embedded VISION SUMMIT 2018

Understanding and Implementing Face Landmark Detection and Tracking

Dakala Jayachandra

22, May 2018

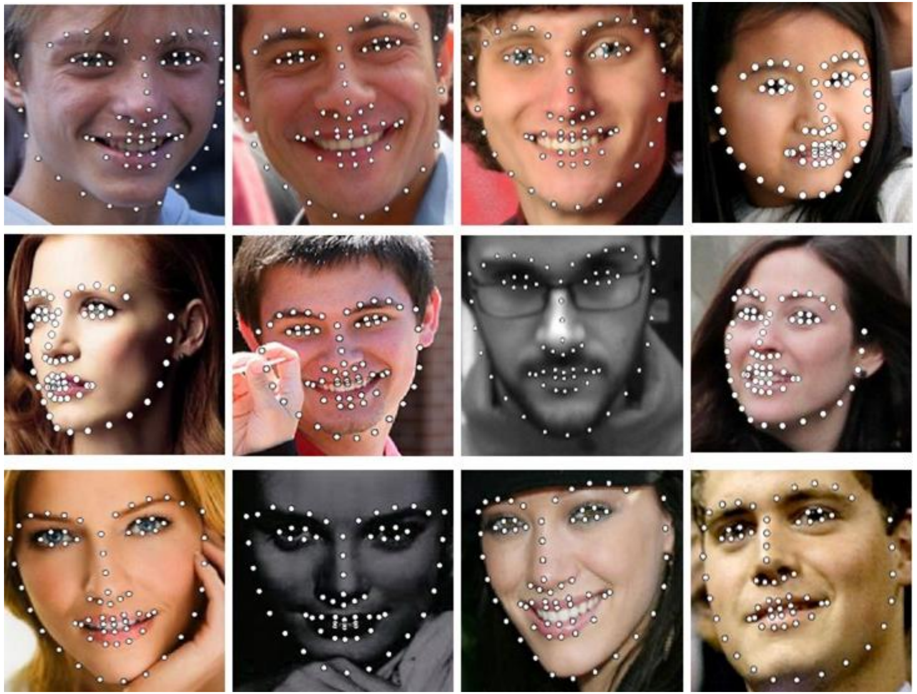
PATHPARTNER

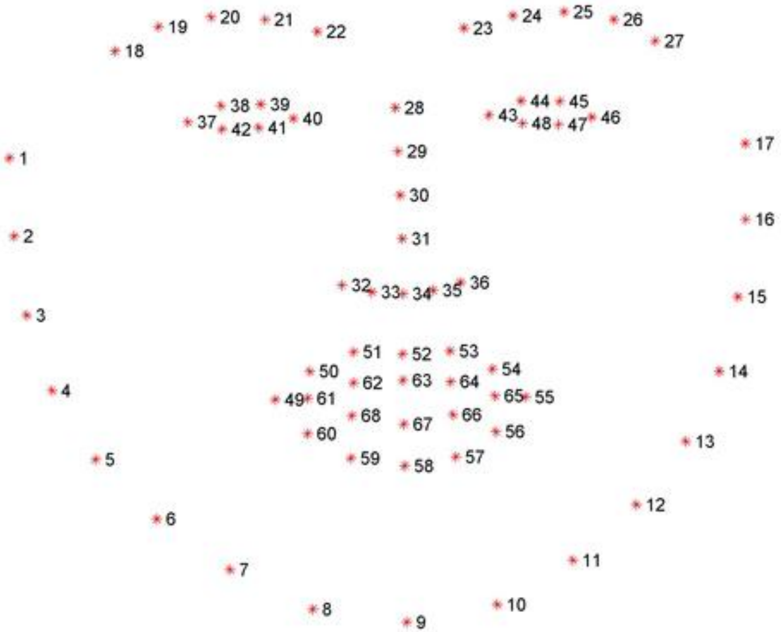
Outline

- Introduction
- Face landmark detection
  - Cascaded shape regressors (CSR)
  - Extension to CSRs
- Face landmark tracking
  - Definition
  - Approaches
- Online learning
- Notes on embedded implementation

Disclaimer: Focus is on ideas rather than on numbers!

What is Face Landmark Detection?

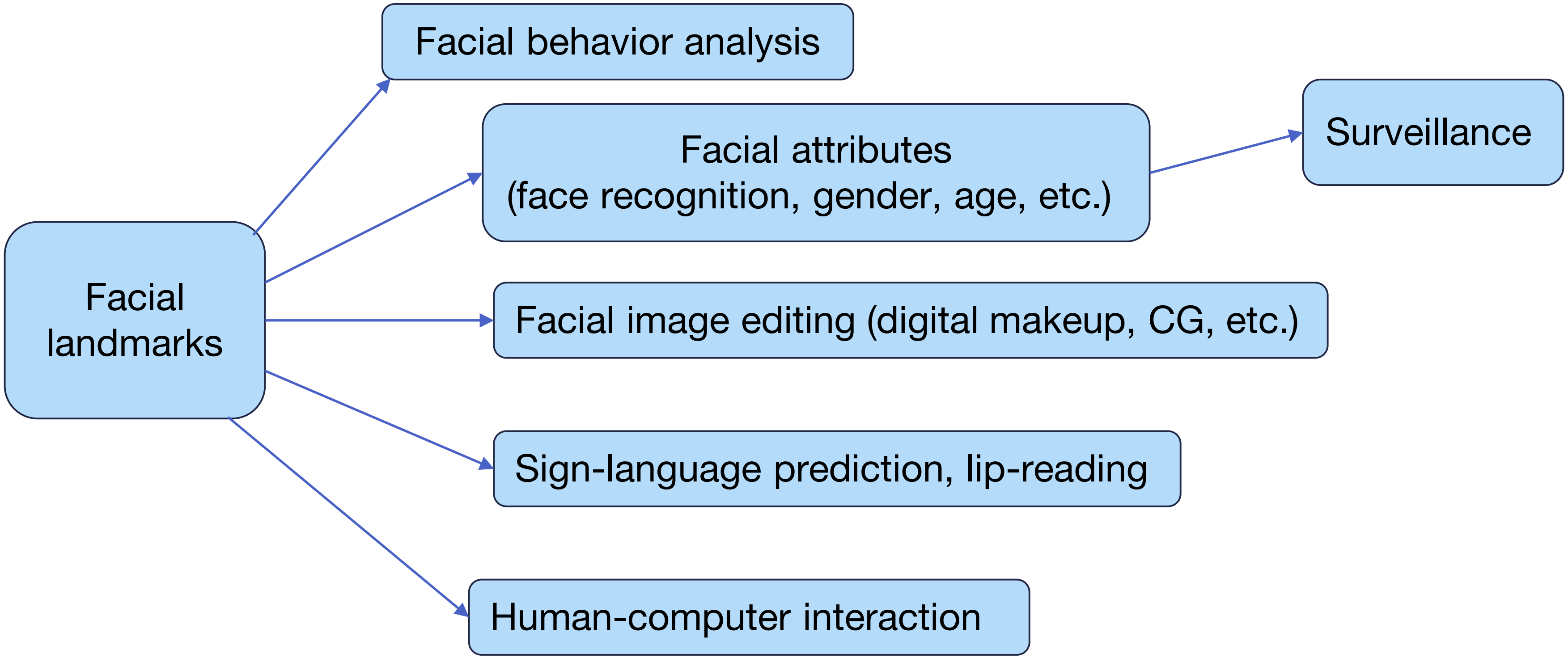




All about finding facial structures such as eyes, nose, mouth etc. by using the face image!

Courtesy: (left) http://mmlab.ie.cuhk.edu.hk/projects/TCDCN/img/2.jpg, (right) http://mmlab.ie.cuhk.edu.hk/projects/TCDCN/img/2.jpg

Applications [1]



Do we need landmarks if we use DNN?

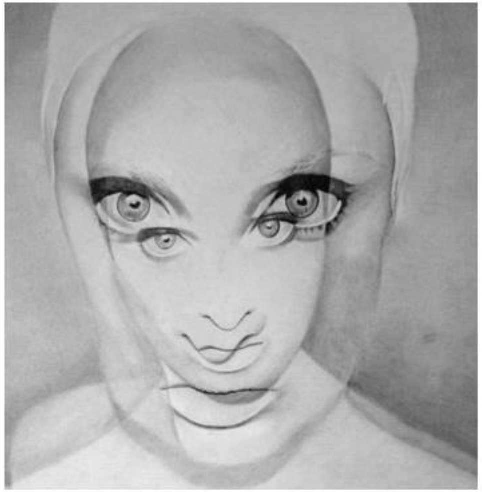
- Without face shape,
  - DNN model complexity may grow significantly
  - May need more amount of labeled data
- Tasks such as digital make-up, eye gaze, etc. are heavily dependent on face landmarks
- For some problems, working on face shape may be sufficient and economical

Do we need landmark if we use DNN?

- For instance, in face recognition, face shape acts a strong priori
- Face recognition accuracy on LFW dataset using Deep Face [10]

With only face detection module	87.9 %
With face and **landmark detection** modules	94.3 %

Look at these pictures!

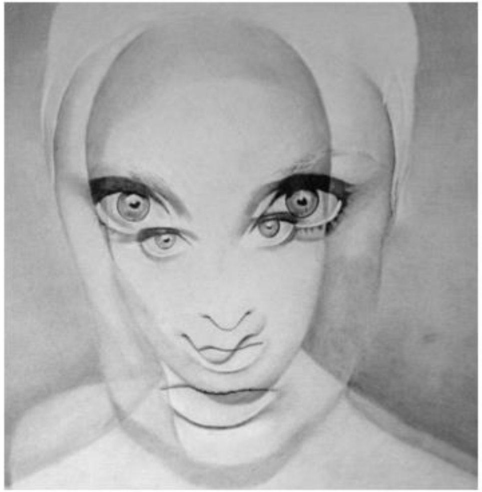


What are we trying to do?

Yes, we are trying to fit a face shape before recognizing identity/expression!

Courtesy: http://www.arts-pi.org.tn/rfmi2016/Zafeiriou_talk.pdf

Look at these pictures!



