigid head pose tracking. They show that the integration of depth and intensity response maps can alleviate the aperture problem. However, CLM-Z does not fully exploit the potential geometric information in the depth data, since the CLM-Z uses the depth image in almost the same manner (with minor difference in patch normalization) as the intensity image. Moreover, in terms of facial landmark fitting accuracy and robustness, the reported results of CLM-Z on depth data are far worse than that of their intensity counterparts.

In [13], Oreifej et al. verify that direct adoption of conventional color-based methods for depth data causes two major issues which degrade the detection performance. The first issue comes from the spatially and temporally discontinuous black regions [14], which we refer as the *missing values*. And the more important issue is that the depth image contains rich surface information, and it can be converted to a powerful image descriptor capturing geometric variations of the object. Hence, we believe that by using histogram-based features which encode the information from normal vectors, the overall performance of the depth CLM is very likely to increase. Therefore, the novelty of our work comes from the utilization of robust 3D geometric features for CLM fitting.

3. ROBUST 3D CONSTRAINED LOCAL MODEL

In this section, we firstly introduce the commonly used 3D geometric features for the recognition/detection tasks. Then, we propose our 3D CLM framework with robust histogram-based 3D features.

3.1. 3D Geometric Feature

The depth sensor provides the distance $d(x, y)$ between pixel in position (x, y) and the sensor, hence each pixel in depth image can be defined as $\mathbf{d} = [x \ y \ d(x, y)]^T$. Assume the normal vector of \mathbf{d} is $\mathbf{n} = [n_x \ n_y \ n_z]^T$, where n_x and n_y are the first-order derivatives of $d(x, y)$ with respect to x and y direction. Generally, the spherical coordinates (Eq. 1) encode the orientation information better than Cartesian coordinates, as zenith angle θ and azimuth angle φ are natural descriptors to the surface of the object. In addition, the normal vector always can be mapped to a plane (i.e. $\mathbf{Z} = 1$), hence n_z is a constant [5], it is redundant to consider radius r in practice.

$$\begin{pmatrix} \varphi \\ \theta \\ r \end{pmatrix} = \begin{pmatrix} \arctan(n_y/n_x) \\ \arctan(n_z/\sqrt{n_x^2 + n_y^2}) \\ \sqrt{n_x^2 + n_y^2 + n_z^2} \end{pmatrix}. \quad (1)$$

Histogram of Oriented Normal Vectors (HONV), proposed by Tang et al. [5], follows the paradigm of the very well engineered Histogram of Oriented Gradients (HOG) [15] features. It divides the search window into m number of *cells*, within which the orientation of normal vector (φ and θ) at each pixel is voted into a 2D histogram. The orientation value is softly binned into the neighbourhood bins using bilinear interpolation so as to avoid boundary effects. Moreover, a 2D Gaussian smoothing is applied to adjacent cells (also called *block*) to minimize spatial boundary effects. The final feature vector would simply be the concatenation of all the normalized histograms. In this paper, we empirically assign 9 bins for φ and 4 bins for θ , 4 pixels as the *cell* size and 3 as the *block* size.

Local Normal Binary Patterns (LNBP)s proposed in [6] is an extension of Local Binary Patterns (LBPs) [16]. Two descriptors $LNBP_{OA}$ and $LNBP_{TA}$ are proposed in [6], both of them use the angle differences between normal vectors to encode the neighbourhood around the center point \mathbf{x}_c . Given a circular neighbourhood of radius r , which covers P points, we compute the unit normal vectors of \mathbf{x}_c and its neighbours, denoted as \mathbf{n}_c and \mathbf{n}_p respectively.

$LNBP_{OA}$ (Eq. 2) computes the inner product of \mathbf{n}_c and \mathbf{n}_p , then compares it with the inner product between \mathbf{n}_c and the vector \mathbf{n}_t that has a certain angle ψ from \mathbf{n}_c . A P -bit binary number is generated for every pixel, and the histogram built from all the numbers becomes a feature vector. Different from $LNBP_{OA}$, $LNBP_{TA}$ (Eq. 4 & 5) considers angle differences of both φ and θ , with ψ_a and ψ_z being the threshold angles respectively. Hence, for each pixel, two binary numbers are generated, from which a 2D histogram is computed. The flattened histogram is used as the feature for further tasks.

