Holistically Constrained Local Model: Going Beyond Frontal Poses for Facial Landmark Detection

KangGeon Kim[1]
kanggeon.kim@usc.edu
Tadas Baltrušaitis[2]
tbaltrus@cs.cmu.edu
Amir Zadeh[2]
abagherz@cs.cmu.edu
Louis-Philippe Morency[2]
morency@cs.cmu.edu
Gérard Medioni[1]
medioni@usc.edu

[1] Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA, USA

[2] Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

Facial landmark detection has received much attention in recent years, with two detection paradigms emerging: local approaches, where each facial landmark is modeled individually and with the help of a shape model; and holistic approaches, where the face appearance and shape are modeled jointly. In recent years both of these approaches have shown great performance gains for facial landmark detection even under "in-the-wild" conditions of varying illumination, occlusion and image quality. However, their accuracy and robustness are very often reduced for profile faces where face alignment is more challenging (e.g., no more facial symmetry, less defined features and more variable background). In this paper, we present a new model, named Holistically Constrained Local Model (HCLM), which unifies local and holistic facial landmark detection by integrating head pose estimation, sparse-holistic landmark detection and dense-local landmark detection. We evaluate our new model on two publicly available datasets, 300-W and AFLW, as well as a newly introduced dataset, IJB-FL which includes a larger proportion of profile face poses. Our HCLM model shows state-of-the-art performance, especially with extreme head poses.

1 Introduction

Facial landmark detection is an essential initial step for a number of facial analysis research areas such as expression analysis, face 3D modeling, facial attribute analysis, and person recognition. It is a well researched problem that has seen a surge of interest in the past couple of years.

However, most state-of-the-art methods still struggle in the presence of extreme head pose, especially in challenging in-the-wild images. Furthermore, as most methods operate

Pages 95.1-95.12
DOI: https://dx.doi.org/10.5244/C.30.95

in a local manner [16, 24, 26], they rely on good and consistent initialization, which is often very difficult to achieve. While some images attempt to combat this by evaluating a number of proposals and initializations, this comes at a computational cost.

In our work, we present a new model, named Holistically Constrained Local Model (HCLM), which unifies local and holistic facial landmark detection by integrating head pose estimation, sparse-holistic landmark detection and dense-local landmark detection. Our method's main advantage is the ability to handle very large pose variations, including profile faces. Furthermore, our model integrates local and holistic facial landmark detectors in a joint framework, with a holistic approach narrowing down the search space for the local one.

We demonstrate the benefits of our model for facial landmark detection through extensive experiments on two publicly available datasets, 300-W [18] and AFLW [10], as well as a newly introduced dataset, IJB-FL (a subset of IJB-A [9]), which includes a larger proportion of profile face poses. Furthermore, we demonstrate the importance of each component of our model in a series of ablation studies, showing the importance of both the head pose estimation and sparse landmark detection.

In the following section, we provide a brief survey of the work done on facial landmark detection and head pose estimation. In Section 3 we describe our novel holistically constrained local model for landmark detection. We follow this with description of our experiments (Section 4) and results (Section 5) demonstrating the benefits of our model. Finally, we summarize our work in Section 6 and propose future directions.

2 Related Work

Facial landmark detection and head pose estimation have made huge progress in the past couple of years. A large number of new approaches and techniques have been proposed especially for landmark detection in faces from RGB images. A full review of work in facial landmark detection and head pose estimation is outside the scope of this paper, we refer the reader to some recent reviews of the field [6, 23].

Facial landmark detection

Modern facial landmark detection approaches can be split into two main categories - *local* and *holistic*. *Local* approaches often model both appearance and shape of facial landmarks with the latter providing a form of regularization. *Holistic* approaches on the other hand do not require an explicit shape model and landmark detection is directly performed on appearance. We provide a short overview of recent local and holistic methods.

Holistic Nowadays majority of the holistic approaches follow a cascaded regression framework, where facial landmark detection is updated in a cascaded fashion. That is the landmark detection is continually improved by applying a regressor on appearance given the current landmark estimate as performed by Cao et al. in explicit shape regression [5]. Other cascaded regression approaches include the Stochastic Descent Method (SDM) [24] which uses SIFT [12] features with linear regression to compute the shape update and Coarse-to-Fine Shape Searching (CFSS) [27] which attempts to avoid a local optima in cascade regression by performing a coarse to fine shape search.

