Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment

Jie Zhang, Shiguang Shan, Meina Kan, Xilin Chen
Institute of Computing Technology, Chinese Academy of Sciences

September 8, 2014

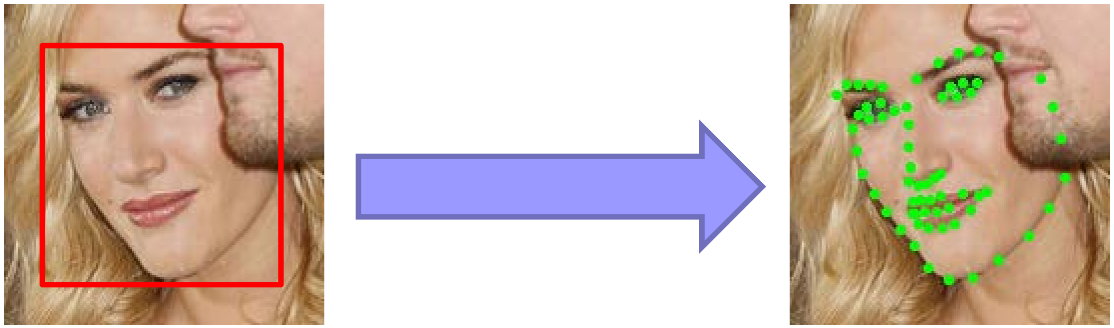
中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

Outline

- Problem
- Related Work
- Motivation
- Method
  - Global SAN
  - Local SAN
  - Coarse-to-fine strategy
- Experiments
- Conclusion

Problem

- Face Alignment：
  - Predict facial landmarks from the face region



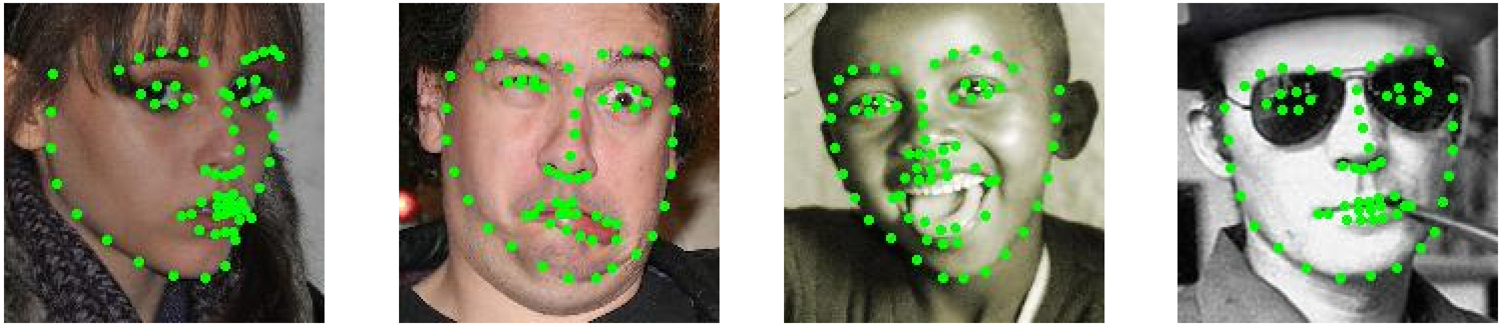
- Application
  - Face recognition
  - Expression recognition
  - Face animation …

Institute of Computing Technology

Challenges

Institute of Computing Technology

- Appearance -> Shape: a complex mapping
- Large appearance & shape variations
  - Head pose
  - Expressions
  - Illumination
  - Partial occlusion



Related Work

- ASM & AAM [Cootes'95; Gu'08; Cootes'01; Matthews'04]
  - Sensitive to initial shapes
  - Sensitive to noise
  - Hard to cover complex variations
- DCNN [Sun'13; Toshev'14]
  - Fragile to partial occlusions
- Shape regression model
  - CPR,ESR,RCPR [Dollar'10; Cao'12; Burgos-Artizzu'13]
  - DRMF [Asthana'13]
  - SDM [Xiong'13]

Motivation

Cascade shape regression models for face alignment

$$\Delta \boldsymbol{S_j} = \boldsymbol{H_j}(\emptyset(\boldsymbol{I}, \boldsymbol{S_{j-1}}))$$

- **Global regression for better initialization $\boldsymbol{S_0}$**

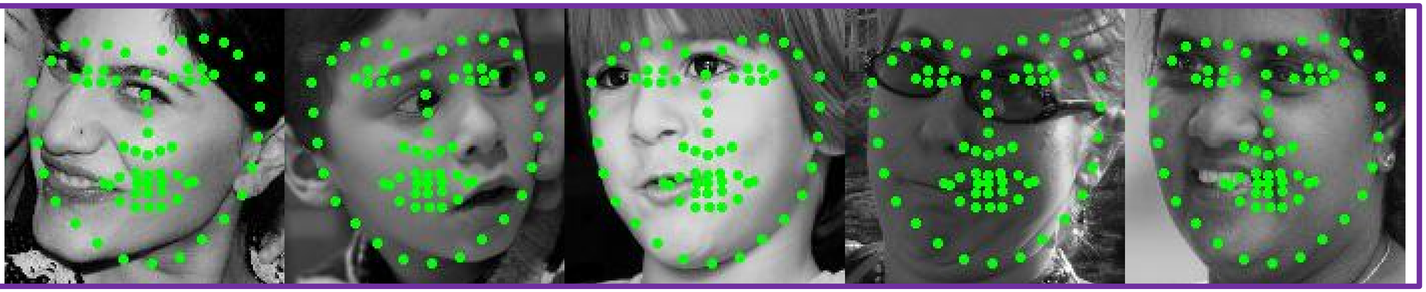
Motivation

Cascade shape regression models for face alignment

$$\Delta S_j = H_j(\emptyset(I, S_{j-1}))$$

- **Global regression for better initialization S_0**

Mean Shape



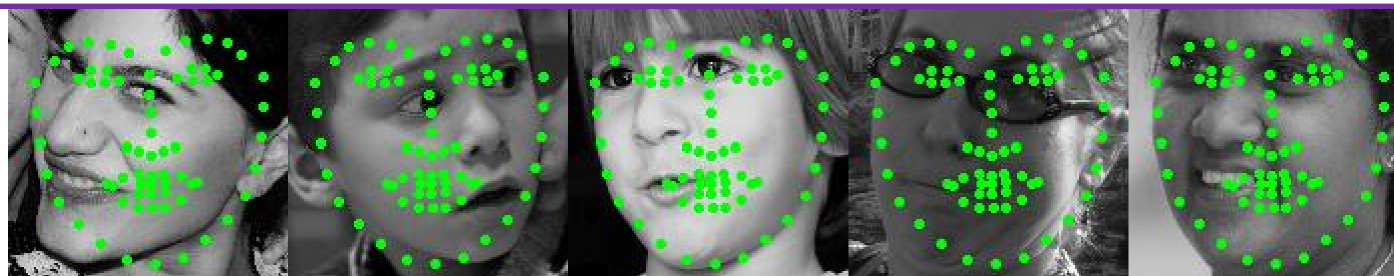
Motivation

Cascade shape regression models for face alignment

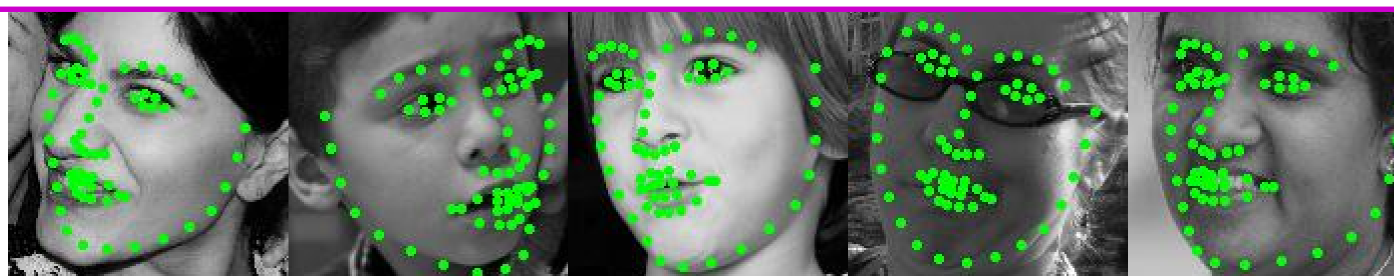
$$\Delta \boldsymbol{S}_j = \boldsymbol{H}_j(\emptyset(\boldsymbol{I}, \boldsymbol{S}_{j-1}))$$

- **Global regression for better initialization S_0**

Mean Shape



Global Regression

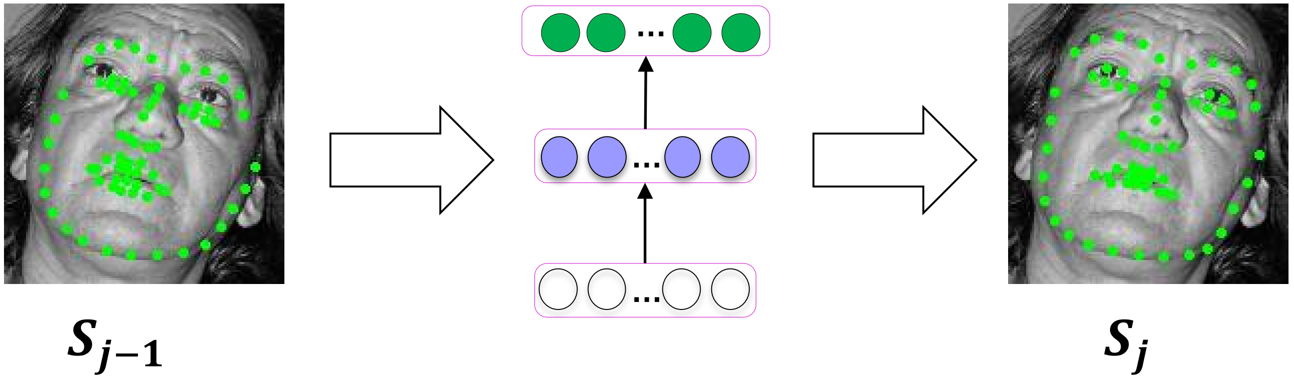


Motivation

Cascade shape regression models for face alignment

$$\Delta S_j = H_j(\emptyset(I, S_{j-1}))$$

- **Global regression** **for better initialization S_0**
- **Deep networks for** **nonlinear regression function H_j**