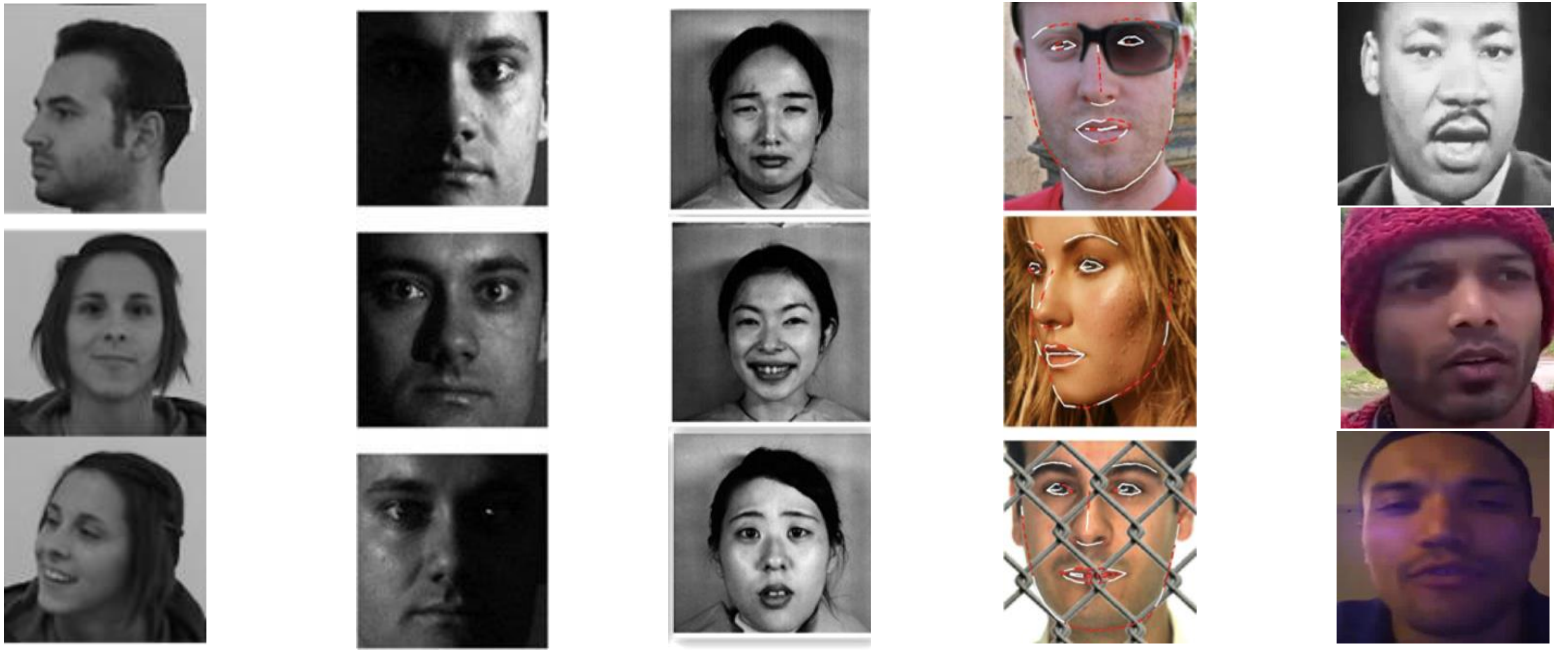
Isn’t it difficult to fit a shape? Why?

Pixel patterns don’t fit statistical models of face appearances we have already learned!

Courtesy: http://www.arts-pi.org.tn/rfmi2016/Zafeiriou_talk.pdf

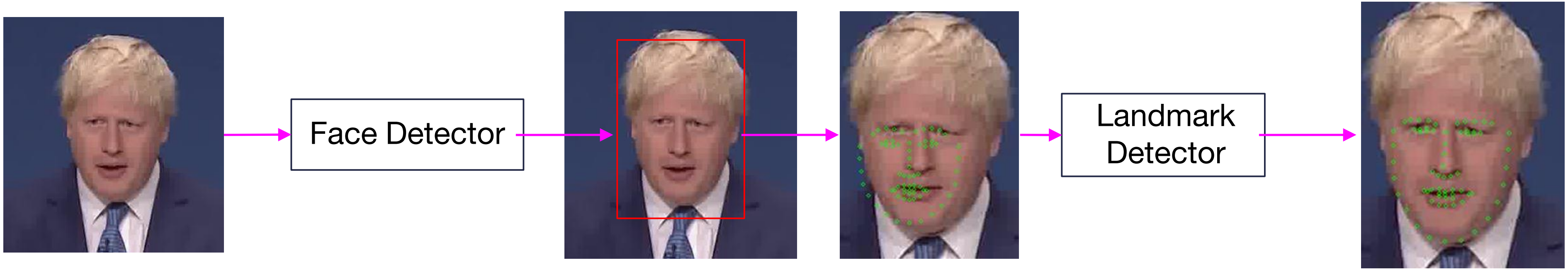
Why is this a hard problem ?

Head pose[a] Illumination[b] Expression[c] Occlusions[d] Sensors[e]



Courtesy: (a). Borghi, Guido et al. “Face-from-Depth for Head Pose Estimation on Depth Images.” *CoRR* abs/1712.05277 (2017), (b). Mitra, Sinjini. (2012). Gaussian Mixture Models for Human Face Recognition under Illumination Variations. Applied Mathematics. 03. 2071-2079. 10.4236/am.2012.312A286 (c). Michael J. Lyons, Shigeru Akemastu, Miyuki Kamachi, Jiro Gyoba, Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205 (1998) (d). http://www.consortium.ri.cmu.edu/data/APF/apf4.jpg , (e). https://ibug.doc.ic.ac.uk/resources/300-VW/

Face Landmark Detection: Typical flow



Essentially, need to learn a function that maps pixels to 2D coordinates!

Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Approaches (broadly) [1]

Generative mels

- Analysis by synthesis methods learn joint distribution, P(x,y)
- Examples: Active Shape Models (ASM), Active Appearance Models (AAM) etc.

Discriminative models

- Regression based methods learn conditional distribution, P(y|x)
- Examples: Cascaded Shape Regression based models like Supervised Descent Methods (SDM), Constrained Local Models (CLMs), etc.

Recently, with availability of data sets, discriminative models surpassed generative models !!

A Regression Problem

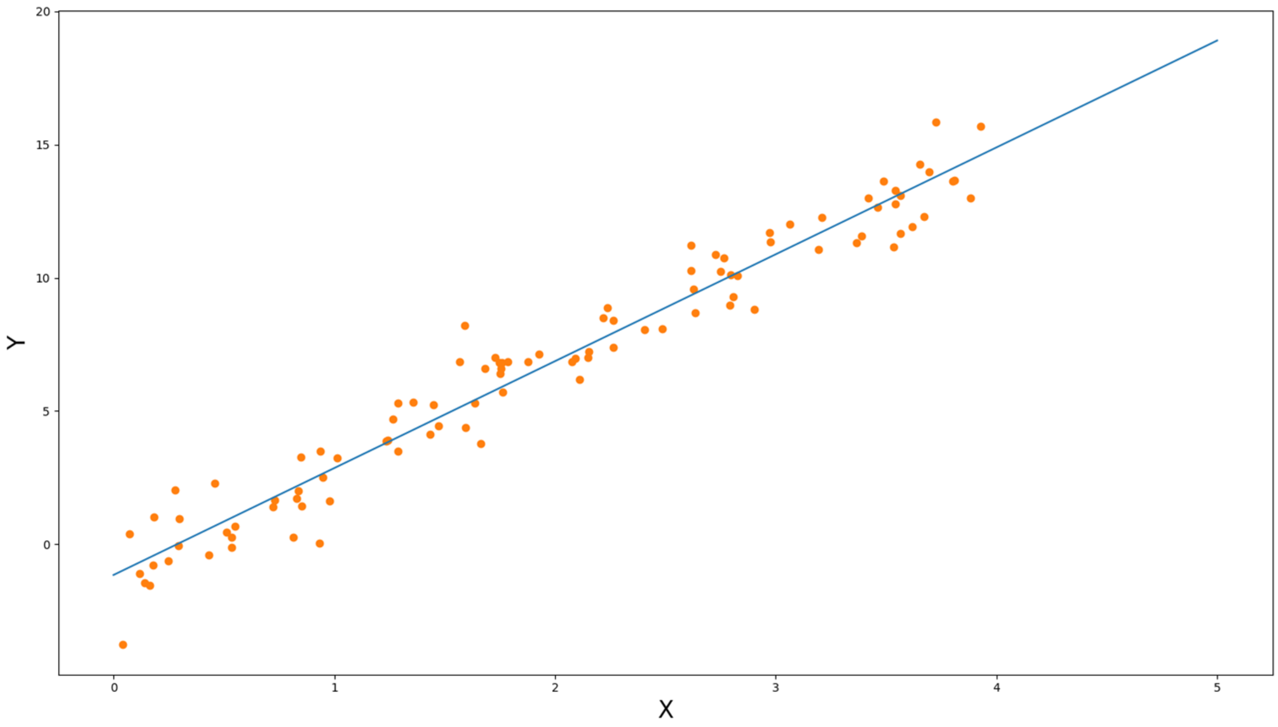
- Given a face image and a face shape initialization (S_0), regress for the shape residual between initial shape (S_0) and manually annotated ground truth shape (S_*)

- Target representation: $Y = \Delta S = S_* - S_0$

- Feature representation: $X = f(I, S_0)$

- Loss / objective function: $\underset{R}{\operatorname{argmin}} \| Y - RX \|_2^2$

Linear Least Squares Solution

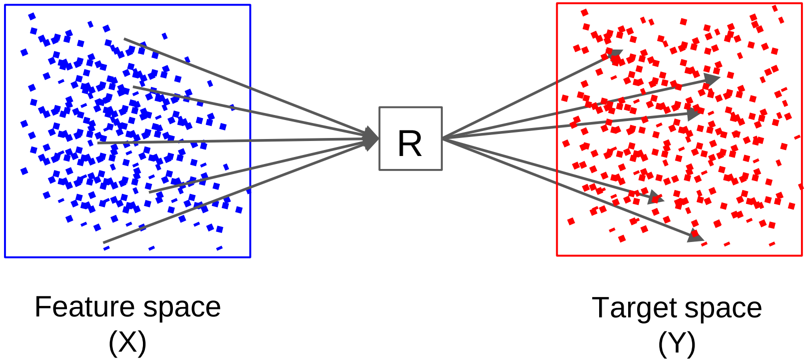


Given a set of scalar variables (X, Y), learn m and c values that best fit the data,

$$y = mx + c$$

Linear Least Squares Solution

- Given a set of multivariate feature (X) and target (Y) vectors, learn a mapping function R.



$$\text{Covariance}(X, Y) = \text{Covariance}(X, X) * R$$

$$YX^T = XX^T * R$$

$$R = YX^T (XX^T)^{-1}$$

Linear Least Squares Solution

- Is linear least squares good enough for landmark detection?

Need for Non-Linear Least Squares Solution

- Inherent mapping function of feature descriptors of face appearance to the target variables is ***non-linear*** in nature

Need for Non-Linear Least Squares Solution

- Inherent mapping function of feature descriptors of face appearance to the target variables is ***non-linear*** in nature

- Due to the wide variety of possible face appearances for the same face shape

Need for Non-Linear Least Squares Solution

- Inherent mapping function of feature descriptors of face appearance to the target variables is ***non-linear*** in nature

- Due to the wide variety of possible face appearances for the same face shape

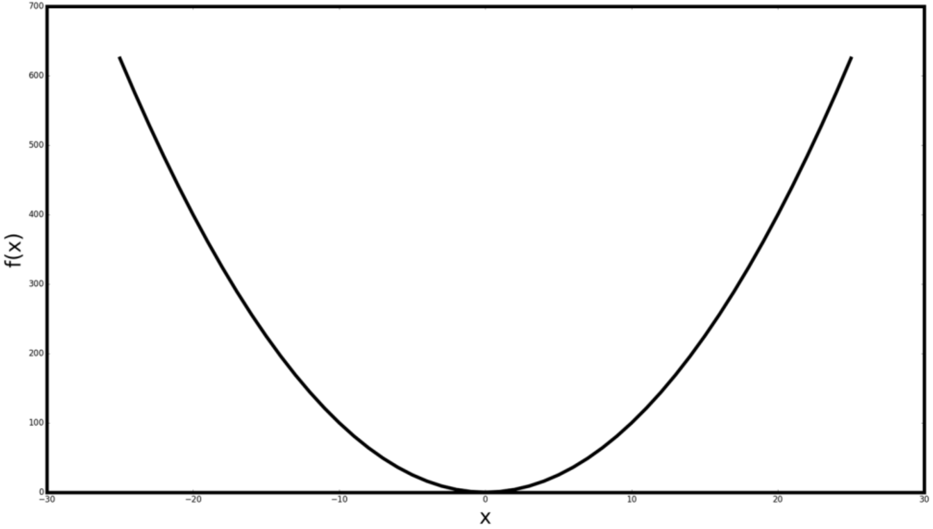
- For instance, pixel values of the same landmark vary wildly with slight illumination changes in the following scenario