Since $LNBP_{OA}$ is reported to achieve a slightly better average detection score than $LNBP_{TA}$ in [6], and the dimensionality of $LNBP_{OA}$ is 15 times less than that of $LNBP_{TA}$, we choose $LNBP_{OA}$ as our 3D feature descriptor in the proposed framework. In all our experiments, we set the radius r and P to 8, and use $\frac{\pi}{12}$ for threshold angle ψ which produces the best result.

$$LNBP_{OA}(\mathbf{x}_c) = \sum_{p=0}^{P-1} 2^p d(\mathbf{n}_c \cdot \mathbf{n}_p, \mathbf{n}_c \cdot \mathbf{n}_t), \quad (2)$$

where $\mathbf{n}_c \cdot \mathbf{n}_t = \cos(\psi)$ and function d is defined as

$$d(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 < x_2 \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

$$LNBP_{TA}^a(\mathbf{x}_c) = \sum_{p=0}^{P-1} 2^p d(\cos(|\varphi_c - \varphi_p|), \cos(\psi_a)), \quad (4)$$

$$LNBP_{TA}^z(\mathbf{x}_c) = \sum_{p=0}^{P-1} 2^p d(\cos(|\theta_c - \theta_p|), \cos(\psi_z)). \quad (5)$$

3.2. CLM on Histogram-based 3D Features

In this section, we explain our CLM framework based on the histogram-based 3D features discussed in Section 3.1. Our main contribution lies in the exploration of histogram-based 3D features from the depth image for CLM framework, as well as the fusion between intensity and depth information. Our CLM framework is defined by the model $M = \{\mathcal{S}, \mathcal{P}\}$. \mathcal{S} models the shape information, it can be either 2D or 3D Point Distribution Model (PDM) [1]. In this paper, we mainly use 3D PDM due to its robustness and compactness [17].

3.2.1. Patch Experts

Model \mathcal{P} includes the patch experts discriminatively trained for each facial landmark. It can be described as a group of linear classifiers $\mathcal{P} = \{\mathbf{w}_i, b_i\}_{i=1}^n$, where \mathbf{w}_i and b_i are the weights and bias of the i^{th} patch expert. The probability for the i^{th} landmark being correctly aligned ($a_i = 1$) at location \mathbf{x} of depth image \mathcal{D} is modelled by a logistic function [9, 4]:

$$p(a_i = 1 | \mathbf{x}, \mathcal{D}) = \frac{1}{1 + e^{-\{\beta_0 + \beta_1(\mathbf{w}_i^T \mathbf{h}(\mathbf{x}; \mathcal{D}) + b_i)\}}}, \quad (6)$$

where β_0 and β_1 are the regression intercept and coefficient respectively. $\mathbf{h}(\mathbf{x}; \mathcal{D})$ denotes the features computed from the local area centered around location \mathbf{x} in \mathcal{D} . In this paper, $\mathbf{h}(\mathbf{x}; \mathcal{D})$ refers to either raw local patch or the 3D features. For each patch expert, we compute features on the positive and negative samples selected from training data, and use Linear Support Vector Machine (SVM) [18] to train the classifier due to its computational advantage. Cross-validation is performed to select the best parameters for SVM.

Fusion of Patch Experts: For the intensity image \mathcal{I} corresponding to the depth image \mathcal{D} , the same training scheme can be applied to train the intensity-based patch experts, with the alignment probability denoted as $p(a_i = 1 | \mathbf{x}, \mathcal{I})$. Furthermore, through a element-wise addition of the response maps (Eq. 6) obtained from the respective intensity and depth patch experts, the fusion response map is computed as:

$$p(a_i | \mathbf{x}, \mathcal{I}, \mathcal{D}) = \frac{p(a_i | \mathbf{x}, \mathcal{I}) + p(a_i | \mathbf{x}, \mathcal{D})}{2}. \quad (7)$$

We would show that the fusion of raw intensity image and robust 3D geometric feature increases the fitting accuracy and robustness by considerable margin, while the the fusion of raw intensity and depth image degrades the performance.

