

Handling Imbalance in Automatic Facial Action Intensity Estimation

Supplemental Material BMVC 2015

Philipp Werner, Frerk Saxen, and Ayoub Al-Hamadi

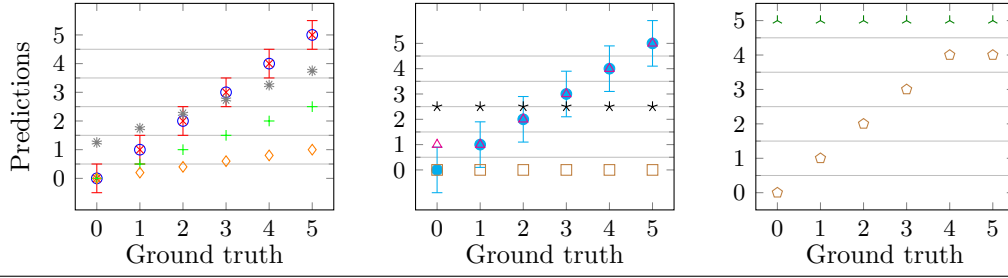
Contents

1	Comparison of Performance Measures using Artificial Data	1
2	MIDRUS Parameter: Single SVR	3
3	MIDRUS Parameters: SVR Ensemble	5
4	AU Performance Comparison	7
5	Qualitative Observations	9
5.1	Underestimation	9
5.2	Overestimation	9
5.3	Challenges with the UNBC dataset	10

1 Comparison of Performance Measures using Artificial Data

This Section supplements Section 3.1 of the paper. Table 1 is similar to Table 2 of the paper, but includes additional performance measures and predictions (in bold).

Macro-averaged MSE^M has similar characteristics to MAE^M , i.e. is lower for predictions at (*) or near (*) the mean of the scale. This is relevant, as the phenomenon simulated in * appears in practice (see Fig. 1, $\alpha = 1$). Further, the macro-averaged error measures (MAE^M and MSE^M) rate \blacktriangle and \blacklozenge identically, i.e. they yield too good values for \blacktriangle (from our perspective). MAE^μ is biased towards under-estimation models (\blacktriangle and \blacklozenge) and has a very low error for the best trivial model (\square), making it quite useless for such imbalanced class distributions. $ICC(3,1)_c$ is very similar to $ICC(3,1)_d$, but we recommend to use the discrete variant due to the reasons listed in the paper. $ICC(1,1)$ puts a larger emphasis on the majority class(es) than $ICC(3,1)$, which is apparent in \blacktriangle and *. In general, it also tends to be lower than $ICC(3,1)$. Further it yields negative performances for trivial models, which is cut off by $ICC(3,1)$. For $F1^M$ we want to mention that it does not take into account the ordinal relationship between classes, i.e. if class 0 sample is misclassified, it does not matter whether it is classified as class 1 or class 5. The accuracy measure is well-known to put high emphasis on the majority class(es). Due to the high performance reachable with a trivial model (\square), it is unsuitable for imbalanced problems.



Prediction	MSE_c^μ	MSE_d^μ	MSE_c^M	MSE_d^M	MAE_c^μ	MAE_d^μ	MAE_c^M	MAE_d^M
○ perfect	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
× perfect + noise 0.5	0.251	0.182	0.250	0.270	0.400	0.178	0.398	0.266
● perfect + noise 0.9	0.812	0.492	0.827	0.686	0.718	0.379	0.726	0.546
+ 0.5 · perfect	0.176	0.119	2.292	1.667	0.133	0.093	1.250	1.000
* 0.5 · perfect + 1.25	1.406	0.933	0.729	0.667	1.155	0.933	0.750	0.667
◇ 0.2 · perfect	0.452	0.448	5.867	5.667	0.213	0.221	2.000	2.000
△ perfect, 0 → 1	0.874	0.874	0.167	0.167	0.874	0.874	0.167	0.167
⬠ perfect, 5 → 4	0.003	0.003	0.167	0.167	0.003	0.003	0.167	0.167
□ always 0	0.706	0.706	9.167	9.167	0.267	0.267	2.500	2.500
* always 2.5	5.623	8.106	2.917	3.167	2.310	2.764	1.500	1.500
⋈ always 5	23.040	23.040	9.167	9.167	4.733	4.733	2.500	2.500
random	8.548	8.548	5.892	5.892	2.395	2.395	1.956	1.956