Recent work has also used deep learning techniques in a cascaded regression framework to extract visual features. Coarse-to-Fine Auto-encoder Networks (CFAN) [26] use visual features extracted by an auto-encoder together with linear regression. Sun et al. [20] proposed a Convolutional Neural Network (CNN) based cascaded regression approach for

sparse landmark detection, however while their approach is robust it is not very accurate.

Local Local approaches are often made up of two steps: extracting an appearance descriptor around certain areas of interest and computing local response maps; fitting a shape model based on the local predictions. Such areas are often defined by the current estimate of facial landmarks. A popular local method for landmark detection is the Constrained Local Model [19] and its various extensions such as Constrained Local Neural Fields [2] and Discriminative Response Map Fitting [1] that use more advanced ways of computing local response maps and inferring the landmark locations. Project out Cascaded regression (PO-CR) [21] is another example of a local approach, but one that uses a cascaded regression to update the shape model parameters rather than predicting landmark locations directly.

Another noteworthy local approach is the mixture of trees model [28] that uses a tree based deformable parts model to jointly perform face detection, pose estimation and facial landmark detection. A notable extension to this approach is the Gauss-Newton Deformable Part Model (GN-DPM) [22] which jointly optimizes a part-based flexible appearance model along with a global shape using Gauss-Newton optimization.

Rajamanoharan and Cootes [15] proposed a local approach that explicitly aims to be more robust in presence of large pose variations. They use landmark detectors trained at orientation conditions to produce more discriminative response maps and explored the best spatial splits for this task. However, they do not propose how such pose information could be acquired to initialize the model, in our work we use similarly trained landmark detectors but also provide a way of initializing the models at extreme angles.

Head pose estimation

Head pose estimation has not received the same amount of interest as facial landmark detection in recent years. Most of the recent work concentrated on exploiting range sensors for the task [7, 14]. However, the limitation of such approaches is that they cannot work on purely RGB images. A large number of head pose estimation approaches rely on explicit prediction of head pose from the image head pose estimate [6]. This is often done through multivariate regression or multi-class classification [13]. It is also possible to use model based rather than discriminative approaches [6]. For example, using the estimated facial landmarks together with a 3D face model to estimate the head pose. However, this requires an estimate of camera calibration parameters and might not be suitable for some applications.

Most similar model to our work is that proposed by Yang et al. [25]. In their approach an estimate of a head pose from a Convolutional Neural Network (CNN) is used to initialize a cascaded shape regression approach for face alignment. Our work integrates the head pose in a similar manner, but also proposes the use of a combined holistic and local approaches for landmark detection.

Our proposed HCLM approach is a method that combines both holistic and local approaches in a joint framework. We use a holistic Convolutional Neural Network (CNN) for initial sparse landmark detection. The sparse landmarks are then used as anchor points for a local CLNF approach. Finally, we integrate a head pose estimation model that allows for even more accurate landmark detection, especially in the presence of non frontal faces.

3 Model

In this section, we introduce our Holistically Constrained Local Model (HCLM) for facial landmark detection. We first start by describing a joint framework for incorporating holistic

and local models (Section 3.1). This is done by refining the sparse landmark predictions (*holistic*) with a dense landmark model (*local*). We follow this by a description of sparse landmark detection and head pose estimation that constrain and refine our model in Section 3.2.

3.1 Holistically Constrained Local Model

Our model integrates coarse and fine landmark detection together in a unified framework. The main goal of this approach is to improve the fine grained dense landmark detection with the help of coarser sparse landmarks.

For a given set of k facial landmark positions $\mathbf{x} = \{x_1, x_2, ..., x_k\}$, our HCLM model defines the likelihood of the facial landmark positions conditioned on a set of sparse landmark positions $X_s = \{x_s,\ s \in S\}$ ($|S| \ll k$) and image $\mathcal{I}$ as follows:

$$p(\mathbf{x}|I, X_s, \mathcal{I}) \propto p(\mathbf{x}) \prod_{i=1}^{k} p(x_i|X_s, \mathcal{I}). \tag{1}$$