Motivation

Cascade shape regression models for face alignment

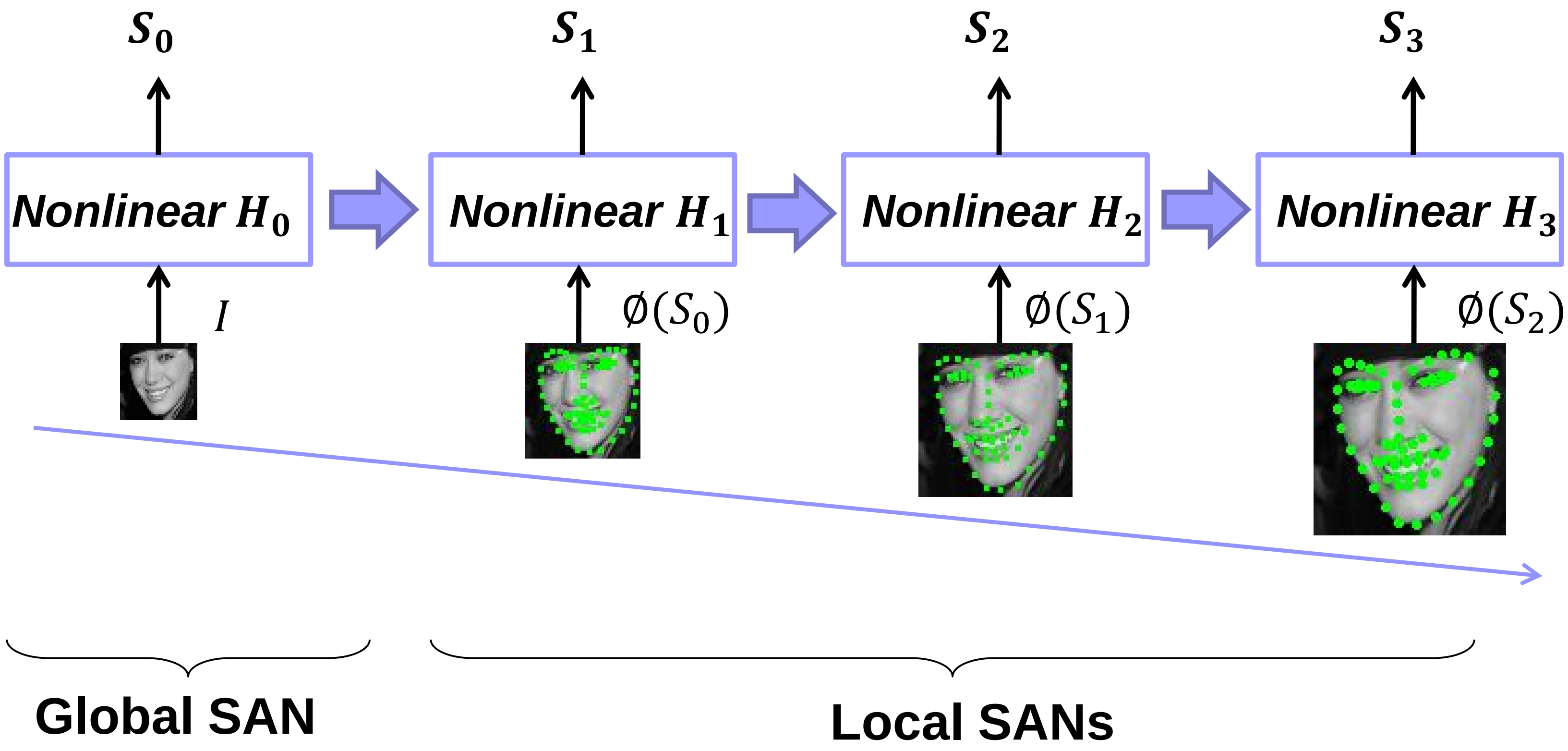
$$\Delta S_j = H_j(\emptyset(I, S_{j-1}))$$

- Global regression for better initialization S_0
- Deep networks for nonlinear regression function H_j
- Cascade H_j in a coarse-to-fine architecture

Coarse-to-fine Cascade

Our Method(1/7)

- Schema of Coarse-to-Fine AE Networks



SAN: Stacked Auto-encoder Network

Our Method(2/7)

Institute of Computing Technology

- Pipeline

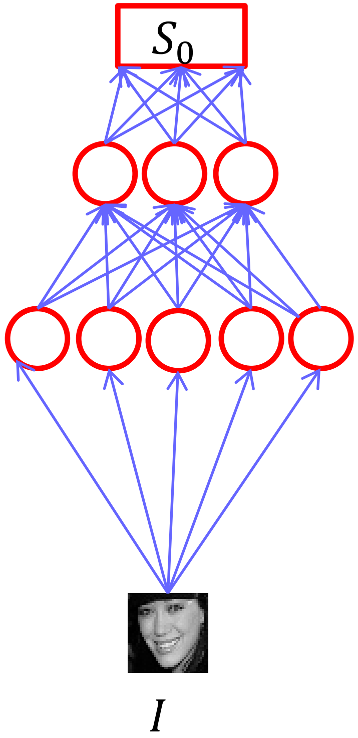
Our Method(2/7)

- Pipeline

I

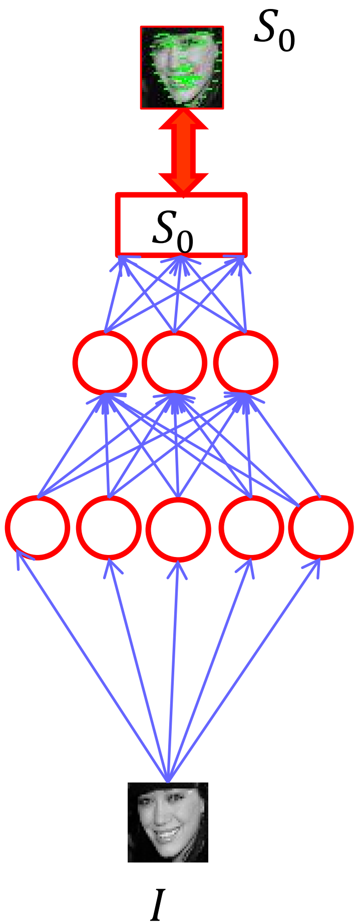
Our Method(2/7)

- Pipeline



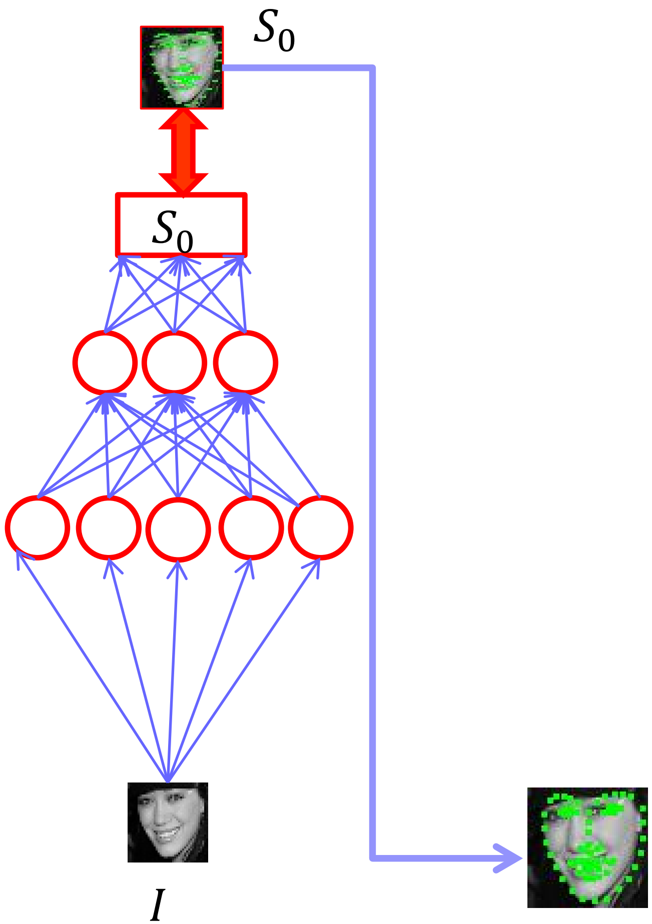
Our Method(2/7)

- Pipeline



Our Method(2/7)

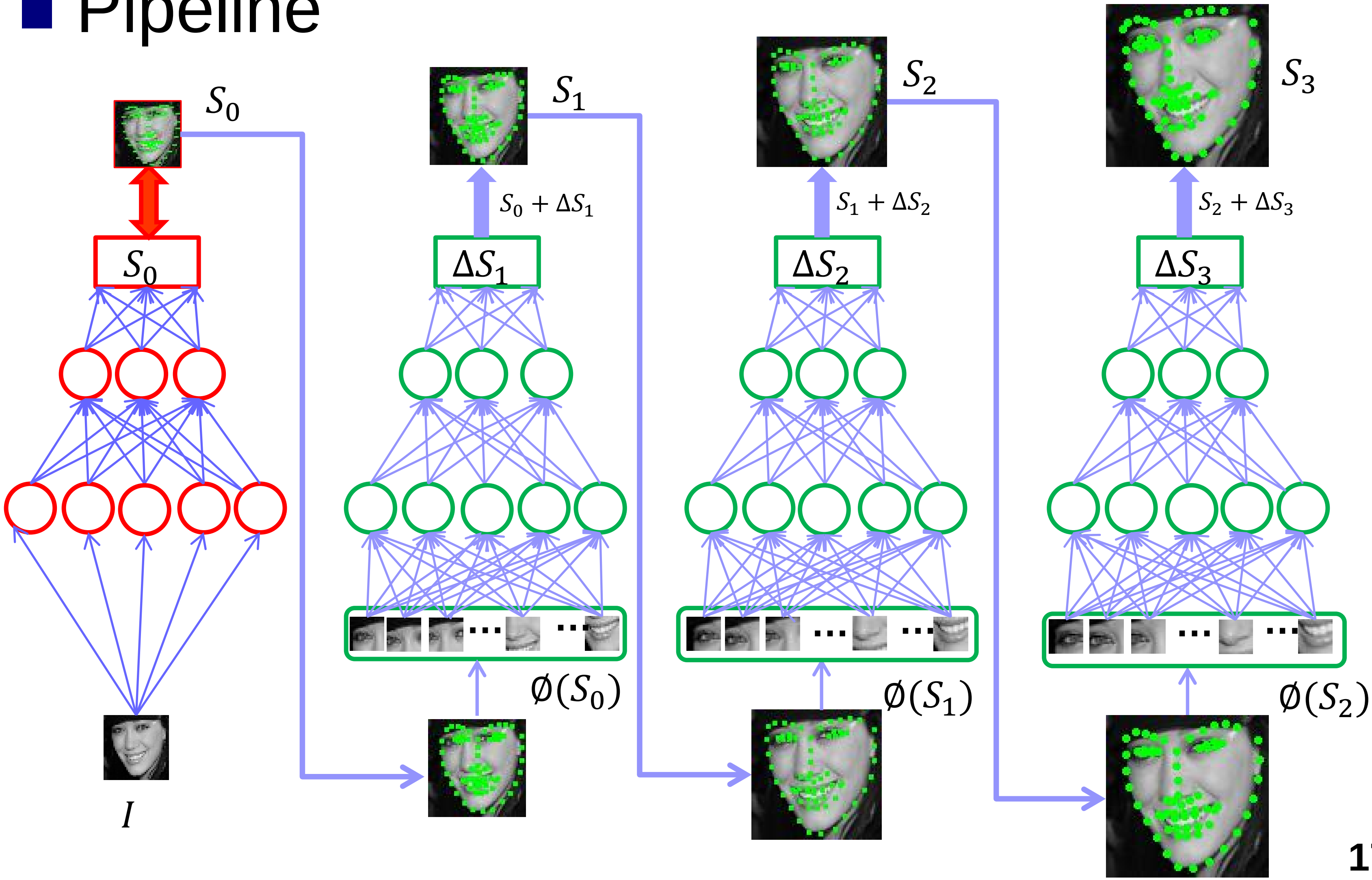
- Pipeline



Our Method(2/7)

