Supplementary Material of Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment

In this supplemental material we provide additional qualitative, but also quantitative results for the Mnemonic Descent Method.

Robustness to varying initialisation

To demonstrate the robustness of MDM to varying initialisations, we provide an additional experiment on the LFPW-test, HELEN-test and IBUG datasets. Here we randomly add different levels of uniform noise to the scale and translation of the initial bounding boxes (up to 10% of the initial bounding box). As is evident in Fig. 1, MDM is much more robust to the initialisation than its SDM counterpart.

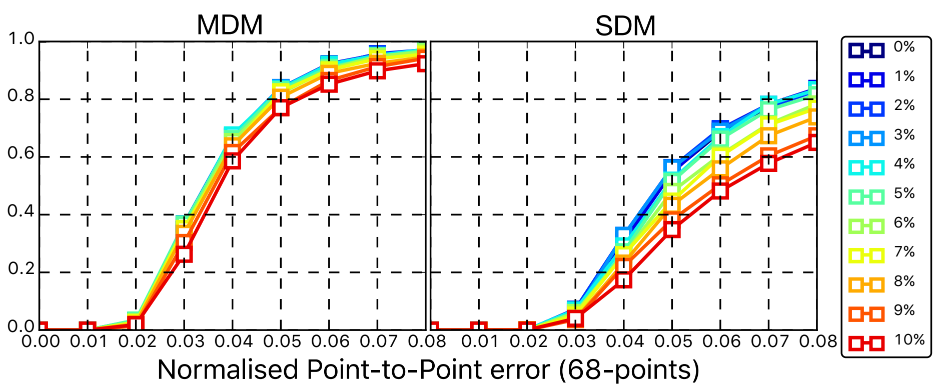

Figure 1: Quantitative results on the LFPW-test, HELEN-test and IBUG datasets for initialisations with varying noise for both MDM (left) and the SDM (right).

Running time

In Tab. 1 we provide the running time and parameters of our model compared to an SDM model. The unoptimised CPU version of our model matches the run time of our SDM implementation, but we should also note that MDM is also easily parallelisable using GPUs ($\sim$ 5ms per image). We emphasise that we did not focus on runtime speed and that these numbers could be improved drastically. Finally, it is noteworthy to mention that model size of an SDM scales linearly with the number of timesteps in contrast to the MDM as also shown in Tab. 1.

Method	**Runtime (ms)**	**# Parameters**
SDM	75.38 ± 2.1	$\mathcal{O}(d \cdot f \cdot T)$
MDM	74.03 ± 1.8	$\mathcal{O}(f \cdot u + u^2 + u \cdot d)$
MDM (GPU)	4.31 ± 0.03	

Table 1: Average time of fitting 1000 images; the model size complexity wrt. number of parameters

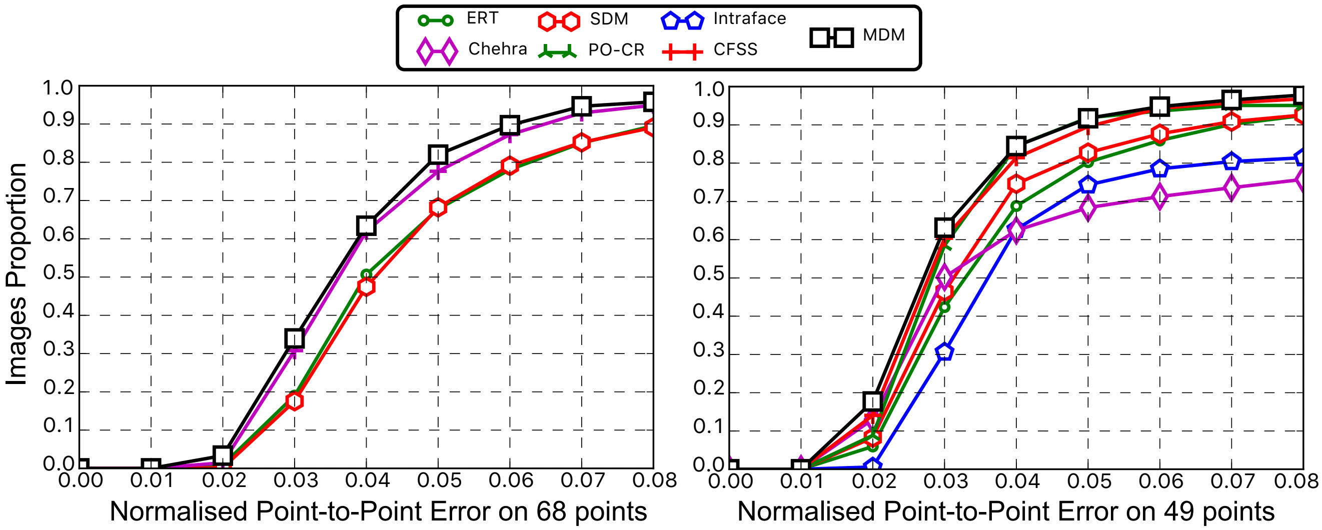

Figure 2: A comparison on the full validation set of the 300W competition (LFPW, HELEN, IBUG), which was used as a validation set for both 68-point (left) and 48-point (right) markups. Results indicate that, due to dramatic advancements in the area of face alignment, the community has reached a point of saturation for these datasets. This is likely due to the fact that the images of the Helen and LFPW datasets have a limited amount of appearance and pose variation.

