Recurrence without Recurrence: Stable Video Landmark Detection with Deep Equilibrium Models

Paul Micaelli
University of Edinburgh
paul.micaelli@ed.ac.uk

Arash Vahdat
NVIDIA
avahdat@nvidia.com

Hongxu Yin
NVIDIA
dannyy@nvidia.com

Jan Kautz
NVIDIA
jkautz@nvidia.com

Pavlo Molchanov
NVIDIA
pmolchanov@nvidia.com

Appendix A: More details on making our WFLW-V dataset

In this section we detail the procedure used to collect, label and curate the 1000 videos that make up the WFLW-V dataset.

Step 1: Video search

We start by producing a list of 100 YouTube search strings, that we think would be correlated with videos conducive to landmark uncertainty. These search strings fall within 7 categories: "Skin care & Makeup" (*e.g. how to put lipstick*), "Hair & Beard care" (*e.g. how to cut your own hair*), "Singing & Podcasts" (*e.g. how to setup your mic*), "Brass instruments" (*e.g. learn to play the French horn*"), "Eating" (*e.g. how to eat fast*), "Smoking" (*e.g. how to smoke the cigar*), and "Miscellaneous" (*e.g. how to brush your teeth*"). Each English search string is translated to 10 languages, to produce more diverse videos: French, German, Spanish, Italian, Portuguese, Catalan, Czech, Danish, Estonian, Dutch.

We use YouTube filters to search for videos less than 4 minutes long, and with a CC BY licence. This licence is the most permissive creator licence. It allows reusers to distribute, remix, adapt the video, and even to use it for commercial use. We only consider videos that have a frame rate between 24 and 31 fps inclusive. This is mostly to exclude all videos like kid cartoons that have very low fps. In total, this step produces around $15,000$ videos.

Step 2: Video cleaning

Our task is now to find 5s of contiguous *clean* face for each video. A *clean* face is a real human face, at least 20% visible, from a single person, without video or camera filters (*e.g.* face filters, jump cuts). We also limit the number of videos that come from the same youtuber, so as not to lower diversity. We use the most popular face detector, the Multi-task Cascaded Convolutional Networks (MTCNN) [6] to help with video cleaning. In total, this leaves around $2,000$ videos.

Step 3: Video annotation

We use an oracle made up of 45 pretrained models, including 15 large Unets (larger than our LDEQ backbone), 15 HRNets-W48, and 15 HRFormer-B. We average the final heatmap of each model to create a mean heatmap, from which we extract our oracle predictions. We found that the bounding box from the MTCNN model are jittery, which in turns facilitates jitter and flicker for landmarks. To fix this we bootstrap our oracle to the bounding box detection. This is done by using the original MTCNN bounding box, finding landmarks, defining a new bounding box based on the smallest/largest landmark coordinates and a scaling margin factor of 1.2, finding landmarks in this new bounding box, and so on. This is repeated for 3 iterations, after which the bounding box values have have converged. We use the landmark predictions on the final bounding box as our oracle predictions.

As our oracle outputs the mean of an ensemble of M independent models, the error on this mean is given by $\sigma/\sqrt{M}$, where σ is the standard deviation of the M predictions. For $M = 45$ we measured this error to be $\sim 0.2\%$ of the mean for the hard subset of WFLW-V.

Step 4: Video verification

We verify each video frame by frame for issues. In some cases, videos are discarded because the degree of uncertainty or occlusion is too high that even a "human best guess" wouldn't be good. This includes videos with very large poses as well. In other cases, particularly for hard videos, some models in the ensemble are visibly mistaken. These models are singled out and removed for problematic