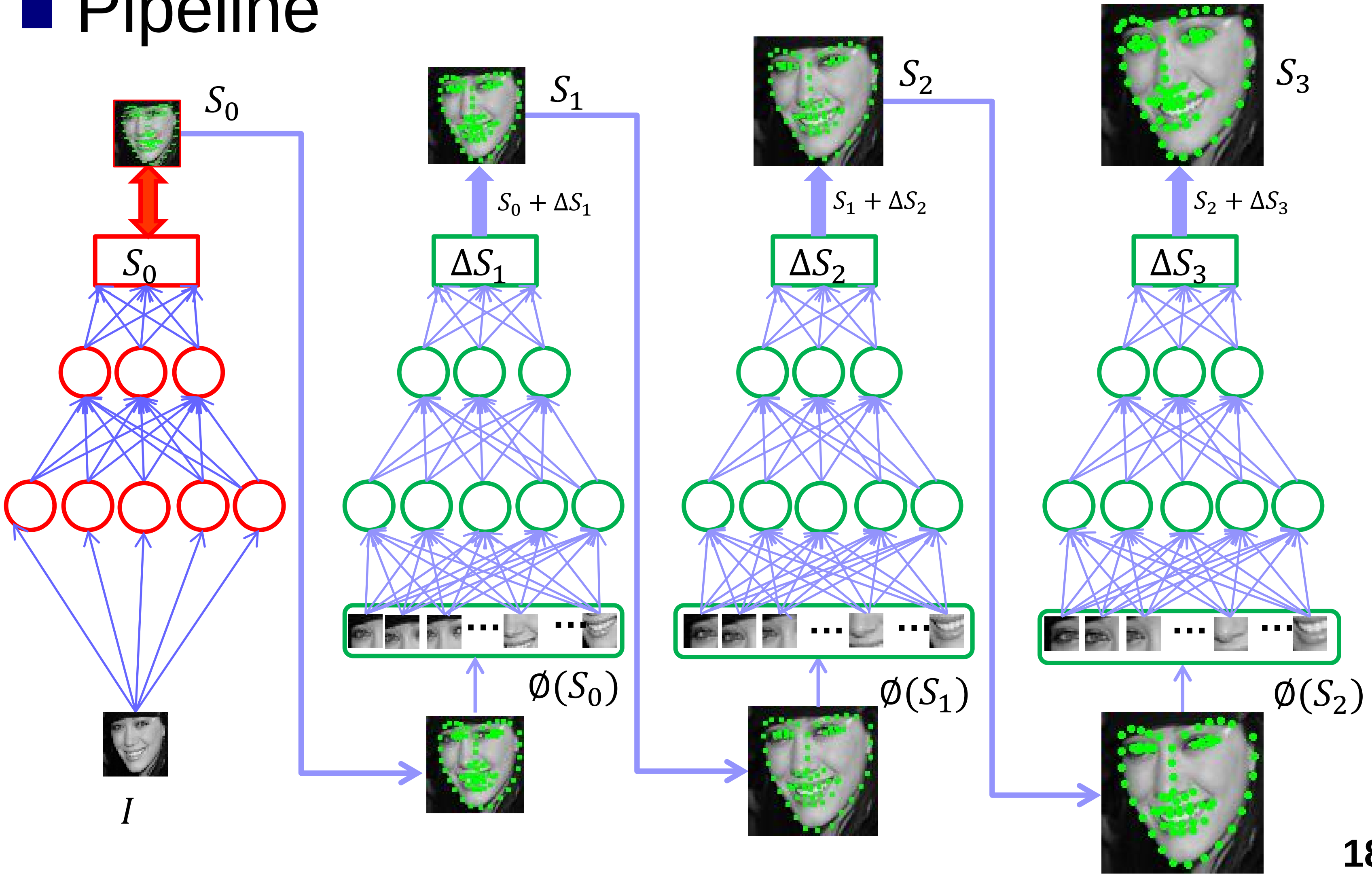
Institute of Computing Technology

- Pipeline



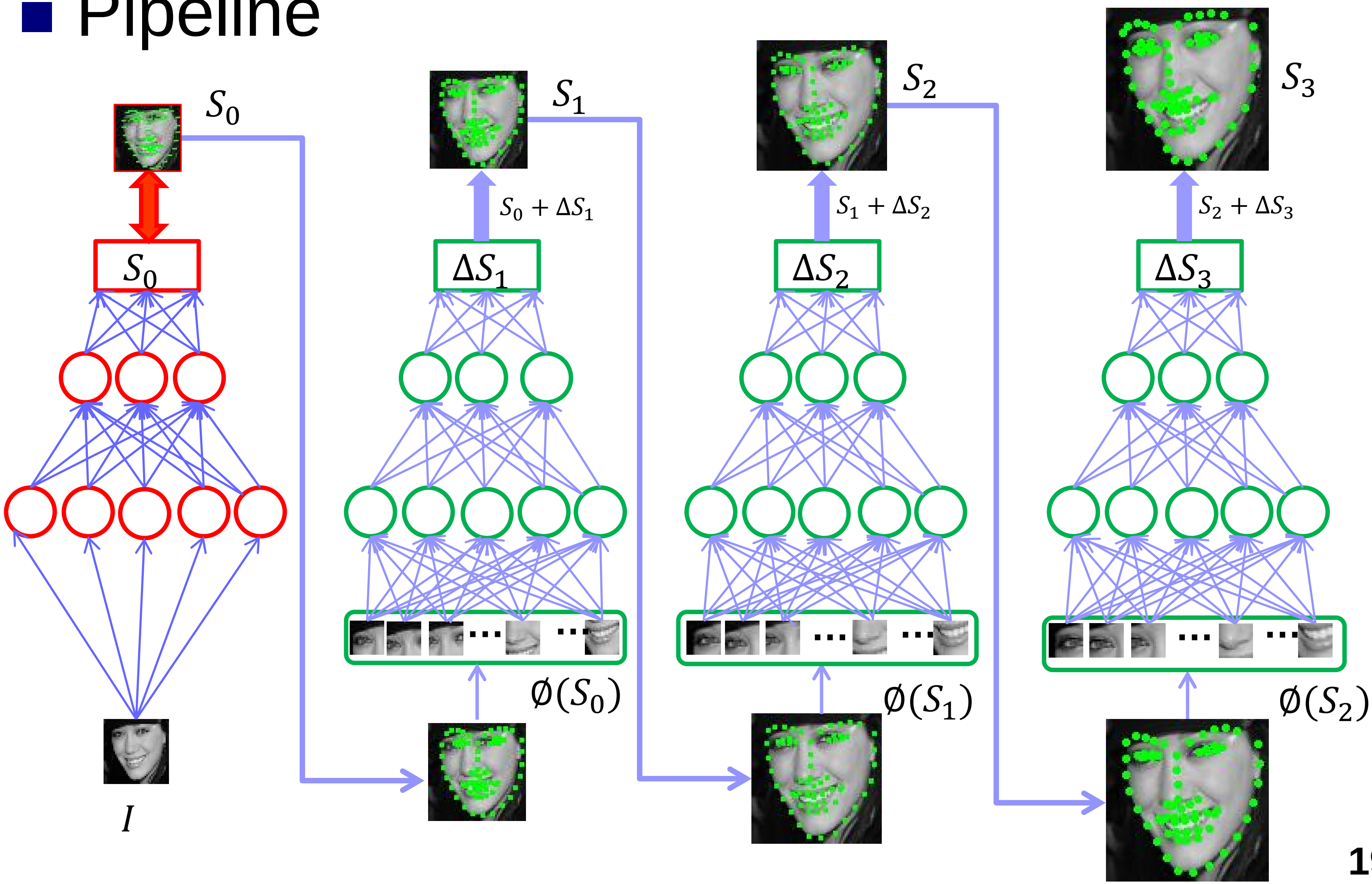
Our Method(3/7)

- Pipeline



Our Method(3/7)

- Pipeline

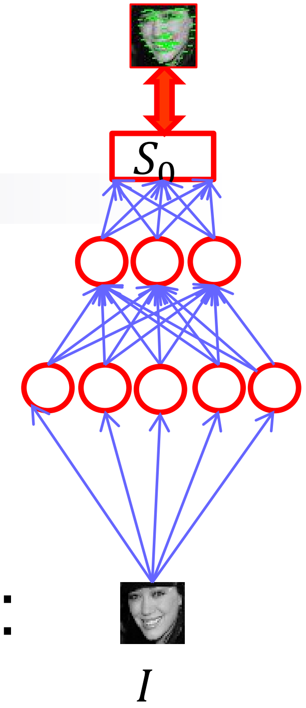


Our Method(4/7)

- Global SAN:
  - Mapping H_0 from image I to shape S.

$$H_0 : S \leftarrow I$$

  - Model H_0 as a Stacked Auto-encoder:

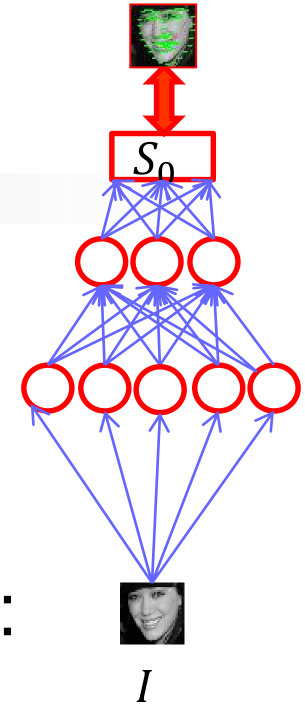


$$H_0^* = \arg\min_{H_0} \|S - f_k(f_{k-1}(\ldots f_1(I)))\|_2^2 + \alpha \sum_{i=1}^{k} \|W_i\|_F^2$$

$$f_i(a_{i-1}) = \sigma(W_i a_{i-1} + b_i) \triangleq a_i, i = 1, \ldots, k-1$$

$$f_k(a_{k-1}) = W_k a_{k-1} + b_k \triangleq S_0$$

Our Method(4/7)



- Global SAN:
  - Mapping H_0 from image I to shape S.

$$H_0 : S \leftarrow I$$

  - Model H_0 as a Stacked Auto-encoder:

$$H_0^* = \arg\min_{H_0} \|S - f_k(f_{k-1}(\dots f_1(I)))\|_2^2 + \alpha \sum_{i=1}^{k} \|W_i\|_F^2$$

Regression — Regularization

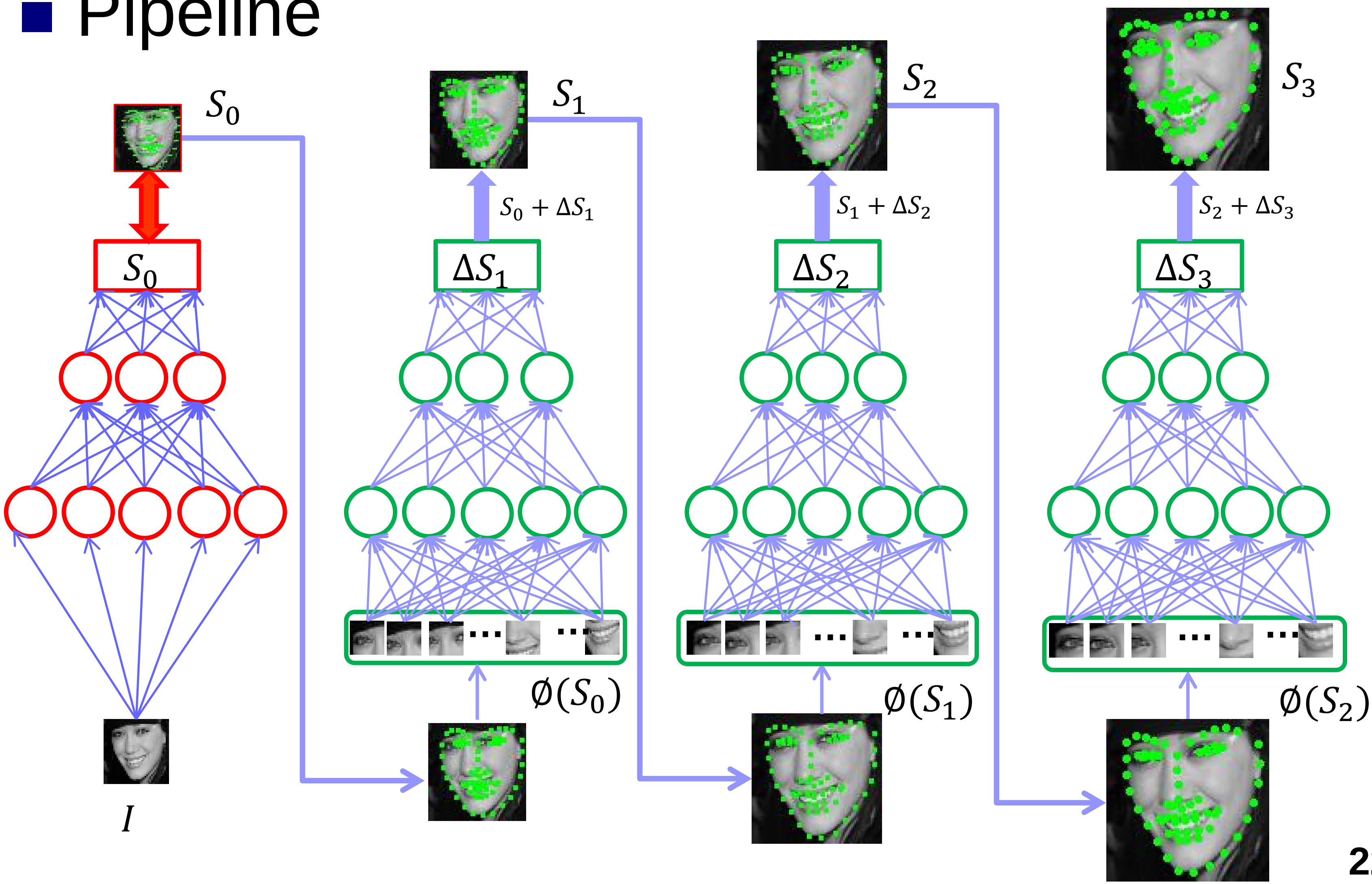
$$f_i(a_{i-1}) = \sigma(W_i a_{i-1} + b_i) \triangleq a_i, i = 1, \dots, k-1$$

$$f_k(a_{k-1}) = W_k a_{k-1} + b_k \triangleq S_0$$

Institute of Computing Technology

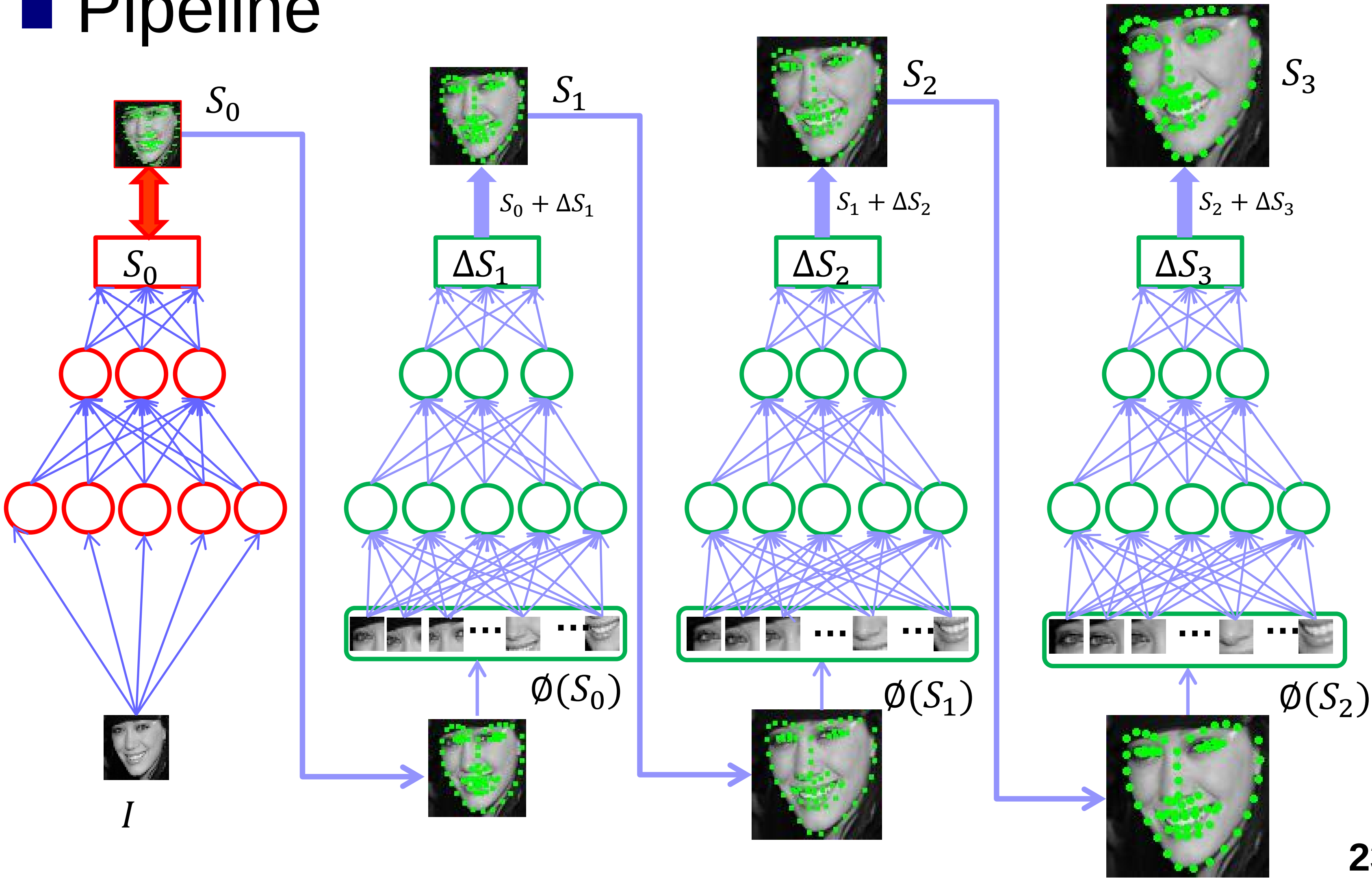
Our Method(5/7)

- Pipeline



Our Method(5/7)

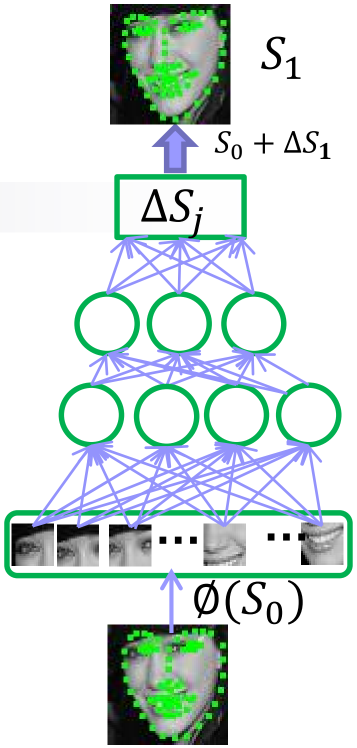
- Pipeline



Our Method(6/7)

Institute of Computing Technology

- Local SAN:
  - Initialize shape S_0 from global SAN.
  - Refine the shape with local feature
    - predict shape deviation with AE



$$H_1^* = \arg\min_{H_1} \left\| \Delta S_1 - h_k^1 \left(\dots h_1^1(\emptyset(S_0)) \right) \right\|_2^2 + \alpha \sum_{i=1}^{k} \left\| W_i^1 \right\|_F^2$$