In Equation 1, $p(\mathbf{x})$ is prior distribution over set of landmarks $\mathbf{x}$ following a 3D point distribution model (PDM) with orthographic camera projection. Similarly to Saragih et al. [19], we impose a Gaussian prior on the non-rigid shape parameters on the model. The probability of individual landmark alignment (response map) is modeled using the following distribution:

$$p(x_i|X_s, \mathcal{I}) = \begin{cases} \mathcal{N}(x_i|\mu = X_s, \sigma^2) & i \in S \\ \\ C(x_i|\mathcal{I}) & i \notin S \end{cases} \tag{2}$$

Above, C is a probabilistic patch expert that describes the probability of a landmark being aligned, while $\mathcal{N}(x_i|\mu, \sigma^2)$ is a bivariate Normal distribution evaluated at x_i with mean - μ and variance - σ^2 in both dimensions. Equation 2 allows our model to place high confidence (controlled by small σ) on a set of sparse landmarks (detected by a holistic model from Section 3.2), while incorporating the response maps from a denser set. C can be any model producing a probabilistic predictions of landmark alignment. In our work, we define C as a multivariate Gaussian likelihood function of a Continuous Conditional Neural Field (CCNF) [3]:

$$C(\mathbf{y}|\mathcal{I}) = p(\mathbf{y}|\mathcal{I}) = \frac{exp(\Psi)}{\int_{-\infty}^{\infty} exp(\Psi)\, dy}, \tag{3}$$

$$\Psi = \sum_{y_i} \sum_{k=1}^{K1} \alpha_k f_k(y_i, \mathcal{W}(y_i; \mathcal{I}), \theta_k) + \sum_{y_i, y_j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j), \tag{4}$$

Above $\mathbf{y}$ is a $m \times m$ area of interest in an image around the current estimate of the landmark (the area we will be searching in for an updated location of the landmark), $\mathcal{W}(y_i; \mathcal{I})$ is a vectorised version of an $n \times n$ image patch centered around y_i and is called the support region (the area based on which we will make a decision about the landmark alignment, typically $m > n$), f_k is a logistic regressor, and g_k is smoothness encouraging edge potential [3]. It can be shown that Equation 3 is a Multivariate Gaussian function [3], making the exact inference possible and fast to compute. The model parameters $[\alpha, \beta, \theta]$ of the CCNF are learned by using Maximum Likelihood Estimation (using BFGS optimization algorithm).

Table 1: *The structure of our convolutional network used for sparse landmark detection and head pose estimation.*

Name	Type	Filter Size	Stride	Output Size
Conv1	convolution	20x4x4	1	20x36x36
Pool1	max-pooling	2x2	2	20x18x18
Conv2	convolution	40x3x3	1	40x16x16
Pool2	max-pooling	2x2	2	40x8x8
Conv3	convolution	60x3x3	1	60x6x6
Pool3	max-pooling	2x2	2	60x3x3
Conv4	convolution	80x2x2	1	80x2x2
Dense1	fully connected			120
Dense2	fully connected			10 (6)

(a) The CNN architecture for sparse landmark detection

Name	Type	Filter Size	Stride	Output Size
Conv1	convolution	32x3x3	1	32x94x94
Pool1	max-pooling	2x2	2	32x47x47
Conv2	convolution	64x2x2	1	64x46x46
Pool2	max-pooling	2x2	2	64x23x23
Conv3	convolution	128x2x2	1	128x22x22
Pool3	max-pooling	2x2	2	128x11x11
Dense1	fully connected			400
Dense2	fully connected			400
Dense3	fully connected			3

(b) The CNN architecture for head pose estimation

In order to optimize Equation 1, we use Non-Uniform Regularized Landmark Mean-Shift which iteratively computes the patch responses and updates the landmark estimates by updating the PDM parameters [2].

The following section describes the method of acquiring the holistic sparse landmarks, X_s, used to constrain our dense landmark detector. It also describes our head pose estimation model that allows for better initialization.