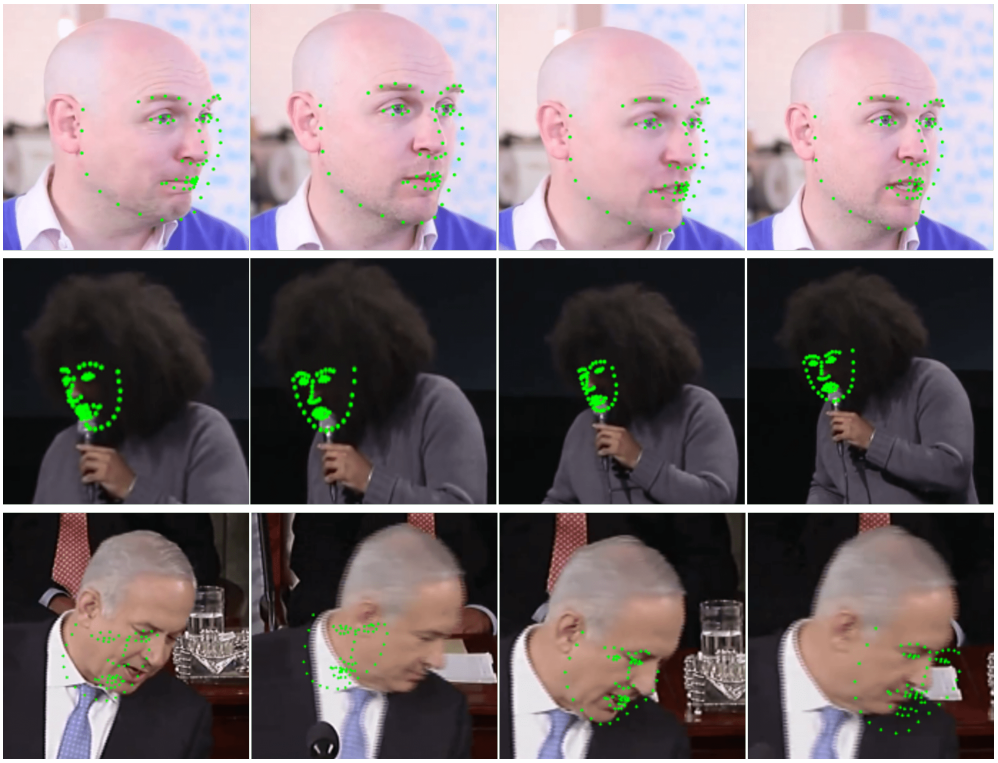

Figure 1. Examples of poorly labelled videos in the 300-VW dataset. We show three levels of labelling errors from top to bottom: medium, bad, very bad. Our new WFLW-V dataset uses much stronger labellers and was checked frame by frame to avoid such errors.

frames. These cases are rare ($\sim$ 25) but worth correcting so we don't bias the dataset by only including videos where our oracle does best.

Step 5: Subset creation

Since we have access to all 45 model predictions in the ensemble, it is easy to see the average variance of these models for each video. This score correlates well with uncertainty, and we use it to rank all videos from hard to easy. We used the top 500 videos for WFLW-hard, and the bottom 500 for WFLW-easy.

Appendix B: Errors in 300-VW

The 300-VW dataset [4] was labelled using the now obsolete models from [2] and [5]. This results in several labelling errors (Fig. 1) that have gone unnoticed. Errors in the ground truth of datasets lead to misleading insights and models that generalize poorly to real-world settings.

We also note that many (perhaps all) of the videos in 300-VW do not have a creative commons licence, and so the legality of their use for industrial research labs may be more ambiguous.

Appendix C: WFLW-V Results

We show the results of Figure 5 in tabular form in Tab. 1. We compare our RwR scheme to the exponential moving average (ema), and show that contrary to ema, our method can improve temporal coherence without lowering accuracy. We tried the following ema weights: [0.005, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. When considering all baselines at once, we found that a weight 0.15 struck the best balance between lowering NMF without increasing NME too much. This was also better than the grid searched Savitzky-Golay filter [3] and One Euro filter [1]. The only exception was the HIH model which is both very jittery and flickery, and for which we used an ema weight of 0.3.

Finally, we found that more augmentations can help performance on WFLW-V while reducing performance on the WFLW test set. This is likely because the WFLW-V dataset is more diverse than the WFLW test set, and unecessary augmentations on WFLW can reduce performance. We therefore retrained LDEQ with more augmentations to get the best performance on WFLW-V.

Method	WFLW-V hard		WFLW-V easy		WFLW-V FULL	
	NME	NMF	NME	NMF	NME	NMF
HIH	3.93	423.11	2.48	294.94	3.20	359.03
+ ema	4.15	313.07	2.54	208.60	3.34	260.84
StackedHourglass	3.93	255.74	2.33	125.91	3.13	190.82
+ ema	3.99	231.52	2.37	119.35	3.18	175.43
HRFormer-S	3.92	289.50	2.29	150.91	3.11	220.21
+ ema	3.98	255.85	2.33	136.37	3.15	196.11
HRNet-W18	3.61	236.96	2.16	127.72	2.89	182.34
+ ema	3.68	215.26	2.20	119.13	2.94	167.20
SDFL	3.13	207.68	1.83	115.22	2.48	161.45
+ ema	3.21	192.77	1.87	108.35	2.54	150.56
Awing	2.90	277.86	1.68	171.48	2.29	224.67
+ ema	2.96	242.95	1.70	146.71	2.33	194.83
HRNet-W32	2.60	203.15	1.45	105.35	2.03	154.25
+ ema	2.71	186.69	1.51	99.62	2.11	143.15
Unet	2.53	189.28	1.38	94.35	1.95	141.81
+ ema	2.65	175.76	1.45	91.58	2.05	133.67
SLPT	2.42	216.56	1.32	105.93	1.87	161.25
+ ema	2.52	195.35	1.39	98.79	1.96	147.07
LDEQ	2.31	197.16	1.24	84.03	1.77	140.59
+ RwR	2.30	172.95	1.24	82.74	**1.77**	**127.85**

Table 1. NME and NMF on the WFLW-V dataset, comparing the effect of an exponential moving average smoothing (ema) with our recurrence without recurrence scheme.

References

[1] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1 € filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 2527–2530, New York, NY, USA, 2012. Association for Computing Machinery. 2

[2] Grigoris G. Chrysos, Epameinondas Antonakos, Stefanos Zafeiriou, and Patrick Snape. Offline deformable face tracking in arbitrary videos. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 954–962, 2015. 2

[3] Abraham. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. 2

[4] Jie Shen, Stefanos Zafeiriou, Grigoris G. Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1003–1011, 2015. 2

[5] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3659–3667, 2015. 2

[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 1