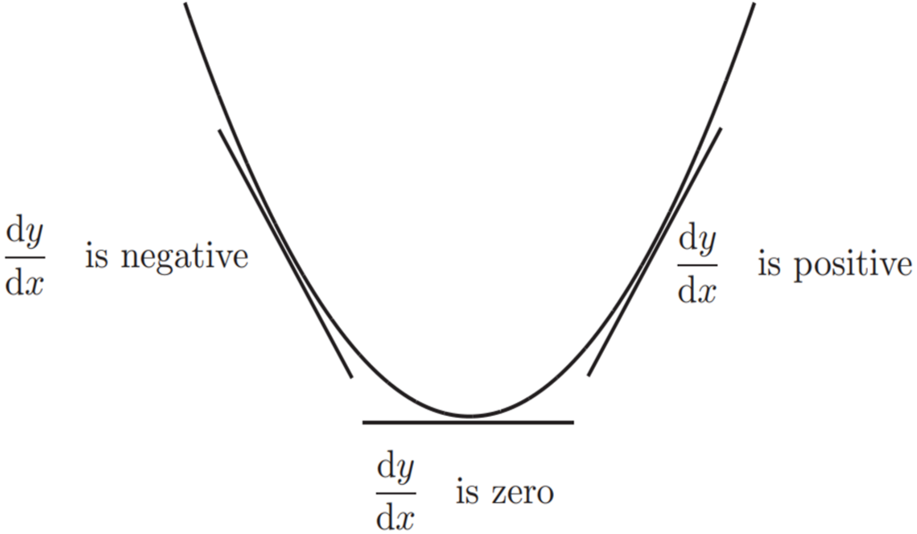



Courtesy: Gross, Ralph & Matthews, Iain & Cohn, Jeffrey & Kanade, Takeo & Baker, Simon. (2010). Multi-PIE. Proceedings of the International Conference on Automatic Face and Gesture Recognition. International Conference on Automatic Face and Gesture Recognition. 28. 807-813. 10.1016/j.imavis.2009.08.002.

Newton's Method: To Solve Non-Linear Least Squares Problem

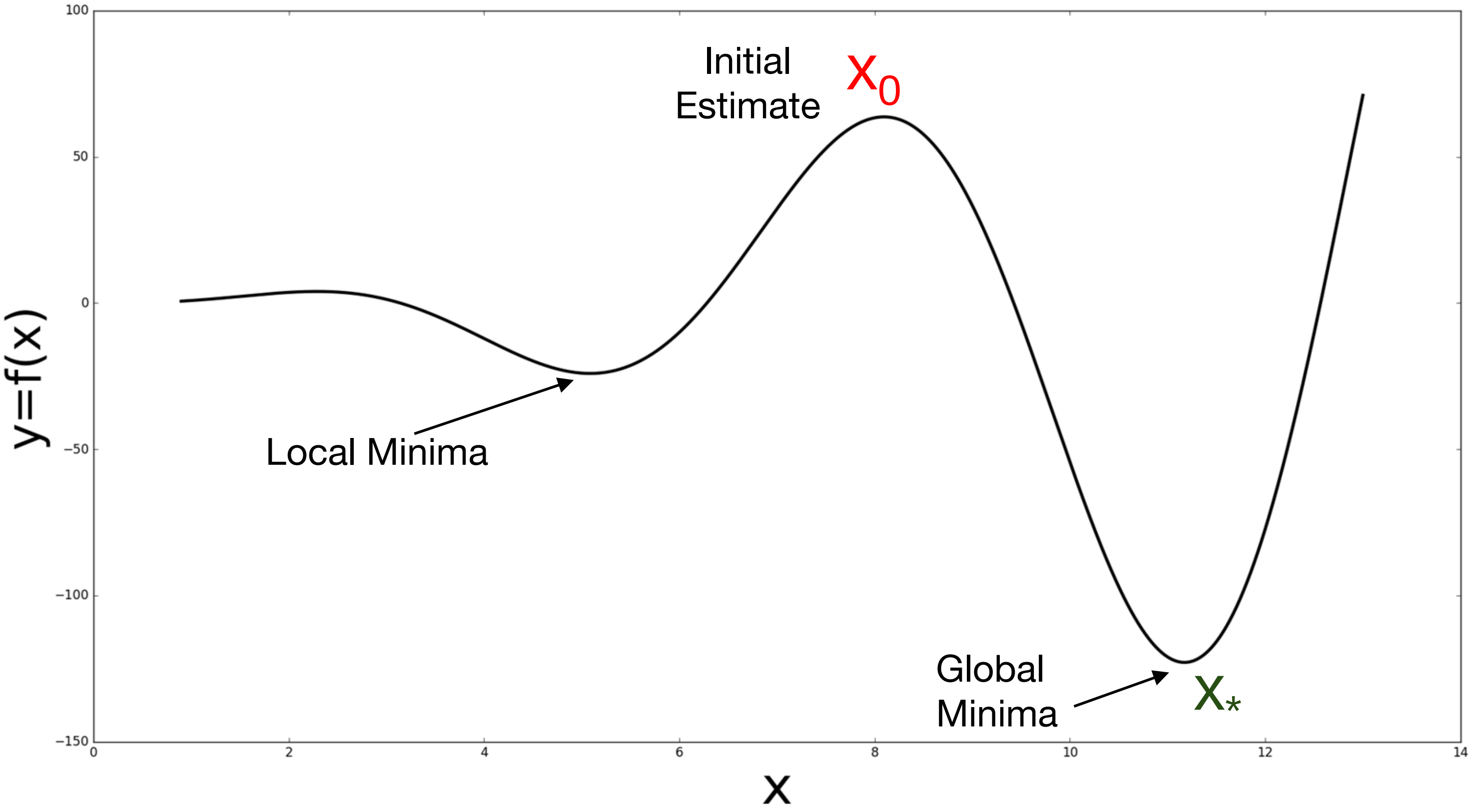
- Given a function f(x) = x^2, find an x value at which f(x) is minimum
- Value of x at which gradient of f(x) is zero is the global minima of f(x)



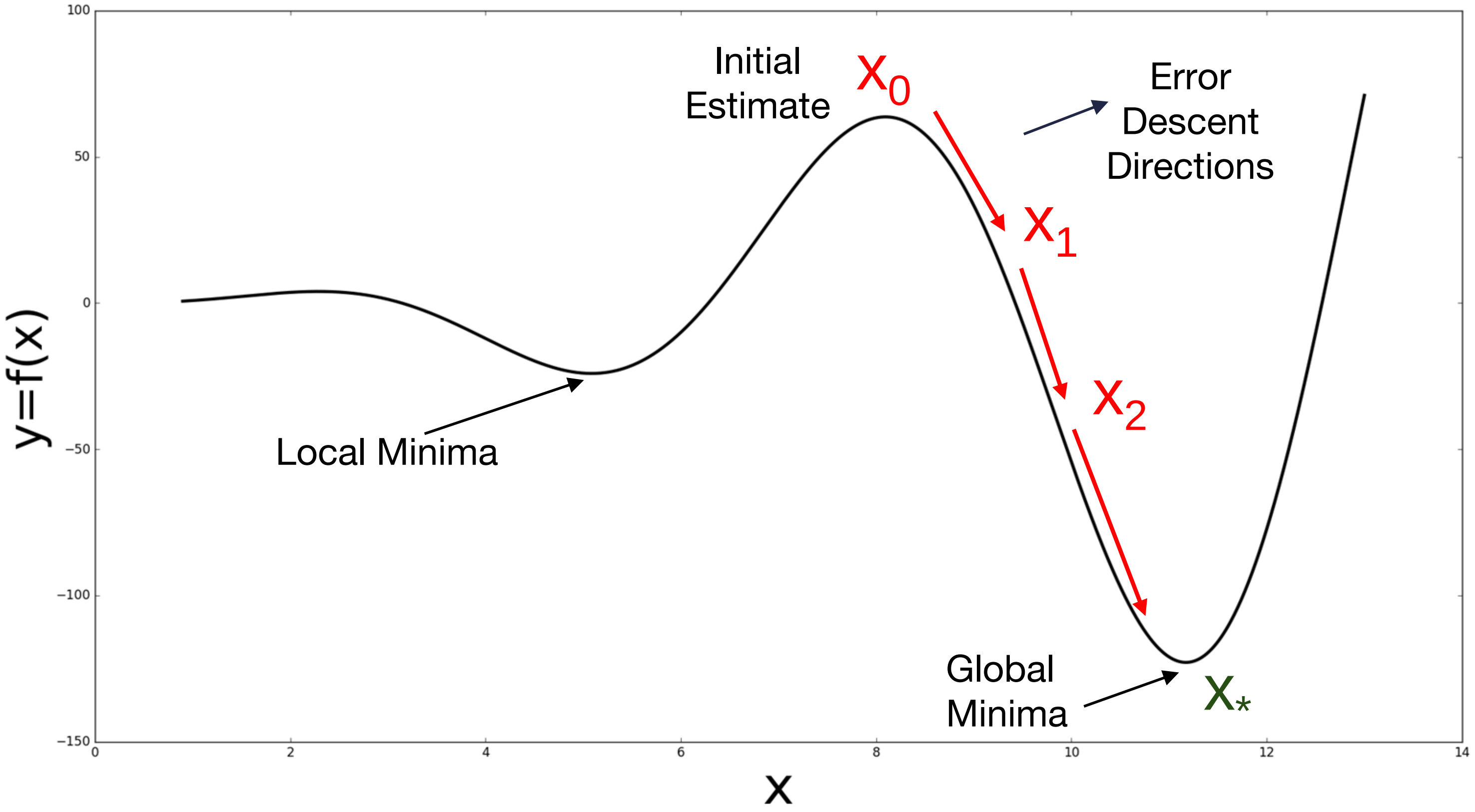


Courtesy: (right) http://www.mathcentre.ac.uk/resources/uploaded/mc-ty-maxmin-2009-1.pdf