Table 2: Quantitative results on the public test set of the 300W competition. (68-points)

Method	Mean	Median	MAD	$\text{AUC}_{0.08}$	Failure Rate (%)
ERT	5.61 ± 6.72	3.98	1.03	43.13	10.45
SDM	5.36 ± 4.87	4.13	1.03	42.94	10.89
CFSS	4.25 ± 2.84	3.57	0.82	49.87	5.08
MDM	4.05 ± 2.73	3.44	0.82	52.12	4.21

Table 3: Quantitative results on the public test set of the 300W competition. (49-points)

Method	Mean	Median	MAD	$\text{AUC}_{0.08}$	Failure Rate (%)
ERT	4.87 ± 7.32	3.22	0.75	52.31	7.55
SDM	4.37 ± 5.24	3.09	0.73	54.58	7.26
Intraface	10.18 ± 16.94	3.52	0.82	46.08	18.58
Chehra	10.97 ± 18.92	2.99	1.00	47.26	24.24
PO-CR	5.40 ± 12.88	2.74	0.53	60.35	4.93
CFSS	3.36 ± 3.02	2.73	0.60	60.52	3.19
MDM	3.19 ± 2.52	2.63	0.59	62.17	2.18

Figure 3: Application of different face alignment methods on challenging images from the public test set of 300W (LFPW-test, HELEN-test, IBUG).

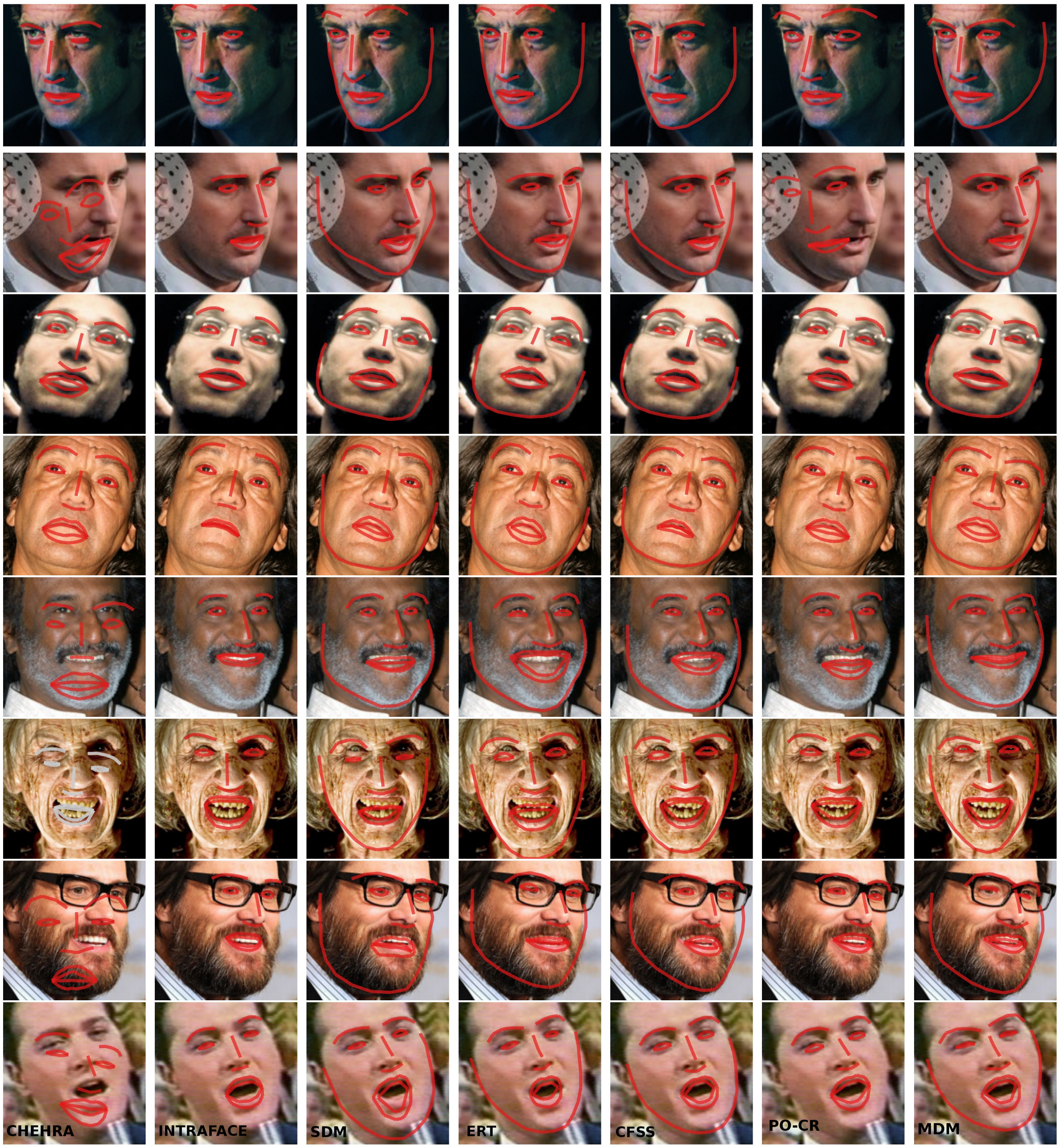

Figure 4: Depicted are the 8 *worst* fitting results for the Mnemonic Descent Method on the LFPW in terms of the interocular error. For comparison we show the estimated shapes from 5 state-of-the-art face landmark localisation techniques. This both demonstrates the saturation of datasets such as LFPW for the task of face alignment but also the highly accurate fittings of our approach.