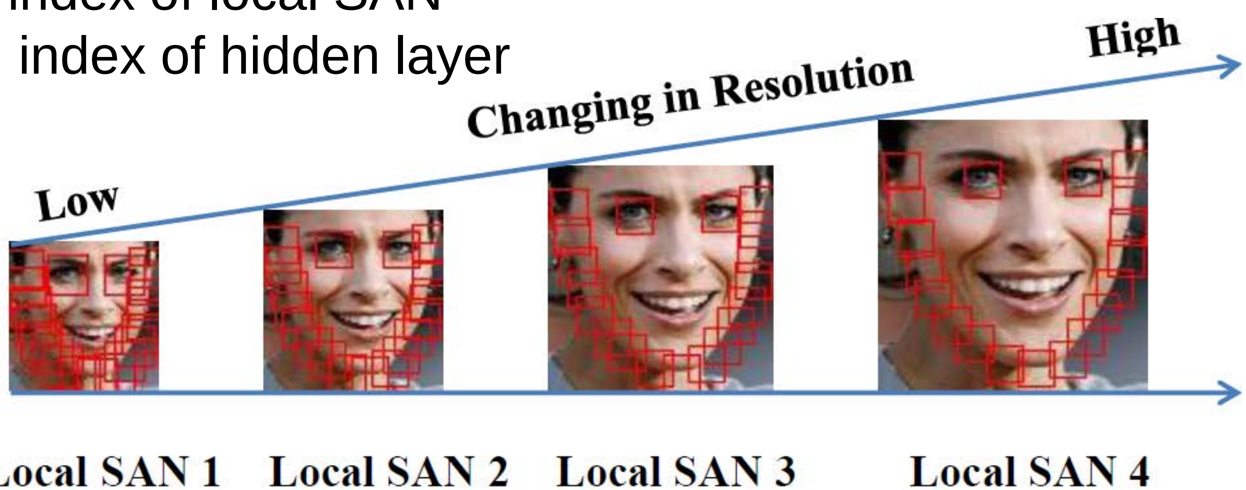
$$\Delta S_1 = S - S_0$$

Our Method(7/7)

- Coarse-to-fine Cascade:

$$H_j^* = \arg\min_{H_j} \left\| \Delta S_j - h_k^j \left(\dots h_1^j \left(\phi \left(S_{j-1} \right) \right) \right) \right\|_2^2 + \alpha \sum_{i=1}^{k} \| W_i^j \|_F^2$$

j: index of local SAN
k: index of hidden layer



S_0 S_1 S_2 S_3

Large Search Region/Step ⟶ **Tiny Search Region/Step**

Experiments(1/8)

- Datasets
  - **XM2VTS** [Messer'99]
    - 2360 face images collected over 4 sessions under the controlled settings
  - **LFPW** [Belhumeur'11]
    - 1132 training images and 300 test images collected from wild condition
  - **HELEN** [Le'12]
    - 2330 high-resolution face images collected from the wild, 2000 images for training and 330 images for test
  - **AFW** [Zhu'12]
    - 205 images with 468 faces collected from the wild

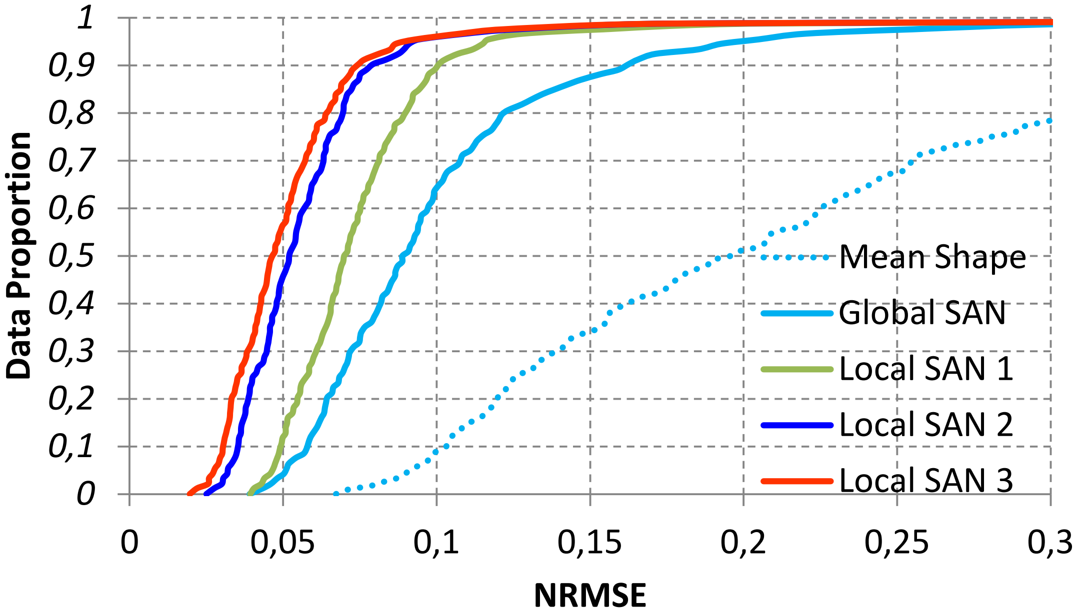
Experiments(2/8)

- Evaluation of Successive SANs

Performance gain of each SAN
(Conduct on LFPW)

Experiments(2/8)

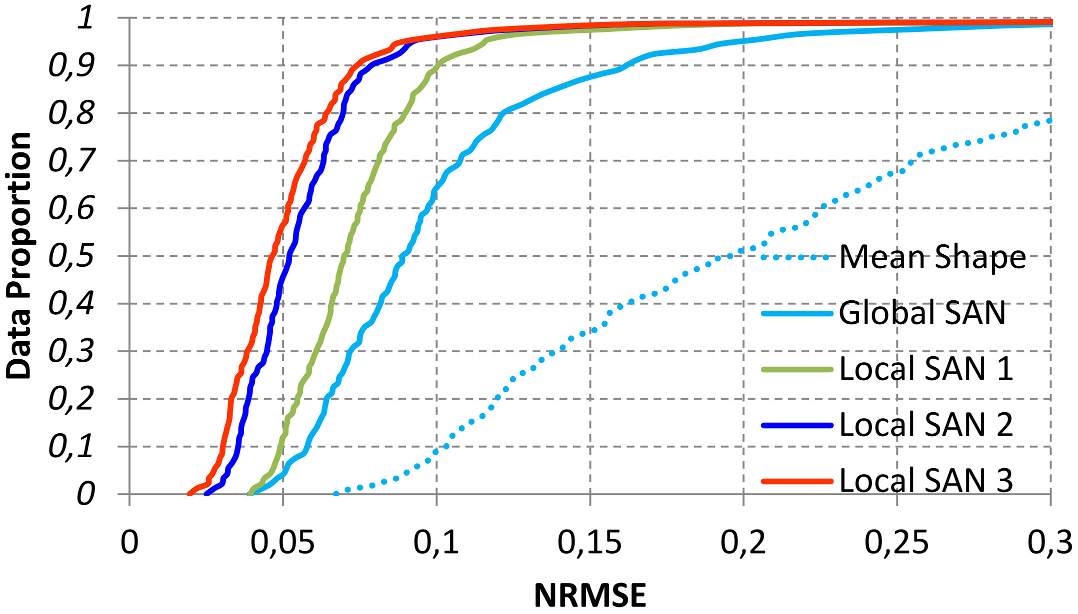
- Evaluation of Successive SANs



Performance gain of each SAN (Conduct on LFPW)

Experiments(2/8)

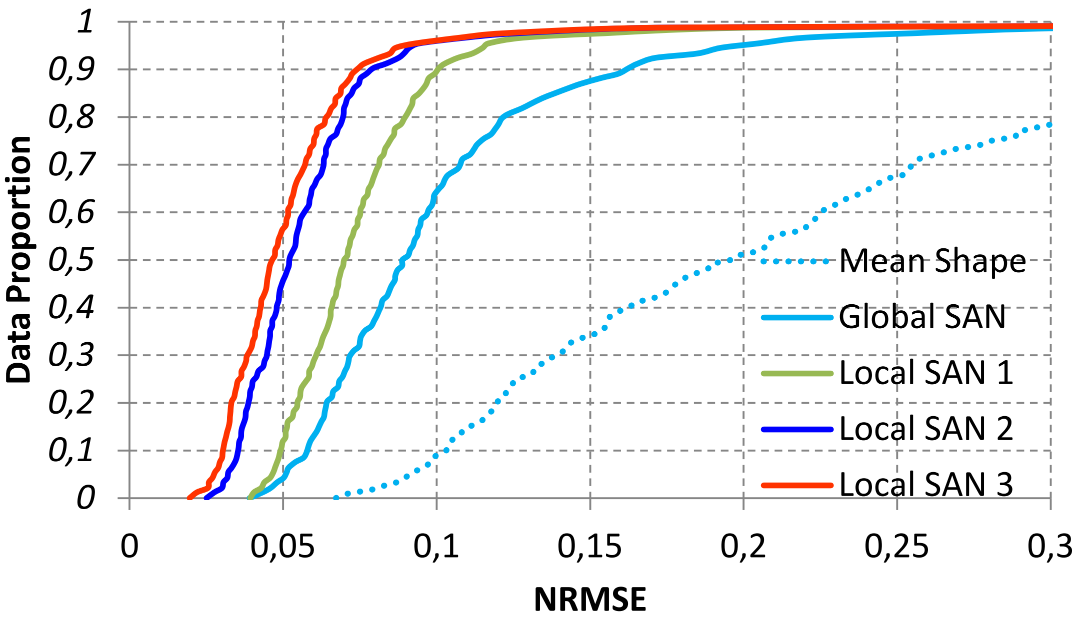
- Evaluation of Successive SANs



Performance gain of each SAN (Conduct on LFPW)

Experiments(2/8)