Newton's Method: To Solve Non-Linear Least Squares Problem



Newton's Method: To Solve Non-Linear Least Squares Problem



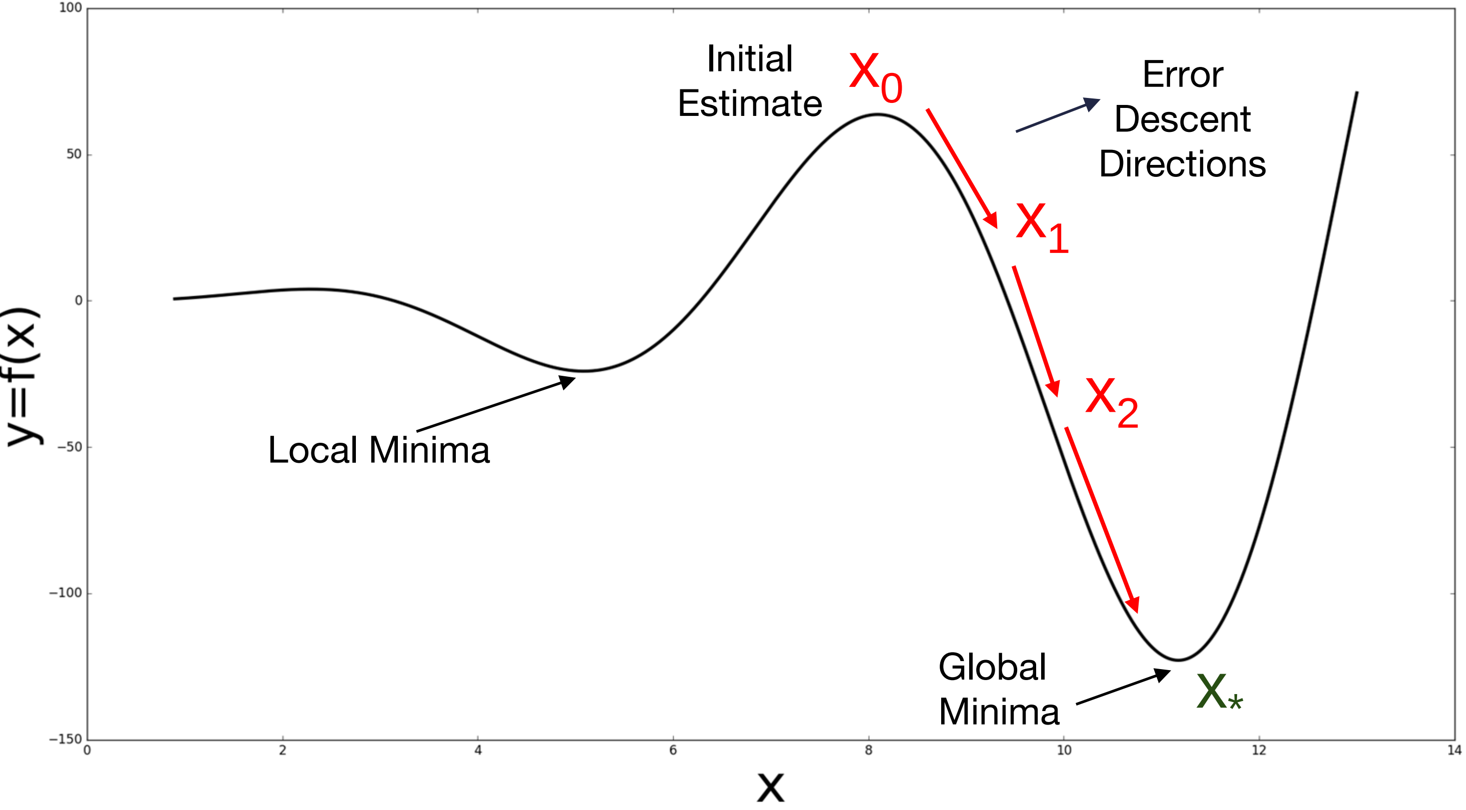
Newton's Method: To Solve Non-Linear Least Squares Problem

- A twice differentiable smooth function f(x) can be approximated using Jacobian and Hessian of f(x)

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

- Given a loss function f(x), Newton's method solves for error descent direction along x that minimizes the loss

Newton's Method: To Solve Non-Linear Least Squares Problem



$$X_1 = X_0 - H_f^{-1}(X_0) * J_f(X_0)$$

where,

H_f : Hessian matrix

- Local curvature of f(x) at $x=x_0$

J_f : Jacobian matrix

- Local slope of f(x) at $x=x_0$

Challenges in Applying Newton's Method to Computer Vision [2]

$$X_1 = X_0 - H_f^{-1}(X_0) * J_f(X_0)$$

- Hessian must be a positive definite in order to compute optimal global minima.
- Appearance feature descriptors may not be differentiable analytically. Numerical gradient and Hessian computation are computationally expensive.
- With very high dimensional features, Hessian matrix is too large. Inverting large matrices is expensive.

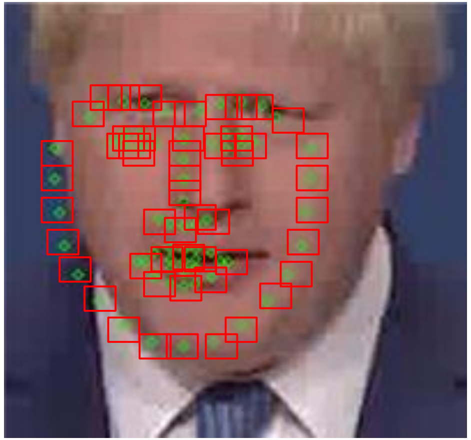
Cascaded Shape Regressors (CSR's) [2]

Idea: Learn a cascade of linear shape regressors!

- Essentially, learn error descent directions from the training data
- Equivalent to piecewise linear approximation of a nonlinear mapping function

Cascaded Shape Regressors [2]

Feature extraction: $f(I,S_0)$



$$Y = \Delta S = S_* - S_0$$

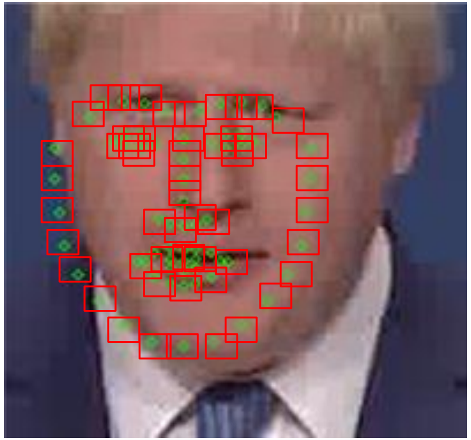
$$X = f(I, S_0)$$

$$R = YX^T(XX^T)^{-1}$$

Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Cascaded Shape Regressors [2]

Feature extraction: $f(I, S_0)$



$$Y = \Delta S = S_* - S_0$$

$$X = f(I, S_0)$$

$$R = YX^T(XX^T)^{-1}$$

Face image with mean shape as init (s_0) → Cascade 1: $S_1 = S_0 - R_0 * f(I_n, S_0)$ → Cascade 2: $S_2 = S_1 - R_1 * f(I_n, S_1)$ → Cascade 3: $S_3 = S_2 - R_2 * f(I_n, S_2)$ → Final shape estimation result (s_3)

Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Cascaded Shape Regressors [2]

What are the optimal representations of face appearance and face shape?

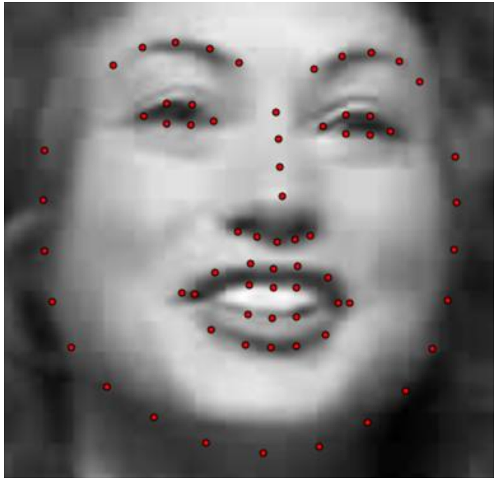
Optimal Face Appearance Representation

- Histogram of oriented Gradients (HoG) and Scale-Invariant Feature Transform (SIFT)
- Local Binary Features (LBF) [7]
- Convolutional Neural Networks for feature learning

Optimal Face Appearance Representation

Histogram of oriented Gradients (HoG) and Scale-Invariant Feature Transform (SIFT)

- Off-the-shelf feature extraction methods
- Heavily used in several CSR approaches, like SDM, Global SDM, etc.

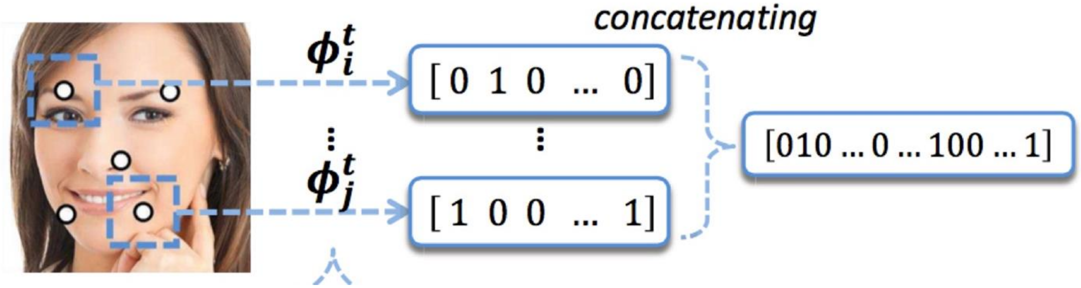


Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Optimal Face Appearance Representation

Local Binary Features (LBF) [7]

- Learns feature mapping functions from data using ensembles of regression trees
- Computationally very efficient during inference time

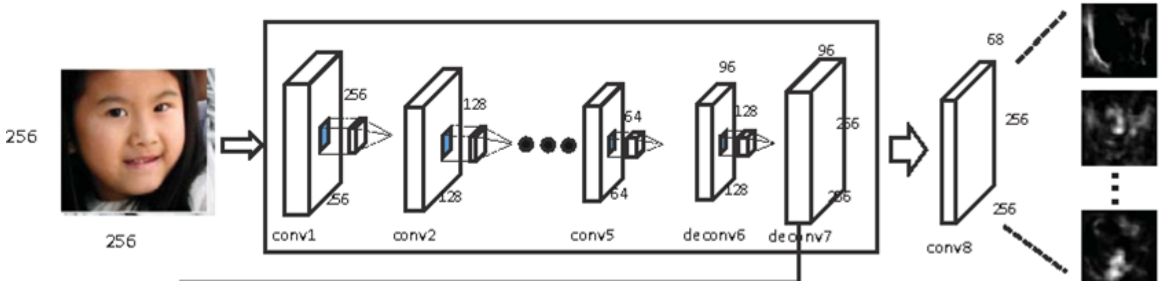


Courtesy: http://freesouls.github.io/2015/06/07/face-alignment-local-binary-feature/index.html

Optimal Face Appearance Representation

Convolutional Neural Networks for feature learning

- Learns optimal non-linear feature representations for the task at hand
- Needs huge amounts of training data
- Run time complexity is quite high on embedded systems



Courtesy: Lai, Hanjiang & Xiao, Shengtao & Pan, Yan & Cui, Zhen & Feng, Jiashi & Xu, Chunyan & Yin, Jian & Yan, Shuicheng. (2016). Deep Recurrent Regression for Facial Landmark Detection. IEEE Transactions on Circuits and Systems for Video Technology. PP. 1-1. 10.1109/TCSVT.2016.2645723.