3.2 Two holistic predictors: sparse landmarks and head pose

Our HCLM model depends on a set of sparse landmarks from a holistic model. In our work we use a similar approach to the CNN model proposed by Sun et al. [20] for such landmark detection. Table 1 (a) shows the CNN architecture used in our work. A gray-scale image of size 39×39 is used for input and pixel values are normalized to the range between 0 and 1. For sparse landmarks, five landmarks (two eyes, one nose, two mouth corners) are used for frontal face and three landmarks (one eye, one nose, one mouth corner) are used for profile face. Consequently, the number of output is ten when the face is frontal or six when it is profile $\{x_1, y_1, x_2, y_2, ..., x_n, y_n\}$ (due to self occlusion). The location of landmarks is shifted with respect to the image center, x and y, and normalized by width and height to be in range between -0.5 and 0.5. We use the Euclidean distance as the network loss:

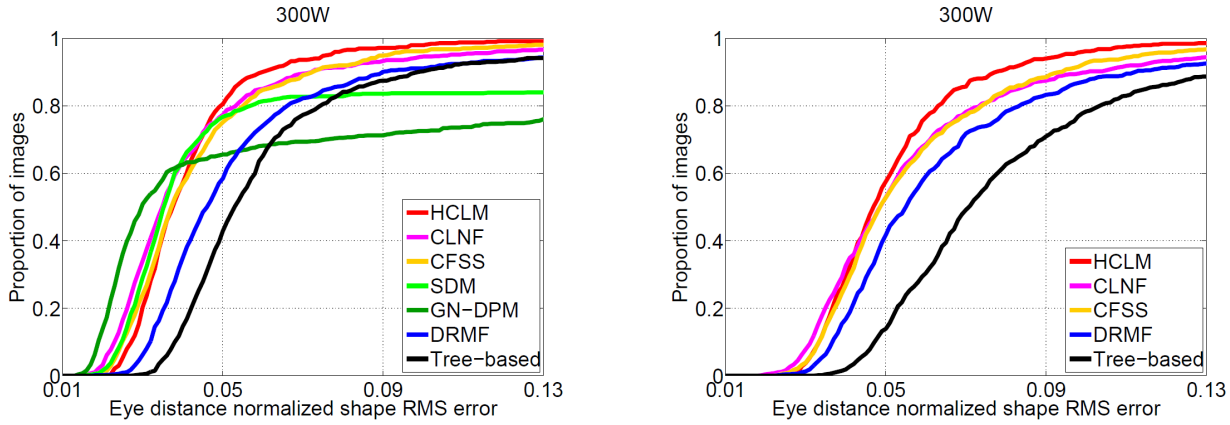
$$loss_{sparse} = \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|_2^2, \tag{5}$$

where $\hat{\mathbf{X}}_i$ is the given ground truth location, and $\mathbf{X}_i$ is the predicted location for training image $\mathcal{I}_i$. Note that $\hat{\mathbf{X}}_i$, and $\mathbf{X}_i$ are normalized locations.

To assist the face alignment and facial landmark detection, it is helpful to know the head pose in advance. Most cases of face alignment failures come from the large head pose variations as the initial shape is often frontal and local approaches are not able to converge onto correct non-frontal landmarks. To avoid this problem, we developed a head pose estimation module which gives an estimate of the three head pose angles: pitch, yaw and roll.

Our implementation of CNN for head pose estimation is based on the work of Yang et al. [25]. Table 1 (b) shows an overview of our CNN architecture. A gray-scale image of size (96×96) is used for input and pixel values are normalized to the range between 0 and 1. The output is three dimensional vector which represents pitch, yaw and roll. The output angles are normalized between -1 and 1. We use the Euclidean distance as the network loss:

$$loss_{head\,pose} = \|\hat{\mathbf{P}}_i - \mathbf{P}_i\|_2^2, \tag{6}$$



(a) 49 point cumulative errors (b) 68 point cumulative errors

Figure 1: *Cumulative error curves on 300-W.* Measured as the mean Euclidean distance from ground truth normalized by the inter-ocular distance.

where $\hat{\mathbf{P}}_i$ is the given ground truth angle, and $\mathbf{P}_i$ is the predicted head pose for training image I_i.