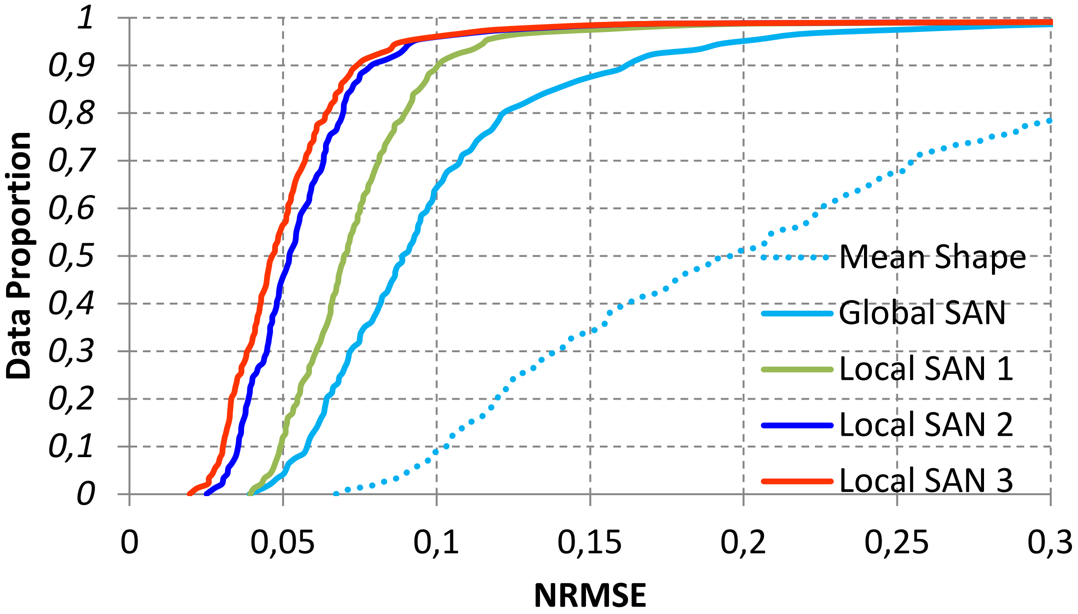
- Evaluation of Successive SANs



Performance gain of each SAN (Conduct on LFPW)

Experiments(2/8)

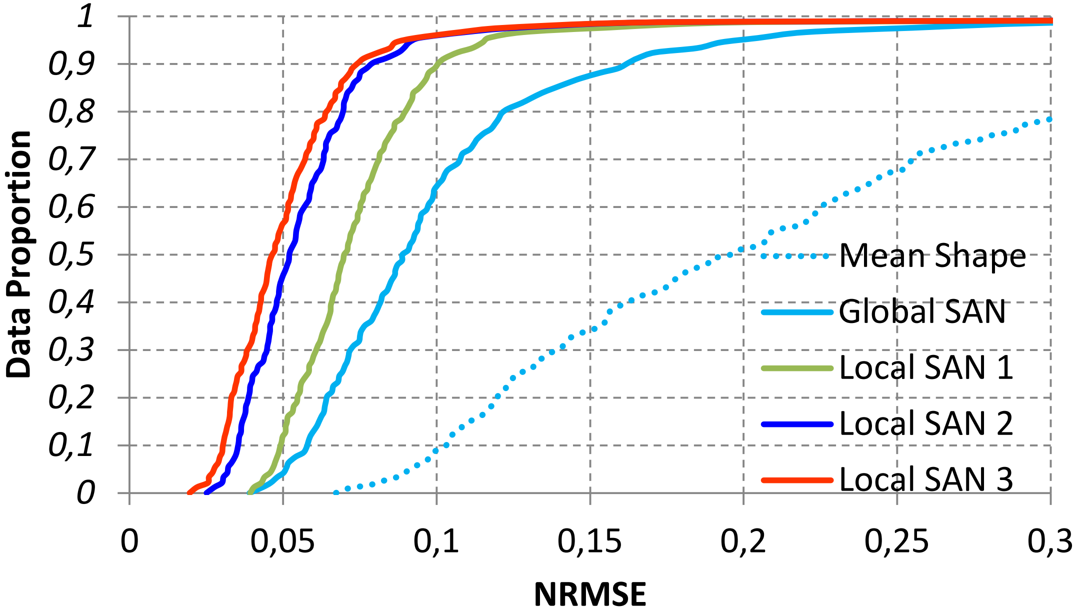
- Evaluation of Successive SANs

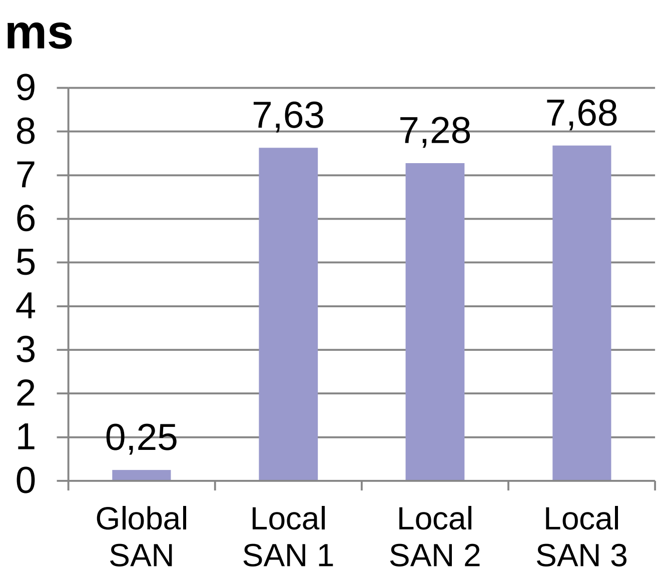


Performance gain of each SAN (Conduct on LFPW)

Experiments(2/8)

- Evaluation of Successive SANs





Performance gain of each SAN (Conduct on LFPW)

Run Time (ms)

Experiments(3/8)

- Comparative Methods
  - Local Models with Regression Fitting
    - SDM [Xiong'13]
    - DRMF [Asthana'13]
  - Tree-structured Models
    - Zhu et al. [Zhu'12]
    - Yu et al. [Yu'13]
  - Deep Model
    - DCNN [Sun'13]

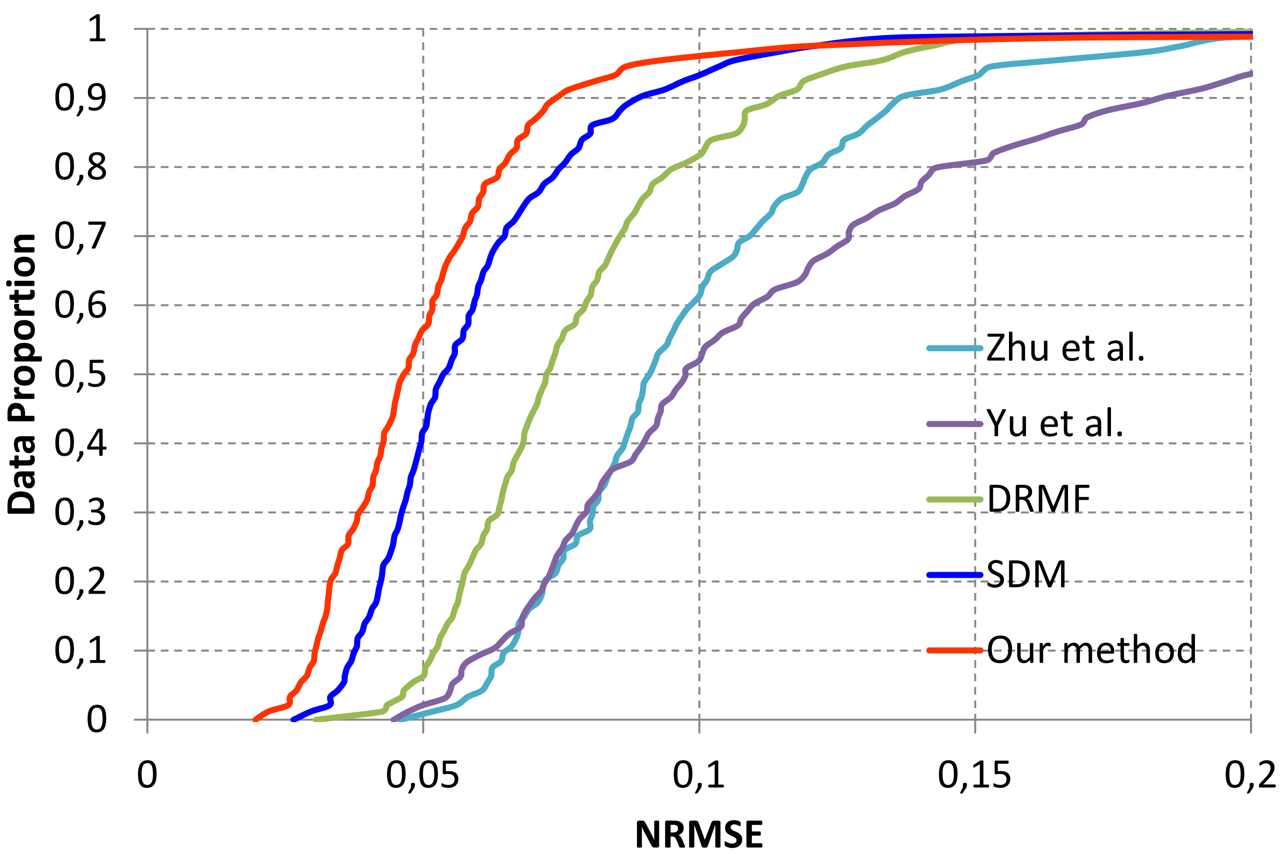
Experimental Result(4/8)