Optimal Face Shape Representation

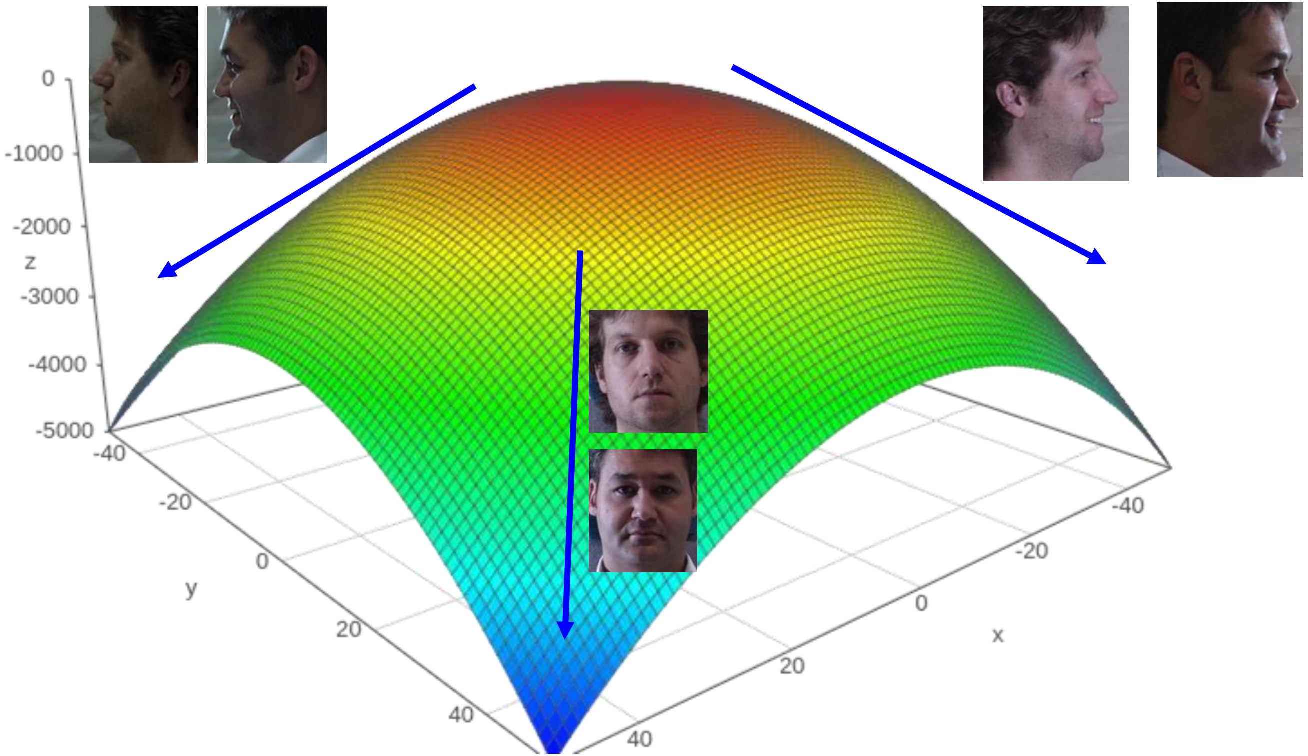
- **Point Distribution Model (PDM):**
  - A shape (S) is parameterized in terms of p = [q, c] where
    - q represents rigid shape parameters and
    - c represents flexible shape parameters

$$S = t_q(S_0 + B_s * c)$$

    - S_0 - mean face shape
    - B_s - Orthogonal basis of flexible face shape deformations

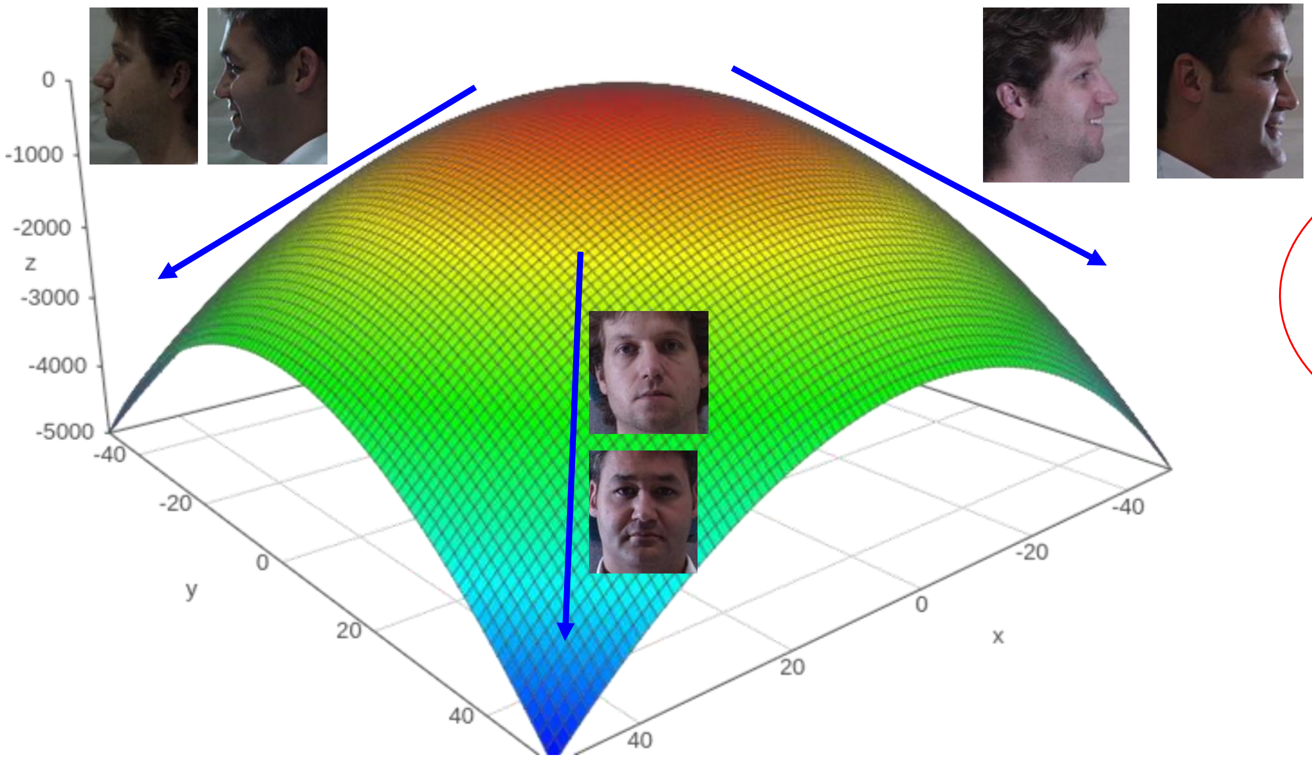
- **Structured Point Distribution Model (SPDM):**
  - In addition to 2D landmark coordinates, **visibility labels** (1,0) of each landmark is also taken into account
  - Combines the PCA bases of rigid, non-rigid and visibility components of face shapes and generates a joint parametric form

Why Cascaded Shape Regressors Can't Solve Multiview Face Alignment? [3]



Courtesy: Gross, Ralph & Matthews, Iain & Cohn, Jeffrey & Kanade, Takeo & Baker, Simon. (2010). Multi-PIE. Proceedings of the International Conference on Automatic Face and Gesture Recognition. International Conference on Automatic Face and Gesture Recognition. 28. 807-813. 10.1016/j.imavis.2009.08.002.

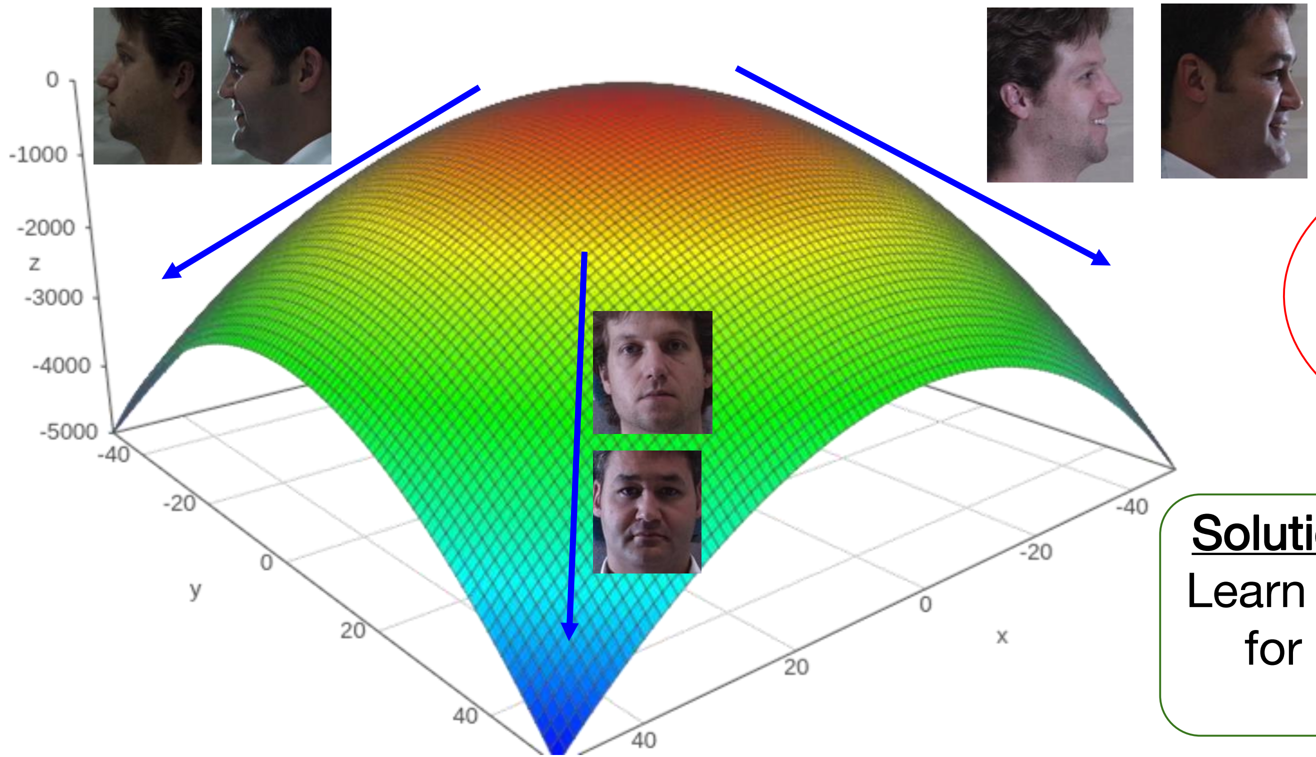
Why Cascaded Shape Regressors Can't Solve Multiview Face Alignment? [3]



Conflicting error descent directions !!!

Courtesy: Gross, Ralph & Matthews, Iain & Cohn, Jeffrey & Kanade, Takeo & Baker, Simon. (2010). Multi-PIE. Proceedings of the International Conference on Automatic Face and Gesture Recognition. International Conference on Automatic Face and Gesture Recognition. 28. 807-813. 10.1016/j.imavis.2009.08.002.

Why Cascaded Shape Regressors Can't Solve Multiview Face Alignment? [3]



Conflicting error descent directions !!!

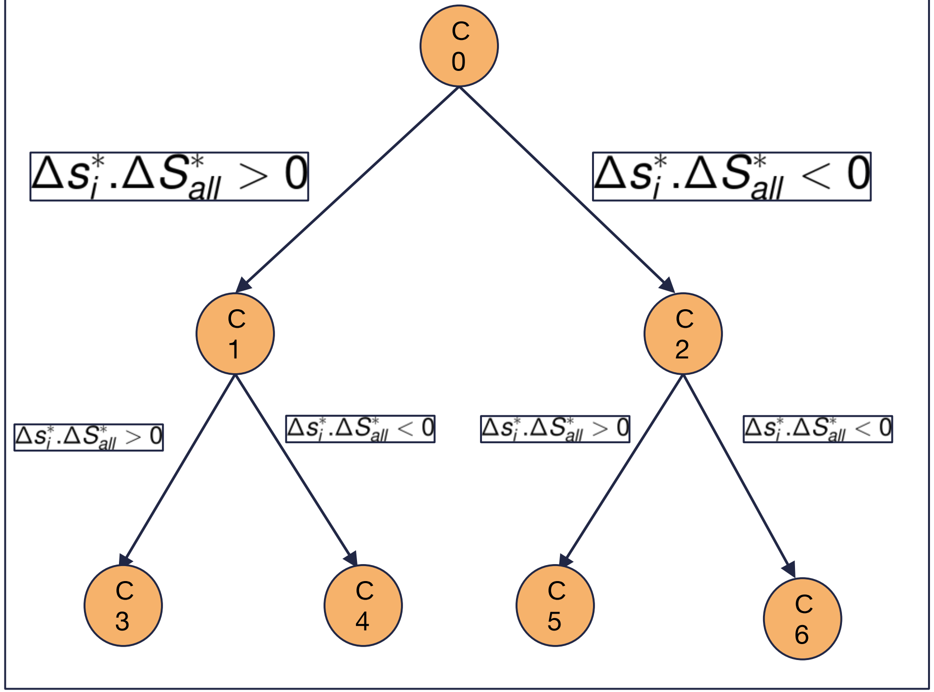
<u>Solution - Extended CSR</u>

Learn different regressors for different descent directions ...

