Mnemonic Descent Method
A recurrent process applied for end-to-end face alignment

Imperial College London

George Trigeorgis* Patrick Snape* Mihalis A. Nicolaou† Epameinondas Antonakos* Stefanos Zafeiriou*

* Department of Computing, Imperial College London, UK
† Department of Computing, Goldsmiths University, UK

Problem Statement

Goal: (*meta-*) optimization of non-linear least squares problems.
Given an initial estimate of the solution $\mathbf{x}_0$, find the optimal updates, $\mathbf{x}_{t+1} = \mathbf{x}_t - \Delta\mathbf{x}_t$, that lead to local minima $\mathbf{x}^*$ of the provided objective function.

Application

Face Alignment. Given an estimate of the N facial landmarks $\mathbf{x}_0 = [x_1, y_1, \ldots, x_N, y_N]^\top$. and given a face image, the goal is to estimate a shape $\mathbf{s}$ that is as close as possible to the true shape $\mathbf{x}^*$.

Input image aligned with the mean face

Final shape estimate

Useful for:

- face recognition,
- face tracking,
- emotion recognition,
- face animation,
- 3D face modelling/Morphable Models.

Contributions

- a non-linear cascaded framework for end-to-end learning of the descent directions of non-linear functions.
- an end-to-end trained model; from pixels intensities to the final predictions.
- the first memory-based descent learning model.
- improve on the state-of-the-art on face alignment by a large margin.

Cascaded Regression

- a cascade of *independent, usually linear* regressors is learnt from $\mathbf{x}_0$ to $\mathbf{x}^*$,
- usually uses *handcrafted features*, or features that are not shared across the cascades.

Usual example is the Supervised Descent Method (SDM) [4] which proposes to learn a series of k linear regressions formulated as:

$$\underset{\mathbf{R}^{(k)}}{\arg\min} \|\Delta\mathbf{X}^{(k)} - \mathbf{R}^{(k)}\left[\mathbf{\Phi}^{(k)}\mathbf{1}\right]\|_F^2,$$

Model

Figure 1: An illustrative example of MDM for a total of $T = 3$ time-steps. Initially the network input consists of a partial image observation, consisting of the patches extracted at the mean face $\mathbf{x}_0$. The extracted patches (30×30) at each time-step are passed through a subsequent convolutional network $f_c(\cdot;\theta_c)$, which in turn produces a representation that is robust to changes in appearance variation. Based on the current state $\mathbf{h}_t$, the mnemonic module (implemented as a recurrent network) generates a new state $\mathbf{h}_{t+1}$ and a new set of descent directions $\Delta\mathbf{x}_{t+1}$ that indicates where the network should focus next. After a total of $T = 3$ time-steps, MDM successfully estimates the landmark locations.

An end-to-end trainable objective function: $\min_\theta \|\mathbf{x}^* - \mathbf{x}_0 + \sum_{t=0}^{T-1} f(I, \mathbf{h}_t; \theta)\|_2^2$

- MDM maintains an internal memory unit with the history of *all* past observations of the input space.
- Alignment of any near profile face from a frontal initialisation will have an extremely similar sequence of descent directions.

Figure 2: A t-SNE depiction of the internal states $(T = 1)$ of MDM when asked to align 2000 randomly selected images of CMU Multi-PIE. Each colour corresponds to a cluster of head pose.

Results

We report state-of-the-art results, even on the more challenging 300W dataset.
We compare **MDM** (□), against:
Constrained-Local Neural Fields [1] (●), **ERT** [2] (◄), **PO-CR** [3] (▲), **Face++** [5] (–), **CFSS** [6] (+), **Intra-face** [4] (▼), and Chehra (△). and Yan et.al. (●),

Figure 3: Quantitative results on the test set of the 300W competition (indoor and outdoor combined) for both 68-point (left) and 51-point (right) plots.

	51-points		68-points	
Method	**AUC$_{0.08}$**	**Failure (%)**	**AUC$_{0.08}$**	**Failure (%)**
CLNF [1]	37.65	17.17	19.55	38.83
ERT [2]	40.60	13.50	32.35	17.00
PO-CR [3]	47.65	11.70	–	–
Face++ [5]	53.29	5.33	32.81	13.00
CFSS [6]	50.79	7.80	39.81	12.30
MDM	**56.34**	**4.20**	**45.32**	**6.80**

Table 1: Quantitative results on the test set of the 300W competition using the AUC (%) and failure rate

References

[1] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency.
Constrained local neural fields for robust facial landmark detection in the wild.
In *International Conference on Computer Vision Workshops (ICCVW)*, pages 354–361. IEEE, 2013.

[2] Vahdat Kazemi and Josephine Sullivan.
One Millisecond Face Alignment with an Ensemble of Regression Trees.
In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874. IEEE, 2014.

[3] Georgios Tzimiropoulos.
Project-Out Cascaded Regression With an Application to Face Alignment.
In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3659–3667, 2015.

[4] Xuehan Xiong and Fernando De la Torre.
Supervised descent method and its applications to face alignment.
In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539. IEEE, 2013.

[5] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin.
Extensive facial landmark localization with coarse-to-fine convolutional network cascade.
In *International Conference on Computer Vision Workshops (ICCVW)*, pages 386–391. IEEE, 2013.

[6] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang.
Face Alignment by Coarse-to-Fine Shape Searching.
In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, 2015.

`http://www.menpo.org`
A Python framework for deformable modelling.