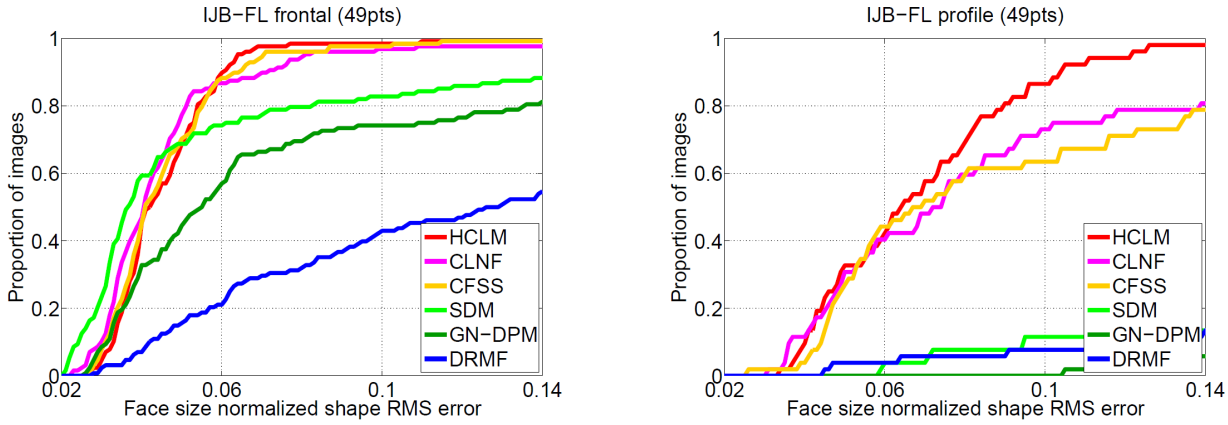
The full pipeline of our HCLM model is as follows: 1) use a CNN head pose predictor to estimate the head pose in the input image; 2) use a view dependent CNN sparse landmark detector; 3) use the HCLM model with the detected sparse landmarks. In the following section, we extensively evaluate our model showing its benefits over other models, especially in presence of extreme poses.

4 Experiments

We designed our experiments to study three aspects of our model. First, we compared at a general level our HCLM with a number of state-of-the-art baselines on two publicly available facial landmark detection datasets. Second, we performed a set of ablation experiments to see how each element of HCLM model affects the final facial landmark detection results. Finally, we demonstrated the performance of the individual holistic models for sparse landmark detection and head pose estimation. In the following sections, the datasets we used and the experimental procedures we followed are presented.

4.1 Comparison with baseline methods

We compared our model with a number of recently proposed approaches for facial landmark detection (both *holistic* and *local* ones). The following models acted as our baselines: **Tree-based** deformable part method of Zhu et al. [28], Gauss-Newton Deformable Part Model (**GN-DPM**) [22], Discriminative Response Map Fitting (**DRMF**) instance of a Constrained Local Model [1], Supervised Descent Method (**SDM**) model of cascaded regression [24], Coarse-to-fine Shape Searching (**CFSS**) extension of cascaded regression [27] and a multi-view version of Constrained Local Neural Fields (**CLNF**) model [3]. We used multi-view initialization for CLNF model. As the GN-DPM and SDM models we used were only trained on 49 landmarks, we only evaluated them for those landmarks. Our comparison against state-of-the-art algorithms is based on the original author's implementations.



(a) frontal (b) profile

Figure 2: *Cumulative error curves on IJB-FL.* Measured as the mean Euclidean distance from ground truth normalized by the face size. Note that we use 49 points for both frontal and profile images.

4.2 Ablation experiments

We performed three ablation experiments to see how the individual elements of our pipeline affect the final results. First, we removed the head pose estimation module and performed sparse landmark detection followed by dense landmarks (three head pose estimation models in parallel). We picked the model with the highest converged likelihood to determine the final landmarks. Second, we did not use sparse landmark detection, instead we used estimated head pose to initialize a CLNF dense landmark detector (Section 3.1). Finally, we performed just a CLNF dense landmark detection.

4.3 Datasets

We evaluated our works on three publicly available in-the-wild datasets:

300-W is a popular dataset which contains images from the HELEN [11], LFPW [4], AFW [28] and IBUG [17]. 300-W provides the ground truth bounding boxes and manually annotated 68 landmarks.