- Performance comparisons on HELEN

Institute of Computing Technology

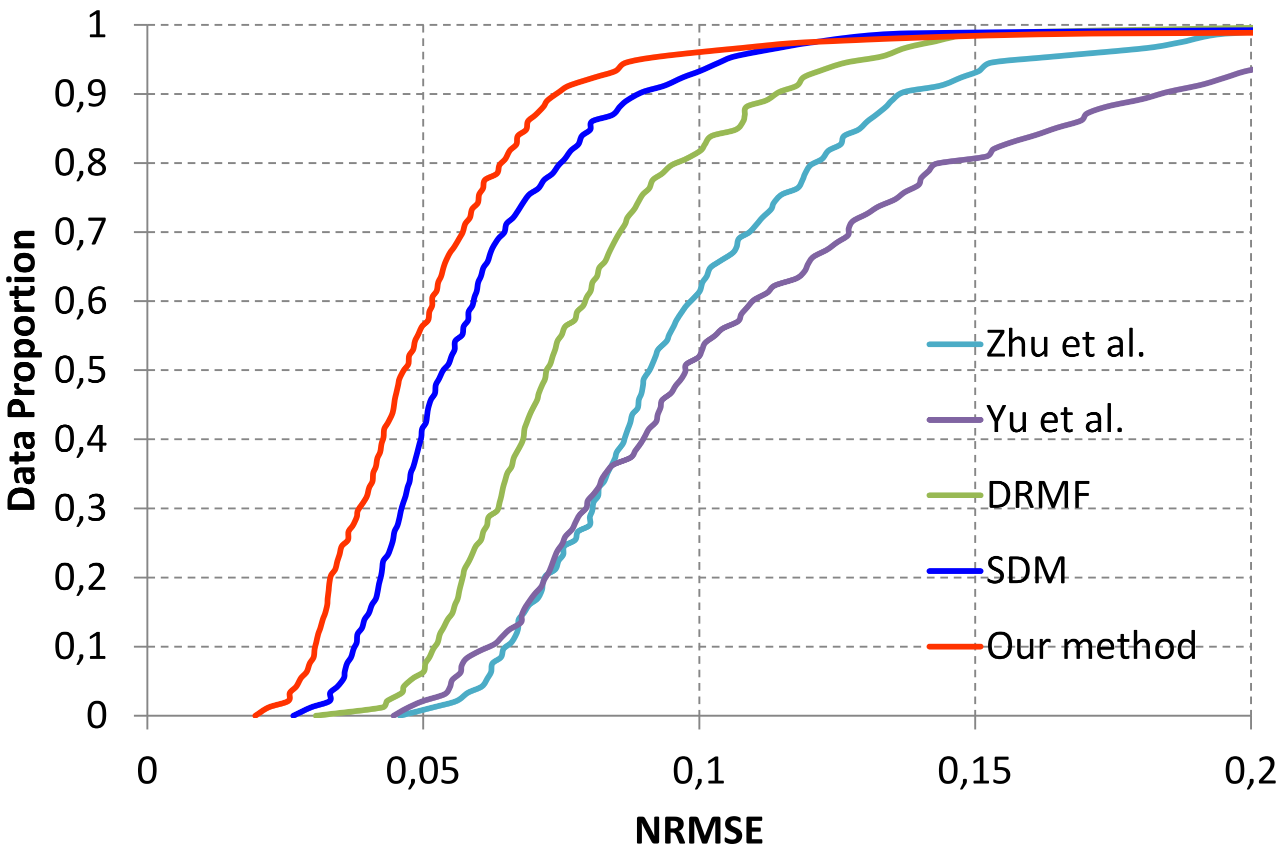
Experimental Result(4/8)

- Performance comparisons on HELEN



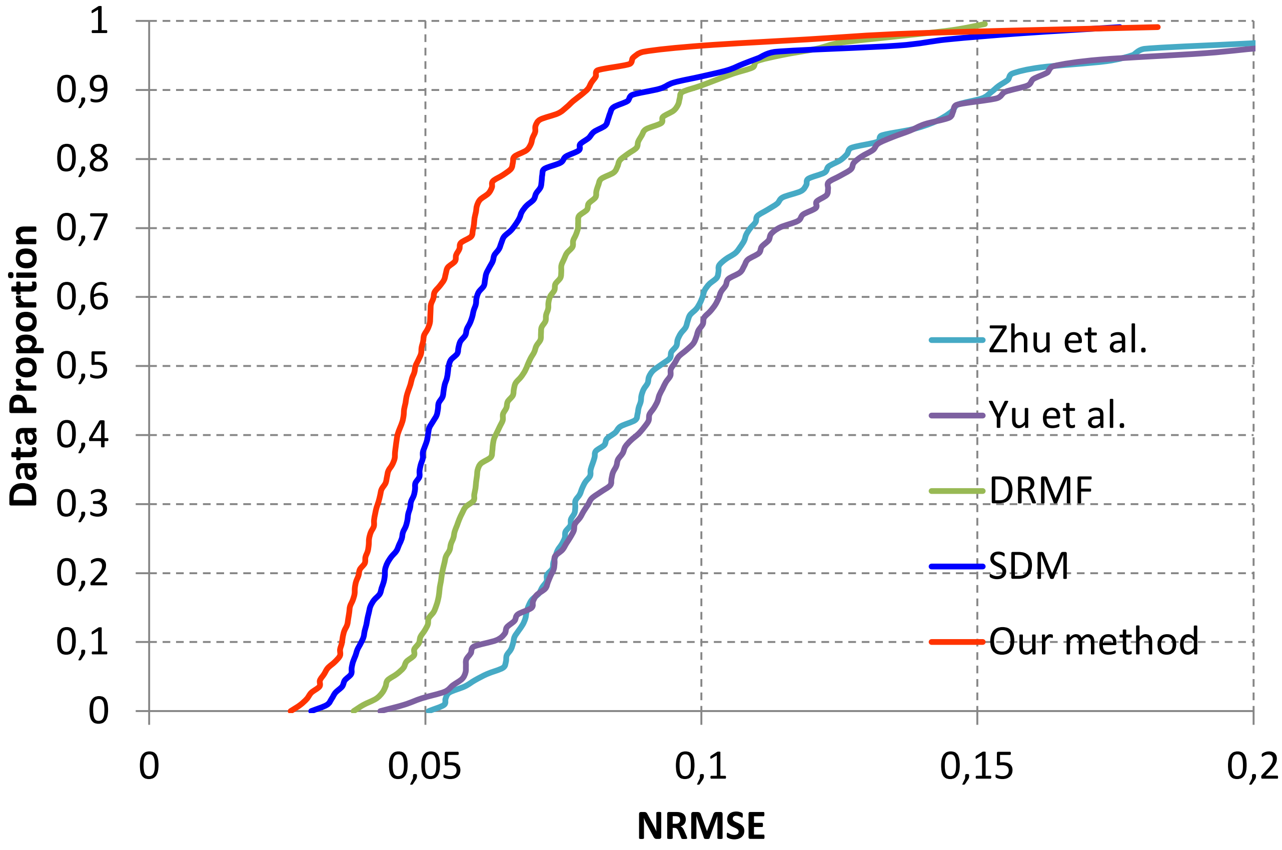
Experimental Result(4/8)

- Performance comparisons on HELEN



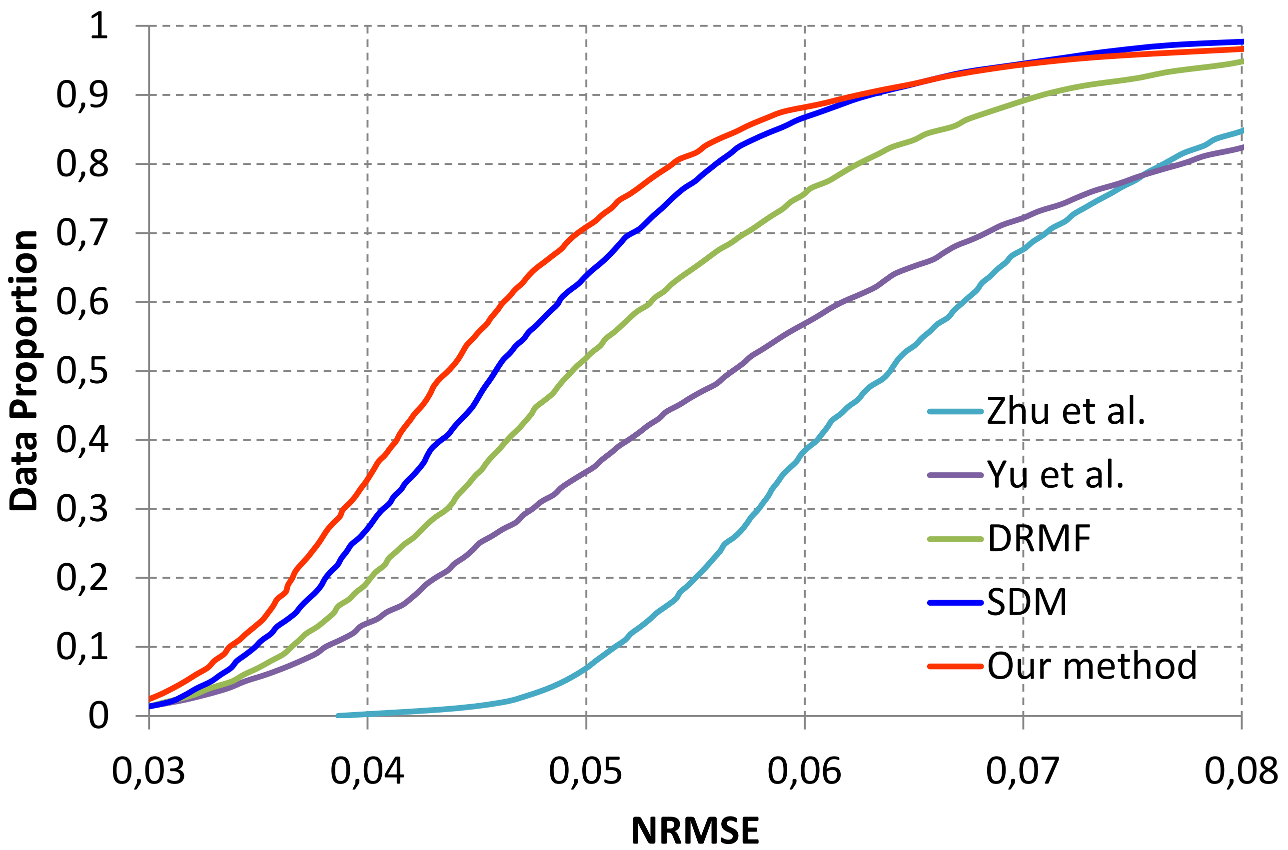
Experimental Result(5/8)

- Performance comparisons on LFPW



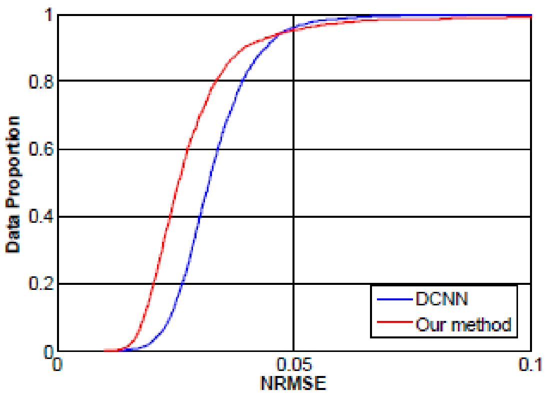
Experimental Result(6/8)

- Performance comparisons on XM2VTS

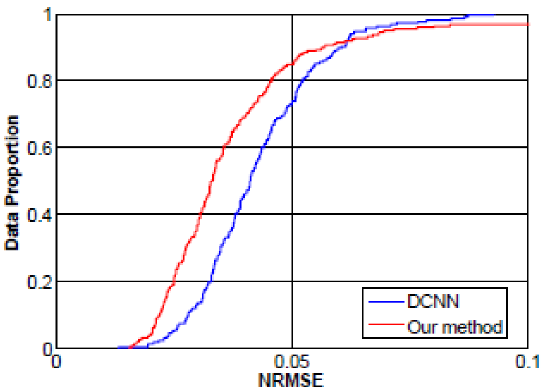


Experimental Result(7/8)