3.2.2. CLM Fitting

Once we obtain the probability of alignment for each landmark point, the objective function of CLM fitting can be formulated as:

$$\begin{aligned} \mathbf{p}_0 &= \arg \max_{\mathbf{p}} \{p(\mathbf{s} | \{a_i = 1\}_{i=1}^n, \mathcal{D})\} \\ &= \arg \max_{\mathbf{p}} \{p(\mathbf{p}) p(\{a_i = 1\}_{i=1}^n | \mathbf{s}, \mathcal{D})\}, \end{aligned} \quad (8)$$

where $\mathbf{s} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$, and \mathbf{x}_i is the i^{th} landmark of the shape. \mathbf{p} is the set of 3D PDM parameter which contains the deformation information of shape \mathbf{s} .

To simplify this problem, it is assumed that the detection of every landmark point is conditionally independent to each other. Therefore we can interpret (8) as:

$$\mathbf{p}_0 = \arg \max_{\mathbf{p}} \{p(\mathbf{p}) \prod_{i=1}^n p(a_i = 1 | \mathbf{x}_i(\mathbf{p}), \mathcal{D})\}. \quad (9)$$

Regularized Landmark Mean-Shift (RLMS): RLMS [4] assumes a homoscedastic isotropic Gaussian kernel density estimate in a set of fixed locations $\{\Psi_i\}_{i=1}^n$ for every part i :

$$p(a_i = 1 | \mathbf{x}_i(\mathbf{p}), \mathcal{D}) = \sum_{\mathbf{y}_i \in \Psi_i} p(a_i = 1 | \mathbf{y}_i, \mathcal{D}) \cdot \mathcal{N}(\mathbf{x}_i(\mathbf{p}) | \mathbf{y}_i, \rho \mathbf{I}), \quad (10)$$

the objective function (9) can be reinterpreted as:

$$\begin{aligned} \mathbf{p}_0 &= \arg \max_{\mathbf{p}} \left\{ p(\mathbf{p}) \prod_{i=1}^n \sum_{\mathbf{y}_i \in \Psi_i} \right. \\ &\quad \left. p(a_i = 1 | \mathbf{y}_i, \mathcal{D}) \cdot \mathcal{N}(\mathbf{x}_i(\mathbf{p}) | \mathbf{y}_i, \rho \mathbf{I}) \right\}. \end{aligned} \quad (11)$$

In 3D PDM, non-rigid shape parameters \mathbf{q} are generally assumed to follow Gaussian distribution, which leads to the regularization term $p(\mathbf{p})$:

$$p(\mathbf{p}) \propto \mathcal{N}(\mathbf{q}; \mathbf{0}, \mathbf{L}), \quad (12)$$

where \mathbf{L} are the eigenvalues obtained from PCA, and ρ is the mean value of \mathbf{L} .

Equation (11) is solved by the Expectation-Maximization (EM) algorithm. The E-step involves computing the *posterior* distribution $\omega_{\mathbf{y}_i} = p(\mathbf{y}_i | a_i = 1, \mathbf{x}_i, \mathcal{D})$ over latent variable $\{\mathbf{y}_i\}_{i=1}^n$. The maximization step minimizes:

$$Q(\mathbf{p}) = \|\mathbf{q}\|_{\mathbf{L}}^{-1} + \sum_{i=1}^n \sum_{\mathbf{y}_i \in \Psi_i} \frac{\omega_{\mathbf{y}_i}}{\rho} \|\mathbf{x}_i(\mathbf{p}) - \mathbf{y}_i\|^2,$$

which can be solved using a Gauss-Newton optimization.

4. EXPERIMENTAL RESULTS

To evaluate the fitting performance of our 3D CLM framework, we conduct the landmark localization experiments on Face Recognition Grand Challenge (FRGC) [7]. FRGC database [7] is initially designed for the task of improving face recognition algorithm, it consists of 50,000 recordings of high quality still images as well as the corresponding 3D scans. For our experiment, we annotate 4910 images (with 66 facial landmarks) from 520 different subjects [19, 20]. The annotations are then split into training and testing sets, with 400 subjects (3993 images) selected for training, while the rest 120 subjects (917 images) are used for testing.