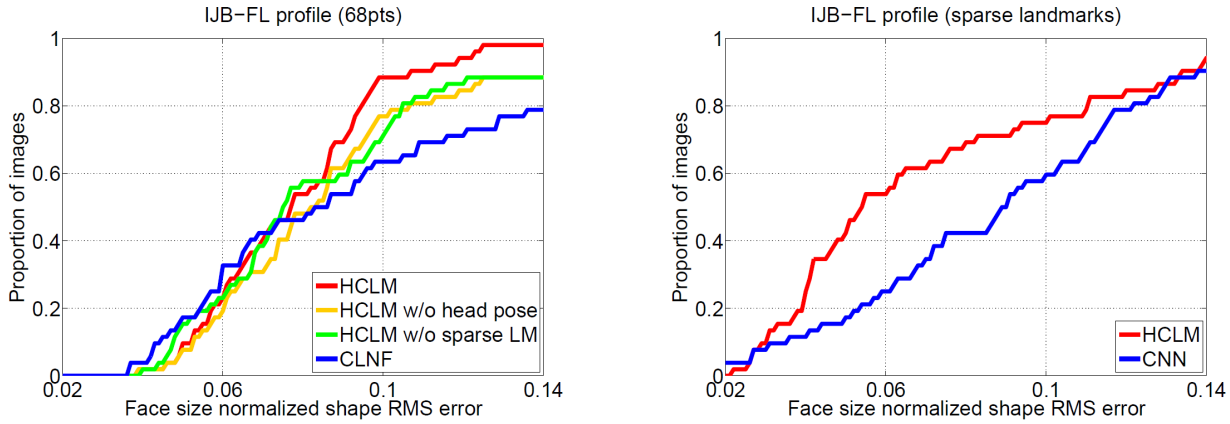
AFLW dataset contains 24,386 face images from Flickr. Each face is annotated with up to 21 landmarks. AFLW provides head pose estimation obtained by fitting a 3D mean face.

IJB-FL is a new dataset which has a substantial proportion of non-frontal images. It is a subset of IJB-A [9] which is a face recognition benchmark and includes challenging unconstrained faces with full pose. We took a sample of 180 images (128 images for frontal and 52 images for profile) from IJB-A, and manually annotated up to 68 facial landmarks in each image (depending on visibility). This is a very challenging subset containing a number of images in non-frontal pose (see Figure 5).[1]

4.4 Methodology

For the CNN training for head pose estimation and sparse landmark detection, we used 300-W and AFLW datasets. More specifically, we used the training partitions of HELEN (2,000 images), LFPW (811 images), AFW (337 images), and AFLW (14,920 images). In addition, to avoid over-fitting, the images were augmented three times with enlargement by 10% and

[1] Annotations of the IJB-FL dataset are available for research purposes.



(a) Ablation experiments on profile (b) Sparse landmarks error on profile

Figure 3: *Ablation experiments and sparse landmarks error on IJB-FL.* (a) We use 68 points for comparison. Our approach is robust in presence of large head pose variation in profile images (b) Cumulative error of sparse landmarks on profile images. Note that sparse landmarks from CNN are refined by HCLM model

20% on the face bounding box. We began training at a learning rate of 0.001 and dropped the learning rate to $1e^{-8}$ with 0.1 step and set the momentum to 0.9. For the CLNF patch expert training, we used Multi-PIE [8] and the training partitions of HELEN and LFPW. Furthermore, we used a multi-view and multi-scale approach as described in Baltrušaitis et al. [3].

For the test of head pose estimation and landmark detection, we used the test set of LFPW (224 images), HELEN (330 images), IBUG (135 images), the remaining AFLW (4,972 images), and IJB-FL (180 images). Faces in the training set were not used in testing.

In case of the 300-W (LFPW, HELEN, AFW, IBUG) datasets, we used the bounding boxes provided by the organizers[17] which were based on a face detector. In case of the AFLW and IJB-FL datasets, we used the bounding boxes based on the ground truth landmarks since the automatic face detection could not detect some faces in the data.