Courtesy: Gross, Ralph & Matthews, Iain & Cohn, Jeffrey & Kanade, Takeo & Baker, Simon. (2010). Multi-PIE. Proceedings of the International Conference on Automatic Face and Gesture Recognition. International Conference on Automatic Face and Gesture Recognition. 28. 807-813. 10.1016/j.imavis.2009.08.002.

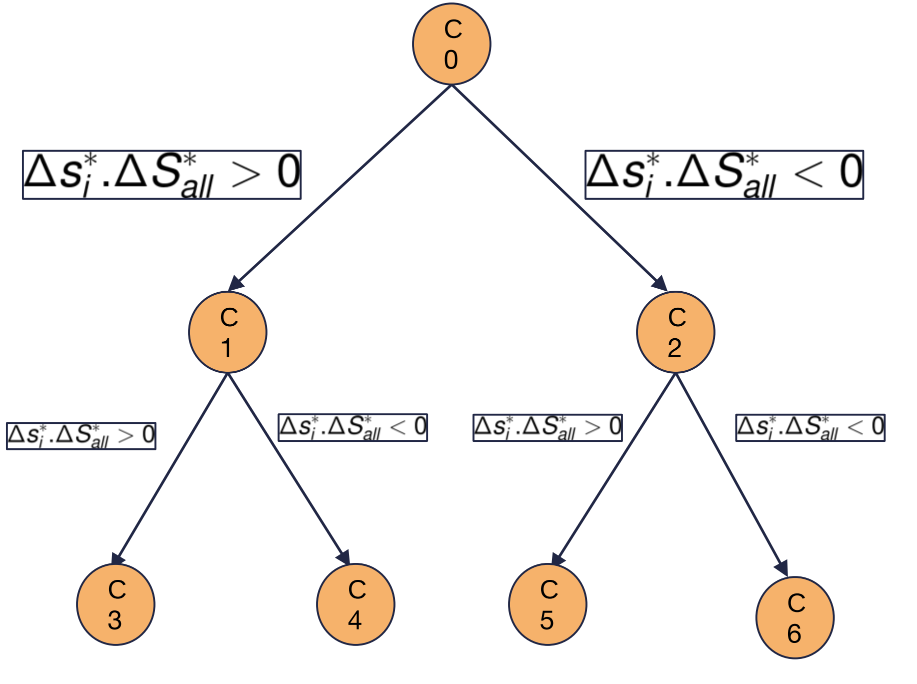
Extended CSR Approaches:

Global Supervised Descent Method (GSDM) [3]

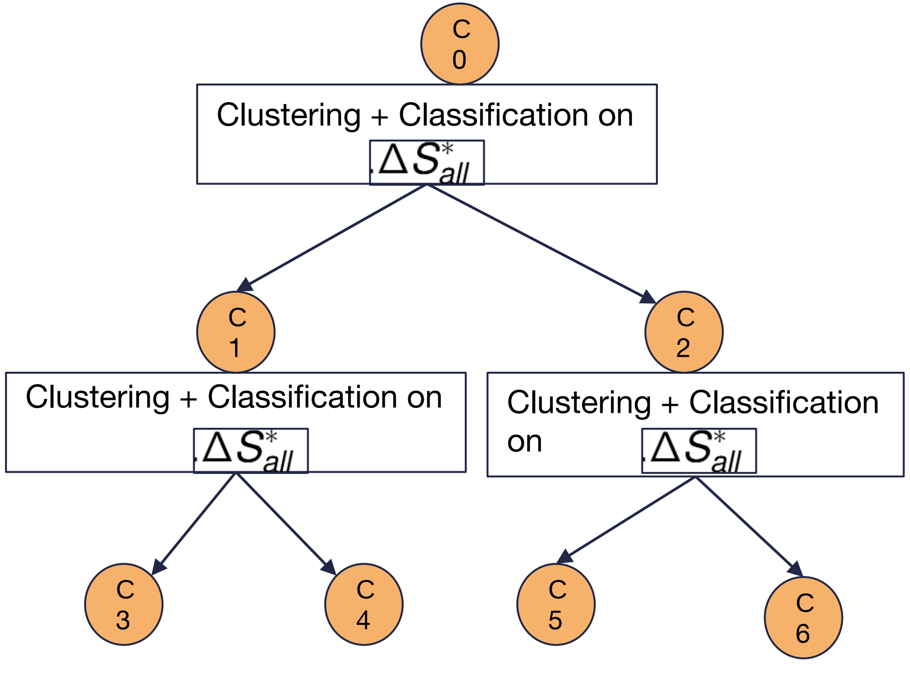


Extended CSR Approaches:

Global Supervised Descent Method (GSDM) [3]



Branching Cascaded Regression (BCR) [4]



Extended CSR Approaches:

Global Supervised Descent Method (GSDM) [3]

- Loss surface partition logic:
  - During training:
    - Groups the training samples based on the sign of dot product of shape residuals
    - Needs ground truth face shape to decide the descent direction
  - During inference:
    - Previous frame's estimated landmarks are used to decide the branching direction

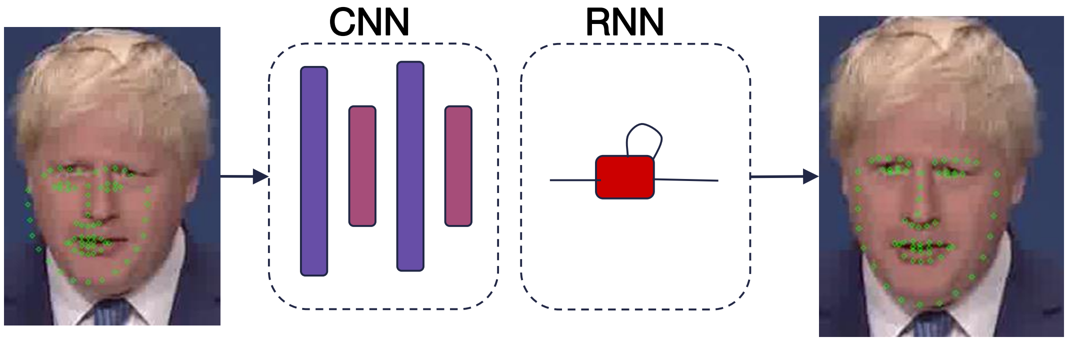
Branching Cascaded Regression (BCR) [4]

- Loss surface partition logic:
  - During training:
    - Applies a clustering algorithm to shape residuals
    - Learns a classifier to build a separating hyperplane between the clusters
  - During inference:
    - Employs the classifier models learned during training to decide which error descent direction to take for a given test sample

Extended CSR Approaches:

Mnemonic Descent Method (MDM) [5]

- Unlike GSDM and BCR, MDM avoids the need of explicit split by using RNNs
- Jointly trains a convolutional recurrent neural network in an end-to-end fashion



Summary: Extended CSR Approaches

Global Supervised Descent Method (GSDM) [3]

- Introduces the idea learning Domain of Homogenous Descent (DHD) directions by partitioning the parameter space
- Doesn’t weight feature descriptors of a landmark based on its visibility

Branching Cascaded Regression (BCR) [4]

- By learning a tree of cascaded shape regressors, this method addresses the problem of loss surface partition
- Weights feature descriptors of a landmark based on its visibility

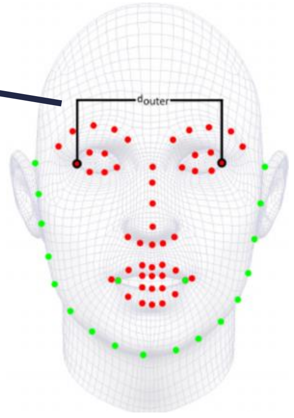
Mnemonic Descent Method (MDM) [5]

- Unlike GSDM and BCR, MDM avoids the need of explicit split by using RNNs
- Jointly trains a convolutional recurrent neural network in an end-to-end fashion

Evaluation Protocols for Benchmarking

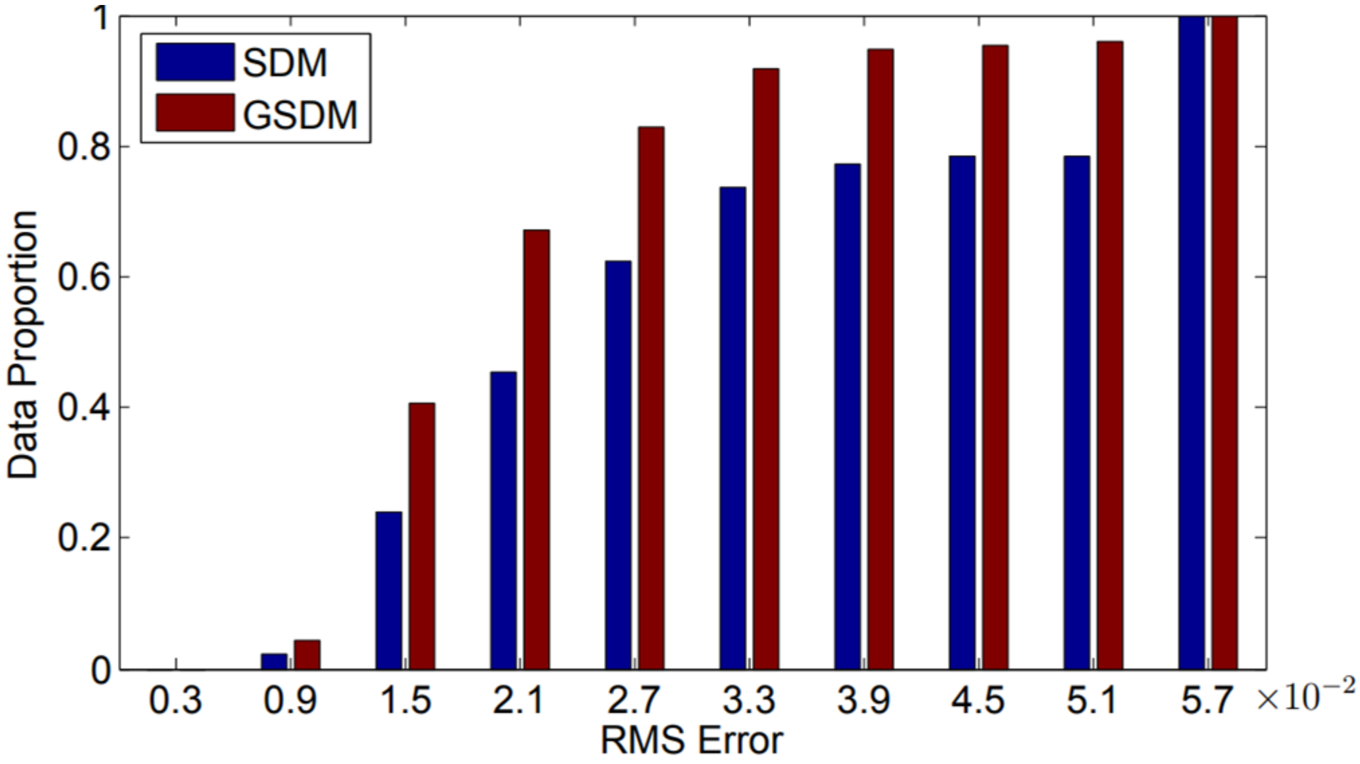
- Euclidean error normalized by Inter Ocular Distance (IOD)

- Area Under Curve (AUC) of Cumulative Error Distribution (CED)



Courtesy: Shen, Jie & Zafeiriou, Stefanos & G Chrysos, Grigorios & Kossaifi, Jean & Tzimiropoulos, Georgios & Pantic, Maja. (2015). The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results.