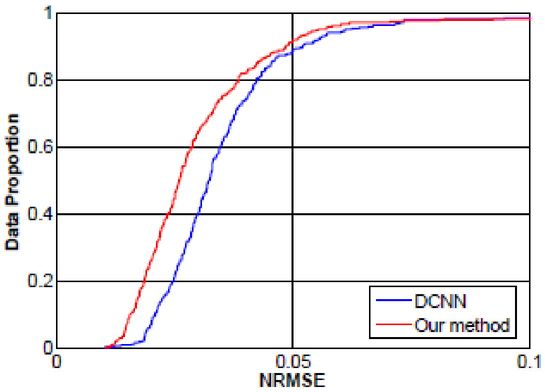
- Comparisons with DCNN



XM2VTS



LFPW



HELEN

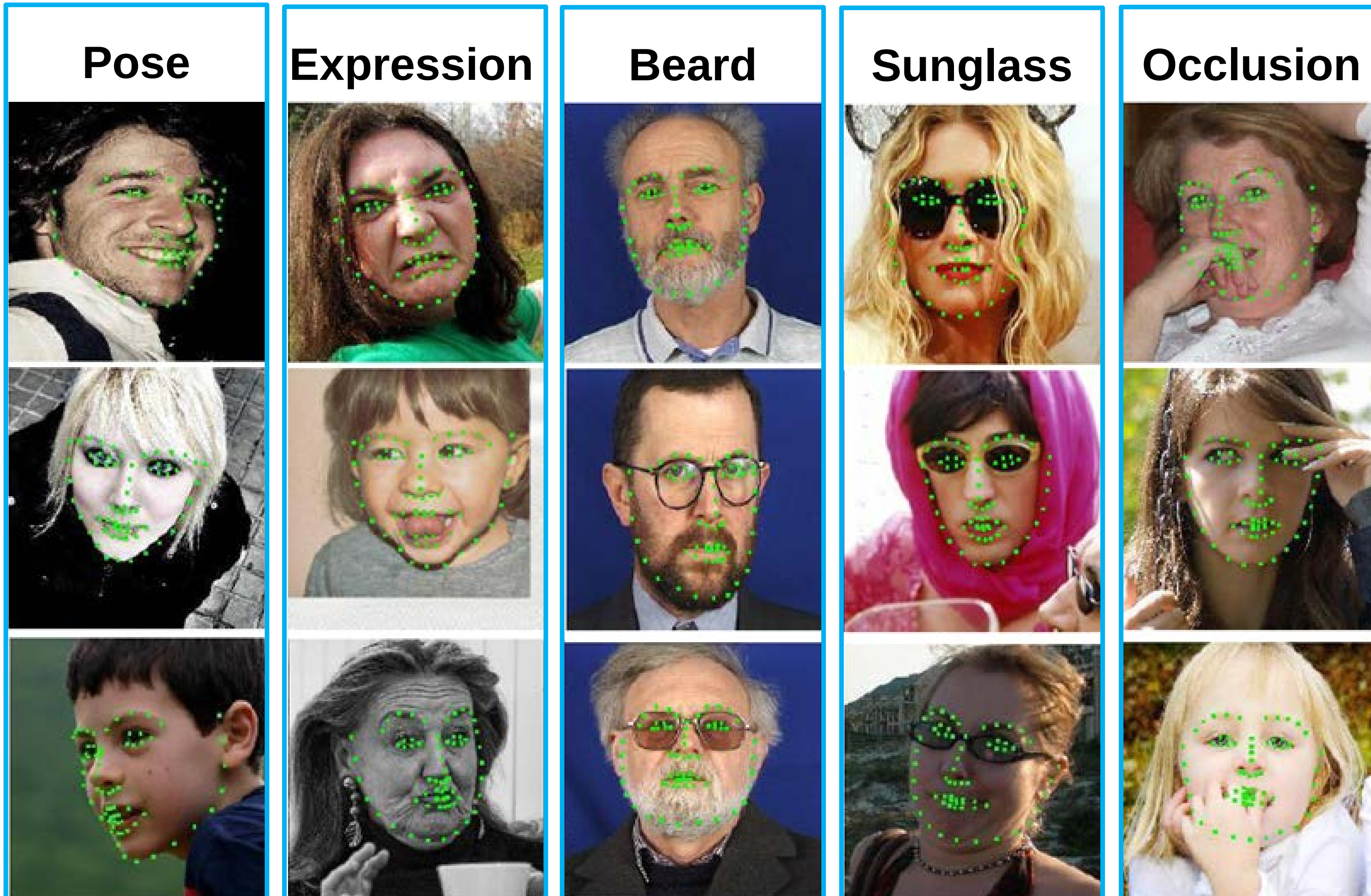
Note: The performance is evaluated in terms of five common landmarks

Experimental Result(8/8)

Institute of Computing Technology

Experimental Result(8/8)

Institute of Computing Technology



Conclusions

Institute of Computing Technology

- Global SAN achieves more accurate initialization
- SAE well characterizes the non-linearity from appearance to face shape
- Coarse-to-fine strategy is effective
  - Alleviate the local minimum problem
- Impressive improvement and real-time performance

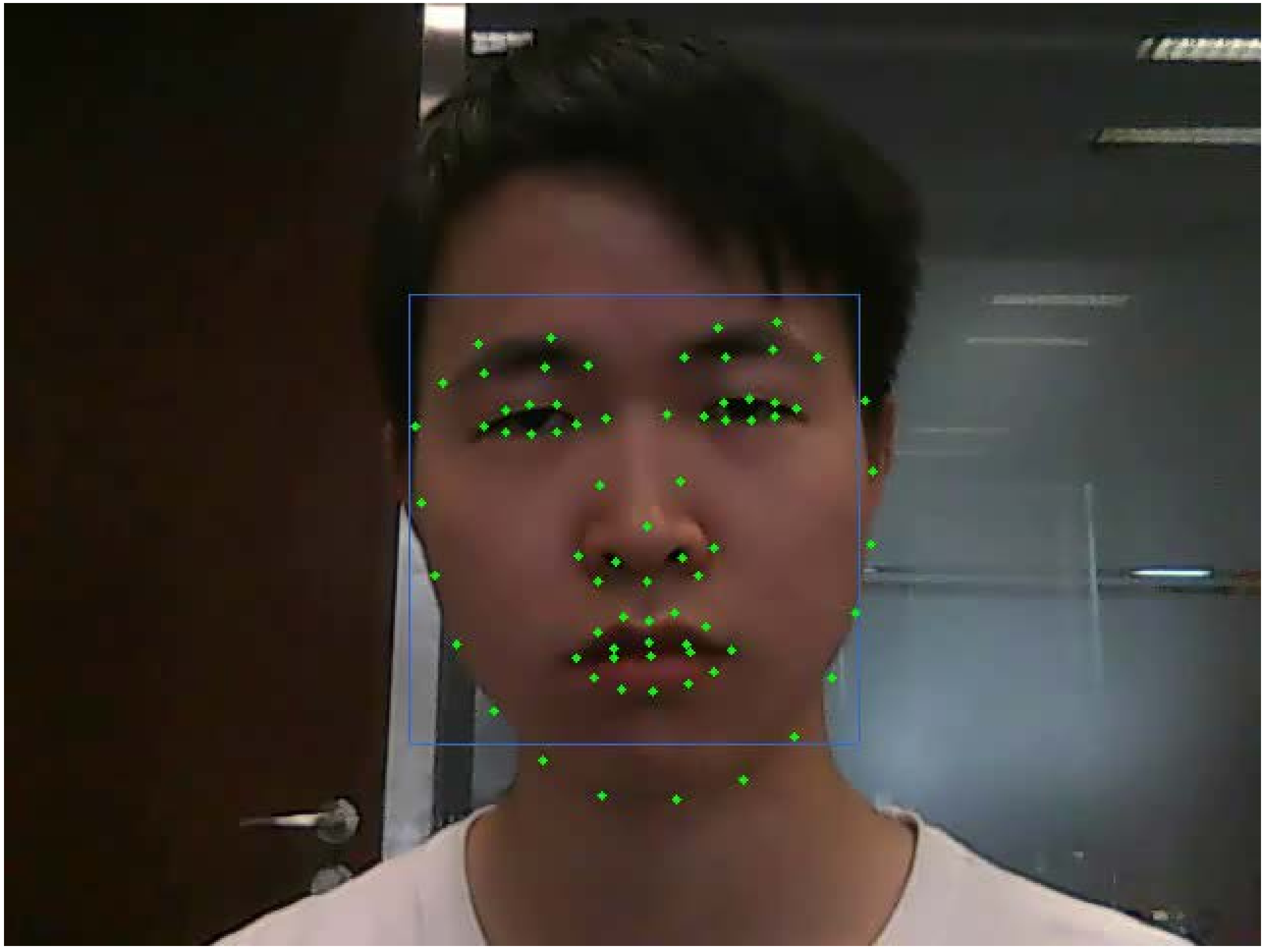
Institute of Computing Technology

Model is available online!

http://vipl.ict.ac.cn/resources

Poster ID: P-2A-50

Demo



References

Institute of Computing Technology

- Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 23(6), 681–685 (2001)
- Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision (IJCV) 60(2), 135–164 (2004)
- Zhao, X., Shan, S., Chai, X., Chen, X.: Locality-constrained active appearance model. In: Asian Conference on Computer Vision (ACCV), pp. 636–647 (2012)
- Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape modelstheir training and application. Computer Vision and Image Understanding (CVIU) 61(1), 38–59 (1995)
- Gu, L., Kanade, T.: A generative shape regularization model for robust face alignment. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 413–426. Springer, Heidelberg (2008)
- Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008
- Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: British Machine Vision Conference (BMVC), vol. 17, pp. 929–938 (2006)
- Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: IEEE International Conference on Computer Vision (ICCV), pp. 1034–1041 (2009)
- Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2887–2894 (2012)
- Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2013)
- Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2013)
- Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3476–3483 (2013)
- Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2014)

References

- Wu, Y., Wang, Z., Ji, Q.: Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3452–3459 (2013)
- Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: The extended m2vts database. In: International Conference on Audio and Video-based Biometric Person Authentication, AVBPA(1999)
- Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 545–552 (2011)
- Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV (2012)
- Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 2879{2886 (2012)
- Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark ftting via optimized part mixtures and cascaded deformable shape model. In: IEEE International Conference on Computer Vision (ICCV) (2013)
- Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085 (2010)
- Burgos-Artizzu, X.P., Perona, P., Doll′ar, P.: Robust face landmark estimation under occlusion. In: IEEE International Conference on Computer Vision, ICCV (2013)