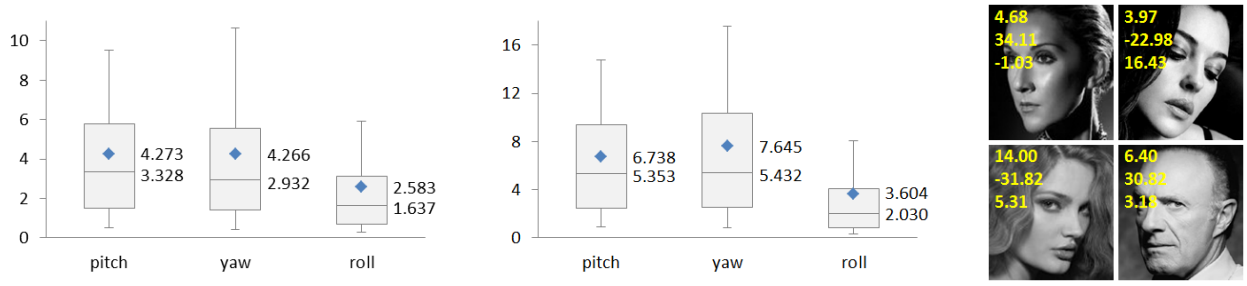
5 Results and discussion

5.1 Facial landmark detection

The results of comparing our HCLM model to the previously mentioned baselines are in Figure 1 (300-W) and Figure 2 (IJB-FL). Our model demonstrates competitive or superior performance to most of the baselines on both datasets. The better performance of our model is particularly clear at higher error rates and on profile images (see Figure 2b). This indicates that our model is more robust than the baselines, and that it is able to fit better on more complex images. This is due to the both better initializations and combination of *holistic* and *local* approaches of our model (see the following sections).

5.2 Ablation experiments and sparse landmark detection

We performed ablation experiments to see how the individual elements of our pipeline affect the final results. Figure 3a shows the performance of our full pipeline compared with the following three cases: only using the dense landmark detector (CLNF), dense landmark detector with head pose estimation (without any sparse landmark anchors), and CLNF with



(a) Error on the 300-W (b) Error on the AFLW (c) Sample images

Figure 4: *Head pose estimation results from 300-W, and AFLW.* (a) and (b) show the mean and median of absolute errors for pitch, yaw and roll respectively, while (c) shows predicted head pose sample images.

sparse landmark anchors (no head pose estimation). The result shows that each individual module - head pose estimation, sparse landmark detection - is important for HCLM model to improve the performance.

In addition, we evaluated the performance of sparse landmark detection methods on the IJB-FL dataset. Figure 3b shows the cumulative error of sparse landmarks on profile images. Sparse landmarks were used to anchor our dense landmark detector and were refined by our HCLM model. The result shows that HCLM model improves the accuracy of sparse landmarks which were predicted by the CNN.

5.3 Head pose estimation

We evaluated our head pose detector on the 300-W and AFLW datasets. The quantitative results are in Figure 4a and Figure 4b, while Figure 4c shows some sample estimations.

In addition, we measured three view classification accuracy based on the head pose estimation results. The ranges of frontal, left and right sides are from -30° to 30°, greater than 30°, less than -30° respectively. The classification accuracy on 300-W and AFLW datasets is 95.2% and 91.5% respectively. This demonstrates that our CNN head pose estimation works well and is useful for the further steps of the pipeline.

6 Conclusions

In this paper, we presented a new model, HCLM which unifies local and holistic facial landmark detection by integrating these three methods: head pose estimation, sparse-holistic landmark detection and dense-local landmark detection. Our new model was evaluated on three challenging datasets: 300-W, AFLW and IJB-FL. It shows state-of-the-art performance and is robust, especially in the presence of large head pose variations.

In the future, we will apply our model to video processing. Face tracking in the video is very challenging in the presence of head pose variations beyond frontal poses. We believe that our model will demonstrate competitive performance, especially with the help of temporal information in the video.