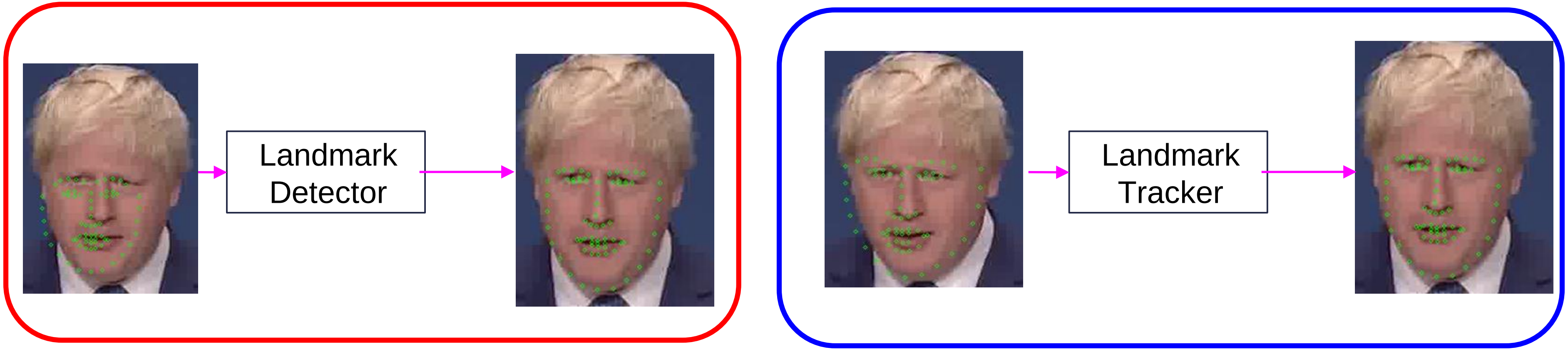
Performance Comparison Between SDM and Global - SDM on Distracted Driver Face (DDF) Dataset



Courtesy: Xiong, Xuehan & De la Torre, Fernando. (2015). Global supervised descent method. 2664-2673. 10.1109/CVPR.2015.7298882.

Face Landmark Tracking [6]

Aim of a landmark tracker is to exploit temporal coherence of faces in a video sequence



Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Face Landmark Tracking

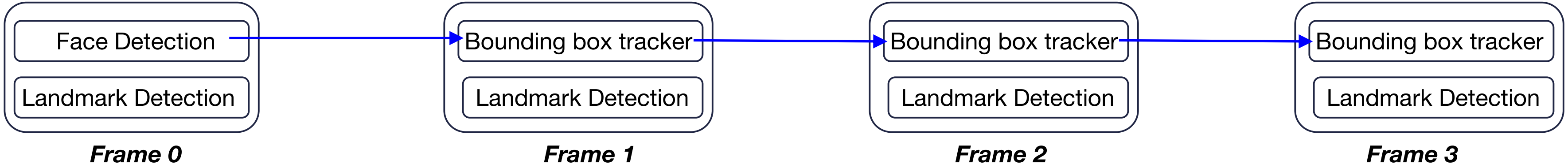
Face Detection and Landmark Detection for each frame:

Frame 0	Frame 1	Frame 2	Frame 3
Face Detection	Face Detection	Face Detection	Face Detection
Landmark Detection	Landmark Detection	Landmark Detection	Landmark Detection

Face Landmark Tracking

Face Detection and Landmark Detection for each frame:

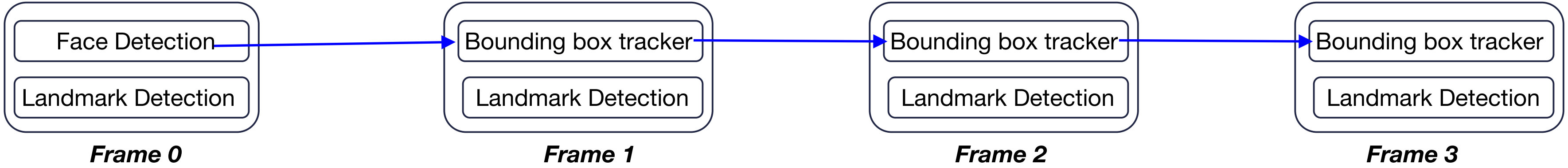
Face Detection only for the first frame and Landmark Detection for each frame:



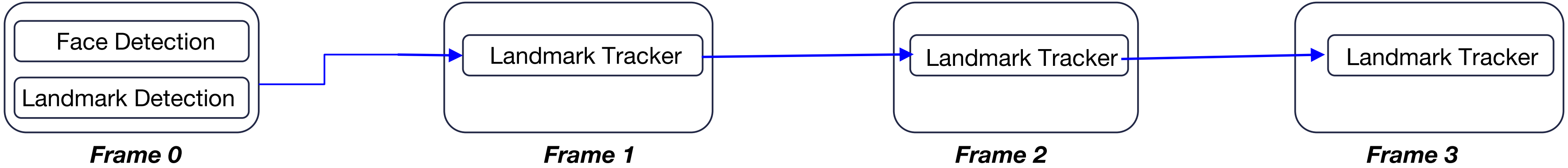
Face Landmark Tracking

Face Detection and Landmark Detection for each frame:

Face Detection only for the first frame and Landmark Detection for each frame:



Face Detection and Landmark Detection only for the first frame:



Landmark Tracking: High-level directions

Tracking long-term temporal dynamics of only shape deformations

Examples: Kalman filters and particle filters, etc.

Tracking short-term temporal dynamics of shape deformations by using face appearance

Examples: Parallel SDM and Continuous Cascaded Regressors, etc.

Tracking long-term temporal dynamics of both shape and appearance

Examples: Dynamic facial analysis using RNNs [8]

**** All the above tracking approaches suffer from drift issues. So, mechanisms for failure checking and re-initializing are imperative.**

Learning a Face Landmark Tracking Model [6]

Training a landmark detector

- Same shape initialization i.e., S_0, for all training samples

$$\boxed{S_{init} = S_0}$$

$$Y = \Delta S = S_* - S_{init}$$

$$X = f(I, S_{init})$$

$$R = YX^T(XX^T)^{-1}$$

Learning a Face Landmark Tracking Model [6]

Training a landmark detector

- Same shape initialization i.e., S_0, for all training samples

$$\boxed{S_{init} = S_0}$$

$$Y = \Delta S = S_* - S_{init}$$

$$X = f(I, S_{init})$$

$$R = YX^T(XX^T)^{-1}$$

Training a landmark tracker

- Shape initializations are drawn from the statistics of frame-to-frame landmark displacement statistics

$$\boxed{\begin{gathered} \Delta S_{error} \sim \mathcal{N}(\mu, \Sigma) \\ S_{init} = S_* + \Delta S_{error} \end{gathered}}$$

$$Y = \Delta S_{error}$$

$$X = f(I, S_{init})$$

$$R = YX^T(XX^T)^{-1}$$

Learning a Face Landmark Tracking Model [6]

- Similar to that of a detector, except for the face shape initializations
- Given a previous frame's detected/tracked landmarks, estimate the current frame's face landmarks
- During training, simulate frame-to-frame landmark displacements through learning the displacement statistics from offline face video sequences
- Add shape perturbations to the ground truth landmarks and use the resultant face shapes as initializations

Online Incremental Learning [6]

- Is it possible to train regression model with all possible variations?
- **May be no!**
- Enabling the model to learn incrementally **on the fly** may help in adapting to the changing environment

Online Incremental Learning [6]

- A naive approach to recomputing the regression model:
  - X = {X_{train}; X_{test}} , Y = {Y_{train}; Y_{test}} , compute $R = YX^T(XX^T)^{-1}$
- But, this approach is computationally quite heavy!
- Incremental Linear Least Squares approaches address the above problem

Online Learning Approaches

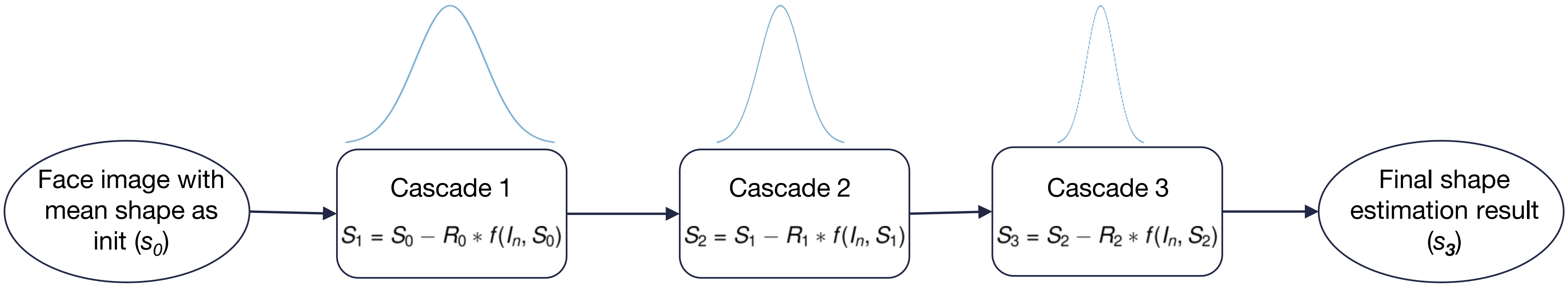
- Online update in CSRs:

Face image with mean shape as init (s_0) → Cascade 1: $S_1 = S_0 - R_0 * f(I_n, S_0)$ → Cascade 2: $S_2 = S_1 - R_1 * f(I_n, S_1)$ → Cascade 3: $S_3 = S_2 - R_2 * f(I_n, S_2)$ → Final shape estimation result (s_3)

A roadblock: Sequential dependency among the stages!

Online Learning Approaches

- Parallel Supervised Descent Method (Parallel SDM):
    - Treats all stages independently by characterizing their ***error correction statistics***



Still need to compute features for multiple initializations!