We initialize the fitting procedure with the mean shape centered at the bounding box perturbed by ± 10 pixels in translation and 10% in scale, w.r.t to the ground-truth bounding box. This replicates the initialization error that will normally be obtained by using any face detector. We employ Shape Root-Mean Square (RMS) error normalized w.r.t. the inter-ocular distance of the face to measure fitting performance, which is unbiased for different size of faces.

4.1. Individual Feature Results

The fitting results of using individual feature (i.e., Depth, Intensity, LNBP and HONV) in Figure 1 show that histogram-based 3D features LNBP and HONV completely outperform the raw depth feature.

As for the improvement achieved by the proposed methods over the raw depth feature, considering the normalized error (i.e. Shape RMS error as the fraction of inter-ocular distance) of 0.03 as the benchmark for accurate landmark localization, LNBP and HONV show astonishing improvements of 50% and 65% over the raw depth descriptor respectively. This proves that by exploiting the potential 3D geometric information and using it in CLM framework, the negative effect of *missing values* in depth data can be minimized, which leads to a much better performance. On the other hand, the use of histogram feature based CLM does not require any additional pre-processing procedure on the depth data. It is interesting to see that HONV shows an improvement of 7% over the intensity, which indicates the richness of information in 3D geometry data. Sample fitting results of HONV CLM can be found in Figure 2, we plot the fitting points in the color image instead of depth image for the ease of visualization.

4.2. Feature Fusion Results

To test the performance of fusion proposed in Section 3.2.1, we fuse the intensity feature based patch expert with those of the raw depth, LNBP and HONV. In Figure 1, we observe that the fusion of intensity and raw depth data lowers the CLM fitting accuracy by 6% comparing with the raw intensity. Meanwhile, the fusions of intensity data and histogram-based 3D features improve the corresponding individual features, among which the fusion with HONV generates the best result (sample fittings are displayed in Figure 2). Considering the normalized error of 0.03 as the benchmark of accurate fitting, the fusion of intensity and HONV achieves a significant improvement of nearly 10% over the single HONV descriptor and 15% over the intensity. This result shows the advantage of combining 3D depth feature and intensity information, and it provides us the motivation on investigating more advanced feature fusion techniques in the future work.

5. CONCLUSIONS

We present a robust 3D Constrained Local Model framework based on the histogram-based 3D facial geometry features

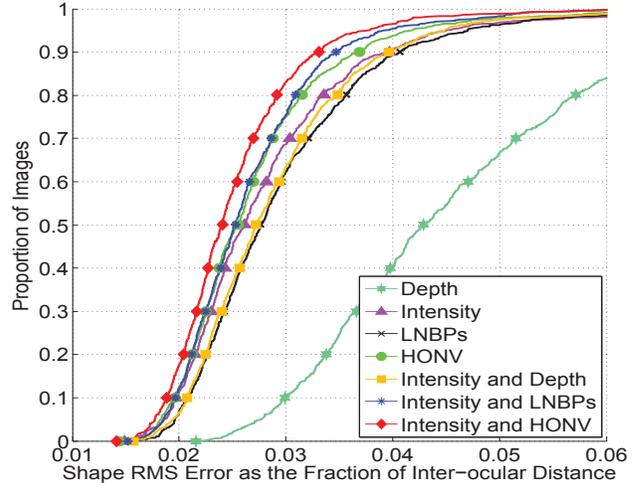


Fig. 1. Experimental results on FRGC database.

(i.e. HONV [5] and LNBP [6]). We conduct facial landmark localization experiments on Face Recognition Grand Challenge (FRGC) [7] database to test the proposed 3D CLM framework. The result shows that by using histogram-based 3D features instead of the raw depth feature, the performance of CLM improves drastically. In addition, we find that a simple fusion of intensity data and 3D feature further improves the performance by a significant margin.

Acknowledgement: The work of S. Cheng and S. Zafeiriou is funded by the EPSRC project EP/J017787/1 (4D-FAB). The work of A. Asthana is funded by Marie Curie Fellowship under FP7-PEOPLE-2011-IIF Grant agreement no. 302836 (FER in the Wild).