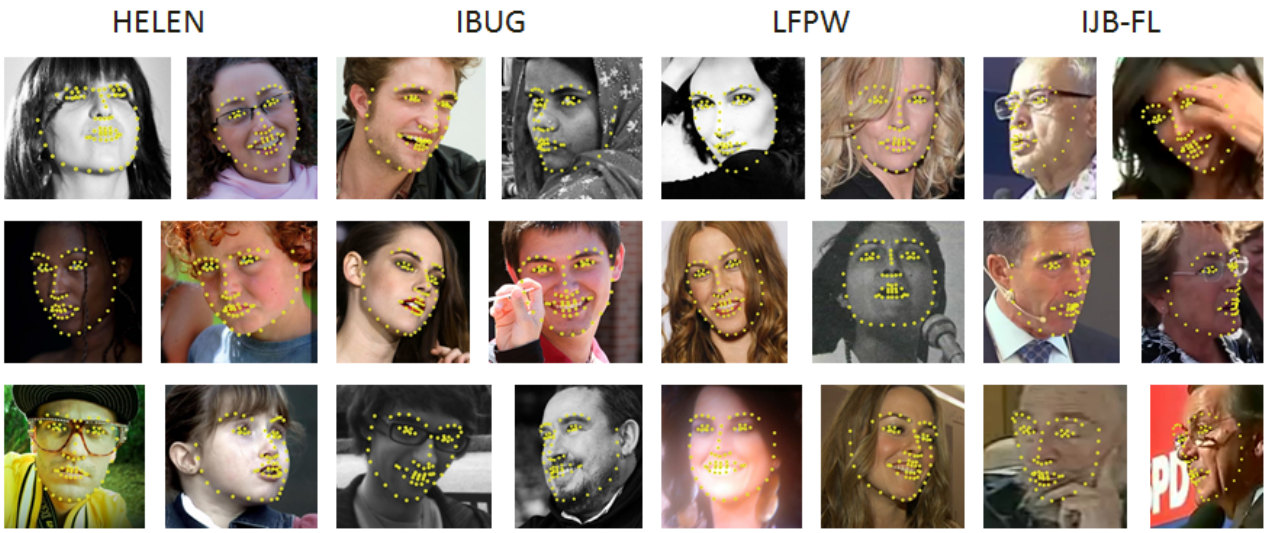


Figure 5: *Example detection results on the 300-W and IJB-FL.* Each column presents images from the subsets of the 300-W (HELEN, IBUG and LFPW) and IJB-FL.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Continuous conditional neural fields for structured regression. In *Computer Vision–ECCV 2014*, pages 593–608. Springer, 2014.

[4] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Narendra Kumar. Localizing parts of faces using a consensus of exemplars. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2930–2940, 2013.

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by Explicit Shape Regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894. Ieee, jun 2012. ISBN 978-1-4673-1228-8. doi: 10.

1109/CVPR.2012.6248015. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6248015.

[6] Błażej Czupryński and Adam Strupczewski. *Active Media Technology: 10th International Conference, AMT 2014, Warsaw, Poland, August 11-14, 2014. Proceedings*, chapter High Accuracy Head Pose Tracking Survey, pages 407–420. Springer International Publishing, Cham, 2014. ISBN 978-3-319-09912-5. doi: 10.1007/978-3-319-09912-5_34. URL http://dx.doi.org/10.1007/978-3-319-09912-5_34.

[7] Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real Time Head Pose Estimation with Random Regression Forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 617–624, 2011.

[8] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2008.

[9] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1931–1939. IEEE, 2015.

[10] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[11] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012.

[12] David G Lowe. Distinctive image features from scale invariant keypoints. *Int'l Journal of Computer Vision*, 60:91–11020042, 2004. ISSN 0920-5691. doi: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94. URL http://portal.acm.org/citation.cfm?id=996342.

[13] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–26, apr 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.106. URL http://www.ncbi.nlm.nih.gov/pubmed/19229078.

[14] Chavdar Papazov, Tim K Marks, and Michael Jones. Real-time 3D Head Pose and Facial Landmark Estimation from Depth Images Using Triangular Surface Patch Features. In *CVPR*, 2015. ISBN VO -. doi: 10.1109/CVPR.2015.7299104.

[15] Georgia Rajamanoharan and Timothy F Cootes. Multi-View Constrained Local Models for Large Head Angle Facial Tracking. In *ICCV*, 2015. ISBN 9780769557205. doi: 10.1109/ICCVW.2015.128.

[16] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[17] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[18] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 2015.

[19] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Deformable Model Fitting by Regularized Landmark Mean-Shift. 91(2):200–215, 2011. doi: 10.1007/s11263-010-0380-4.

[20] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013. ISSN 10636919. doi: 10.1109/CVPR.2013.446.

[21] Georgios Tzimiropoulos. Project-Out Cascaded Regression with an application to Face Alignment. In *CVPR*, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298989.

[22] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[23] Nannan Wang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Facial Feature Point Detection: A Comprehensive Survey. page 32, 2014. URL http://arxiv.org/abs/1410.1037.

[24] Xuehan Xiong and Fernando Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[25] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, and Peter Robinson. Face Alignment Assisted by Head Pose Estimation. In *BMVC*, pages 1–13, 2015.

[26] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

[27] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[28] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.