Online Learning Approaches

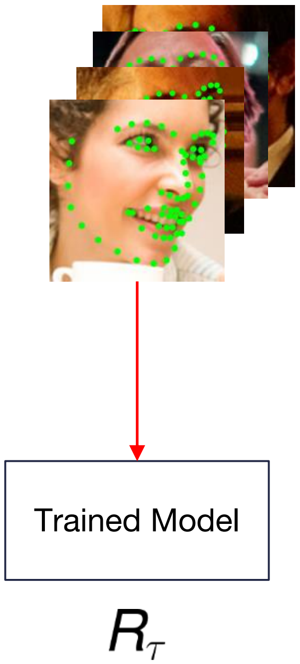
- Incremental Continuous Cascaded Regression (iCCR):
  - Approximate feature space using first-order Taylor series expansion:

$$f(\mathbf{I}_j, \mathbf{s}_j^* + \delta\mathbf{s}) \approx f(\mathbf{I}_j, \mathbf{s}_j^*) + \mathbf{J}_j^* \delta\mathbf{s},$$

where J_j^* is the Jacobian matrix of feature descriptors at ground truth landmarks

Online Learning Approaches

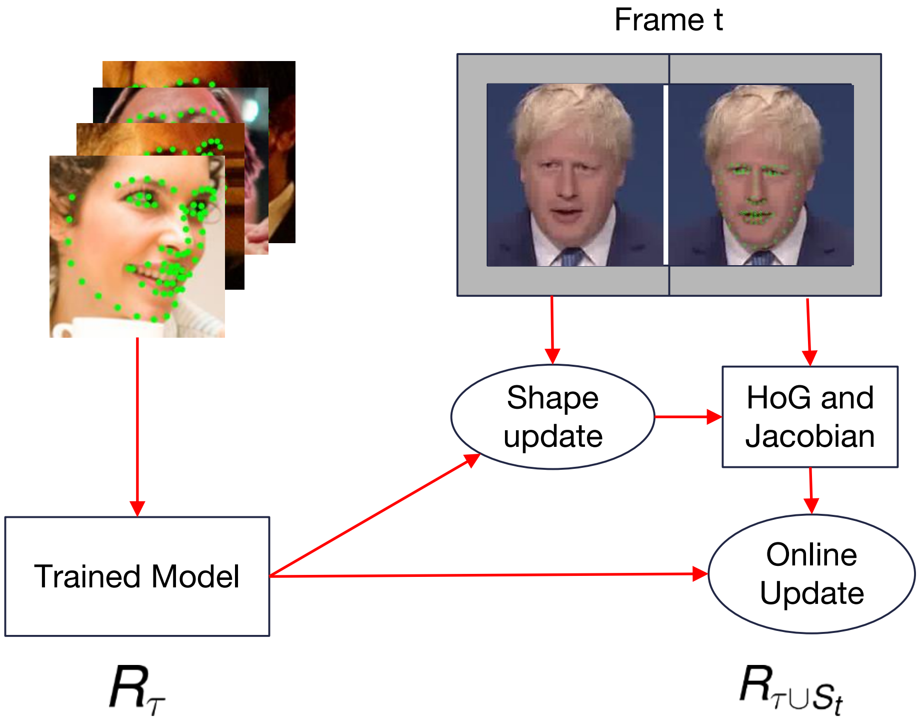
- Incremental Continuous Cascaded Regression (iCCR):



Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Online Learning Approaches

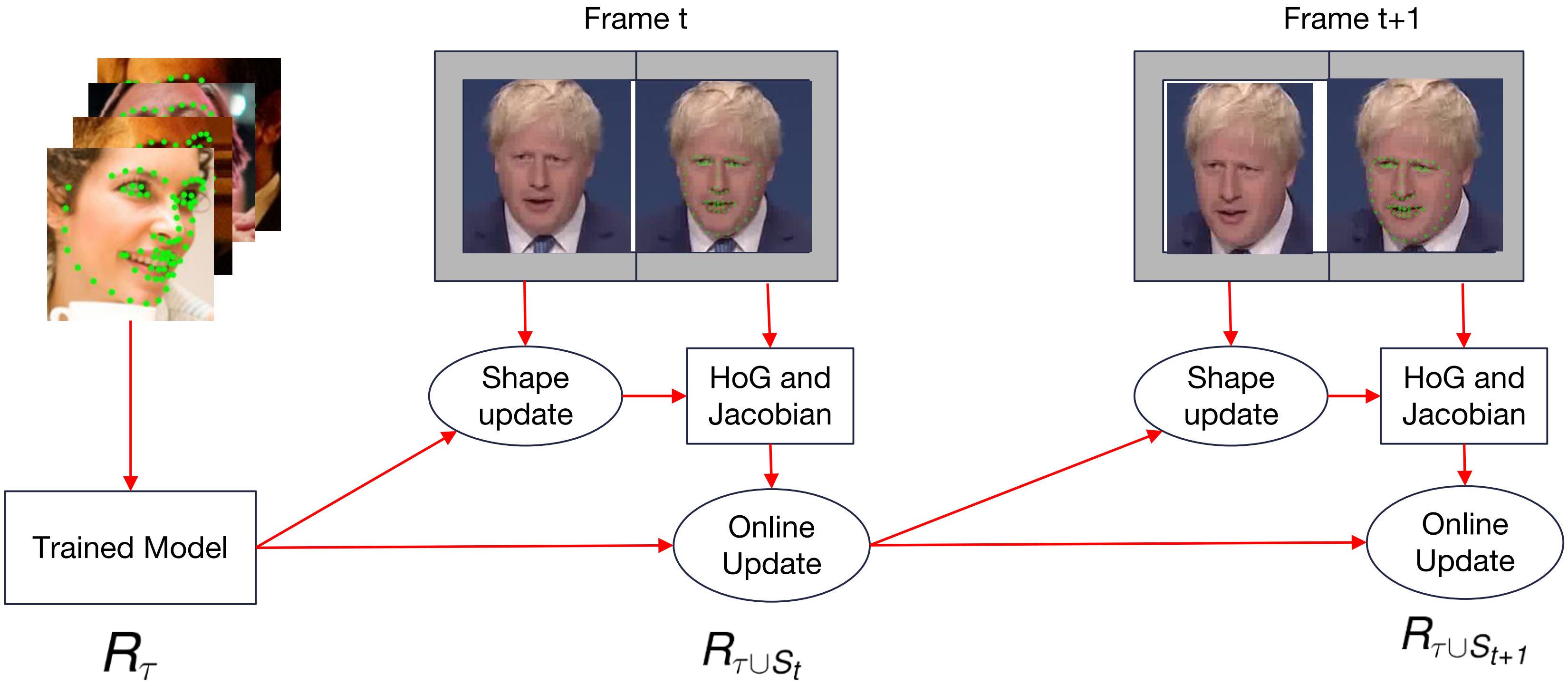
- Incremental Continuous Cascaded Regression (iCCR):



Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Online Learning Approaches

- Incremental Continuous Cascaded Regression (iCCR):



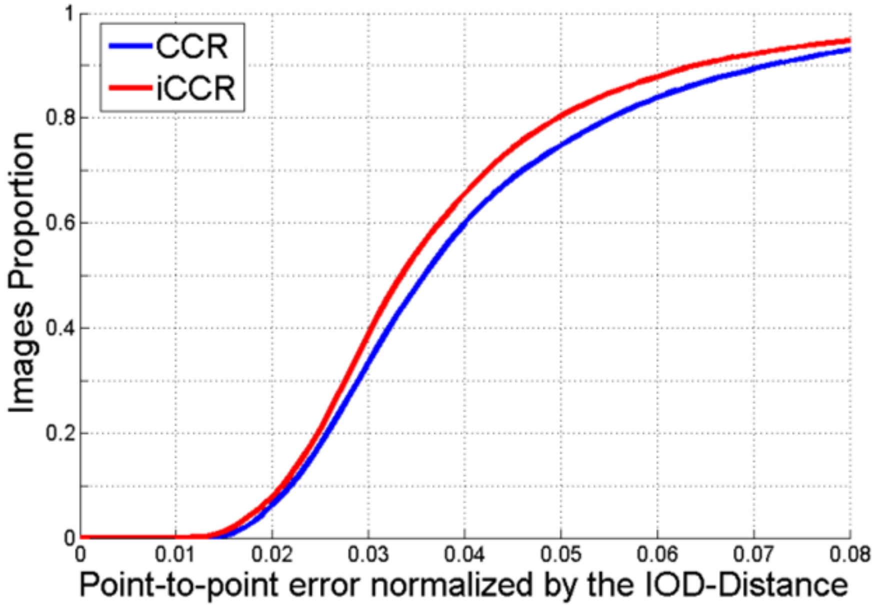
Courtesy: https://ibug.doc.ic.ac.uk/resources/300-VW/

Sample results with / without online learning

AUC values for three different categories of 300 VW benchmark

	CCR	iCCR
Category 1	0.4807	0.5171
Category 2	0.4680	0.5232
Category 3	0.3198	0.4044

CED curves for the most challenging category of 300 VW benchmark



Courtesy: Sánchez-Lozano, Enrique & Martinez, Brais & Tzimiropoulos, Georgios & Valstar, Michel. (2016). Cascaded Continuous Regression for Real-time Incremental Face Tracking, https://arxiv.org/pdf/1608.01137.pdf

In summary, notes for embedded

- Feature representation:
  - Use computationally less intensive descriptors: HoG, SIFT, LBF
- Target representation:
  - Use a compact representation of (x,y) coordinates such as PDM and SPDM
  - This reduces the cost of an inference call by reducing the size of the regression weight matrix
- Incremental online learning:
  - Use feature space approximation to support real-time online update
- Inexpensive model-free tracking methods for improving the accuracy:
  - Employ model-free trackers such as Kalman to get better shape initializations for landmark tracking

References

1. G. G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A comprehensive performance evaluation of deformable face tracking ”in-the-wild”. CoRR, abs/1603.06015, 2016.
2. Xiong, X., la Torre, F.D.: Supervised descent method for solving nonlinear least squares problems in computer vision, arXiv abs/1405.0601, 2014.
3. Xiong, X., De la Torre, F.: Global supervised descent method. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. , 20152664– 2673, 2015.
4. Smith, B. M., and C. R. Dyer. Efficient Branching Cascaded Regression for Face Alignment Under Significant Head Rotation. ArXiv e-prints 1611.01584, 2016.
5. G. Trigeorgis, P. Snape, E. Antonakos, M. A. Nicolaou, and S. Zafeiriou. Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2016.
6. E. Sánchez-Lozano, B. Martinez, G. Tzimiropoulos, M. Valstar, Cascaded continuous regression for real-time incremental face tracking, European Conference on Computer Vision - ECCV 2016 Part VIII, pp. 645-661, 2016.1611.01584, 2016.
7. S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
8. Gu J., Yang X., De Mello S., Kautz J., Dynamic Facial Analysis: from Bayesian Filtering to Recurrent 385 Neural Network, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1531-1540, 2017.
9. A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, "Incremental face alignment in the wild", *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1859-1866, 2014.
10. Taigman, Yaniv & Yang, Ming & Ranzato, Marc'Aurelio & Wolf, Lior. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 10.1109/CVPR.2014.220.

About PathPartner

PathPartner is a top-notch design engineering services company and works with several semiconductor companies, OEMs and ODMs in embedded media-centric devices and intelligent systems.

- Incorporated in July 2006; HQ in Bangalore, India
- R&D Centers: Bangalore, India, and California, USA
- Marketing representatives in USA, Europe, Japan and India
- PathPartner is a member of Embedded Vision Alliance and partner of various semiconductor companies

- Present company strength is ~280
- Quality : ISO 9001:2015, 27001:2013
- R&D Workforce : >10%

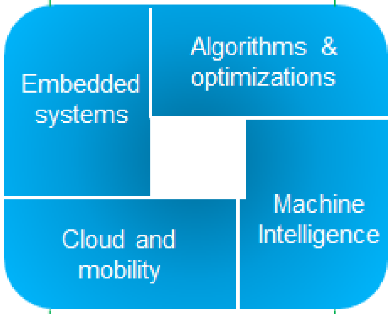
- Semiconductor Companies
- OEMs and ODMs

+91 80 6772 2000 |
+1 408 242 7411 |
+81 9055466819

sales.india@pathpartnertech.com |
sales.usa@pathpartnertech.com |
sales.japan@pathpartnertech.com

Hardware design
FPGA design
System software
Testing and validation

Computational imaging
Multimedia algorithms
Screen-casting algorithms
Connectivity protocols



Cloud integration
Application development

Computer vision
Deep neural networks

Thanks!!

Questions?