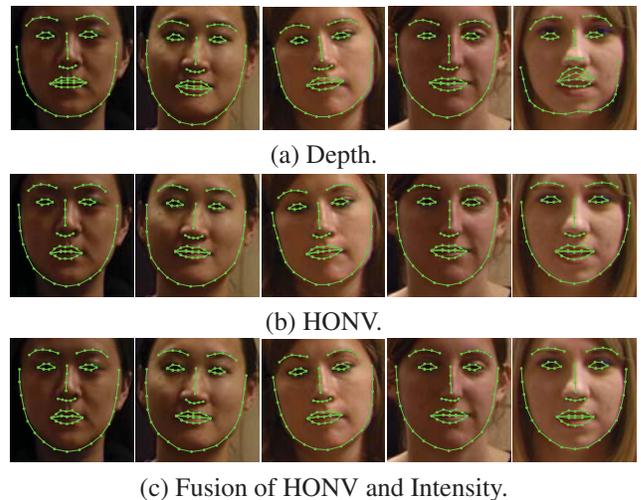


Fig. 2. Examples of CLM fittings on FRGC database. Top row is raw depth, middle row is HONV and bottom row is fusion of Intensity and HONV.

6. REFERENCES

- [1] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and applications," *CVIU*, 1995.
- [2] G. Edwards, C. Taylor, and T. Cootes, "Interpreting Face Images Using Active Appearance Models," in *IEEE FG*, 1998.
- [3] David Cristinacce and Tim Cootes, "Feature detection and tracking with constrained local models," in *BMVC*, 2006.
- [4] Jason Saragih, Simon Lucey, and Jeffrey Cohn, "Deformable model fitting by regularized landmark mean-shift," *IJCV*, vol. 91, no. 2, pp. 200–215, Jan. 2011.
- [5] Shuai Tang, Xiaoyu Wang, Xutao Lv, Tony X. Han, James Keller, Zhihai He, Marjorie Skubic, and Shihong Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part II*, Berlin, Heidelberg, 2013, ACCV'12, pp. 525–538, Springer-Verlag.
- [6] G. Sandbach, S. Zafeiriou, and M. Pantic, "Local normal binary patterns for 3d facial action unit detection," in *Proceedings of the IEEE International Conference on Image Processing (ICIP 2012)*, Orlando, FL, USA, October 2012, pp. 1813–1816.
- [7] P. Jonathon Phillips, Patrick J. Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek, "Overview of the face recognition grand challenge," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, Washington, DC, USA, 2005, CVPR '05, pp. 947–954, IEEE Computer Society.
- [8] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE PAMI*, vol. 25, no. 9, pp. 1063–1074, Sept. 2003.
- [9] A. Athana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *IEEE CVPR*, 2013.
- [10] Xuehan Xiong and Fernando De la Torre Frade, "Supervised descent method and its applications to face alignment," in *IEEE CVPR*, 2013.
- [11] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," in *CVPR*, 2012.
- [12] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *IEEE CVPR*, 2012.
- [13] Omar Oreifej and Zicheng Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," June 2013.
- [14] M. Camplani and L. Salgado, "Efficient spatio-temporal hole filling strategy for Kinect depth maps," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Feb. 2012, vol. 8290 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*.
- [15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
- [16] Timo Ojala, Matti Pietikinen, and David Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, pp. 51–59, 1996.
- [17] I. Matthews, J. Xiao, and S. Baker, "2D vs. 3D Deformable Face Models: Representational Power, Construction, and Real-Time Fitting," *IJCV*, vol. 75, no. 1, pp. 93–113, Oct. 2007.
- [18] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *JMLR*, vol. 9, pp. 1871–1874, 2008.
- [19] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of IEEE Intl Conf. on Computer Vision (ICCV-W 2013), 300 Faces in-the-Wild Challenge (300-W)*, Sydney, Australia, December 2013.
- [20] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proceedings of IEEE Intl Conf. Computer Vision and Pattern Recognition (CVPR-W13), 5th Workshop on Analysis and Modeling of Faces and Gestures (AMFG2013)*, Portland Oregon, USA, June 2013.