Prediction	PCC_c	PCC_d	$ICC(3,1)_c$	$ICC(3,1)_d$	$ICC(1,1)_c$	$ICC(1,1)_d$	$F1^M$	Accuracy
○ perfect	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
× perfect + noise 0.5	0.846	0.882	0.835	0.880	0.835	0.867	0.645	0.823
● perfect + noise 0.9	0.664	0.743	0.611	0.733	0.611	0.683	0.406	0.675
+ 0.5 · perfect	1.000	0.976	0.800	0.875	0.780	0.865	0.287	0.920
* 0.5 · perfect + 1.25	1.000	0.976	0.800	0.875	0.008	0.277	0.245	0.067
◇ 0.2 · perfect	1.000	0.841	0.385	0.411	0.339	0.362	0.159	0.874
△ perfect, 0 → 1	0.955	0.955	0.880	0.880	0.325	0.325	0.682	0.126
⬠ perfect, 5 → 4	0.998	0.998	0.998	0.998	0.998	0.998	0.815	0.997
□ always 0	undef	undef	0.000	0.000	-0.053	-0.053	0.155	0.874
* always 2.5	undef	undef	0.000	0.000	-0.797	-0.855	0.010	0.032
⋈ always 5	undef	undef	0.000	0.000	-0.946	-0.946	0.001	0.003
random	-0.003	-0.003	-0.002	-0.002	-0.412	-0.412	0.081	0.167

Table 1: Comparison of performance measures on artificial predictions. Compare with Table 2 in the paper; additional measures and predictions are inserted in bold. The background shading encodes the ranking of predictions in the respective measure.

2 MIDRUS Parameter: Single SVR

In this Section we supplement the single SVR results presented in Sec. 4.1 and Fig. 3a of the paper. In Table 2 we report the results of varying α for the performance measures that we evaluated with artificial data. For most of the error measures we get best results when using either the original imbalanced training set ($\alpha = 0$) or the ‘balanced’ set ($\alpha = 1$, absence class is balanced with the second most frequent class). In contrast, the correlation measures score the results best for moderate damping ($0.2 < \alpha < 0.5$). In Fig. 1 we show some corresponding confusion matrices and box plots to investigate what are the results beyond the summarizing numbers. As to be expected, $\alpha = 0$ yields the lowest error rate for class 0 (the most frequent class). But it also promotes under-estimation of the intensity, which is particularly evident for AU 11 (nasolabial deepener). With $\alpha = 1$ the performance on class 0 is very poor, especially for AU 11 and 34 (cheek puff). Here we get a combination of over-estimation for the low intensity classes and under-estimation for the high intensity classes, i.e. the predictions tend to be closer to the mean of the scale. That is why the macro-averaging error measures are lower for increasing α , as we have also seen in Table 1 for $+$ and $*$. In contrast, the other error-based measures (MSE^μ and MAE^μ) are better for $\alpha = 0$. The correlation-based measures score the results obtained with $\alpha = 0.35$ better. The results of $\alpha = 0$ and $\alpha = 0.35$ are similar for the well-performing AU models, but $\alpha = 0.35$ tends to be better for the more challenging AUs. In contrast to well-performing AUs, most of the challenging AUs are characterized by a stronger class imbalance, which is compensated by MIDRUS.

α	MSE_c^μ	MSE_d^μ	MSE_c^M	MSE_d^M	MAE_c^μ	MAE_d^μ	MAE_c^M	MAE_d^M
0 (imbalanced)	0.607	0.534	2.809	2.813	0.566	0.373	1.289	1.220
0.2	0.584	0.542	2.333	2.351	0.560	0.391	1.168	1.106
0.35	0.597	0.575	2.081	2.125	0.573	0.423	1.090	1.039
0.5	0.659	0.660	1.865	1.901	0.610	0.483	1.039	0.984
0.65	0.749	0.772	1.739	1.784	0.662	0.560	1.004	0.952
0.8	0.894	0.941	1.621	1.641	0.743	0.672	0.977	0.925
1 (‘balanced’)	1.207	1.274	1.521	1.614	0.911	0.875	0.974	0.947

α	PCC_c	PCC_d	$\text{ICC}(3,1)_c$	$\text{ICC}(3,1)_d$	$\text{ICC}(1,1)_c$	$\text{ICC}(1,1)_d$	F1^M	Accuracy
0 (imbalanced)	0.542	0.564	0.541	0.556	0.541	0.542	0.265	0.694
0.2	0.573	0.591	0.572	0.586	0.568	0.563	0.277	0.674
0.35	0.581	0.591	0.581	0.589	0.565	0.552	0.280	0.645
0.5	0.577	0.580	0.575	0.577	0.537	0.517	0.282	0.596
0.65	0.567	0.563	0.564	0.559	0.485	0.462	0.276	0.534
0.8	0.557	0.545	0.554	0.540	0.405	0.383	0.265	0.449
1 (‘balanced’)	0.540	0.516	0.537	0.512	0.245	0.231	0.222	0.305

Table 2: Cross-validated performances of SVR with MIDRUS with varied parameter α in different measures. See Fig. 1 to get an impression of the underlying predictions and class confusion.

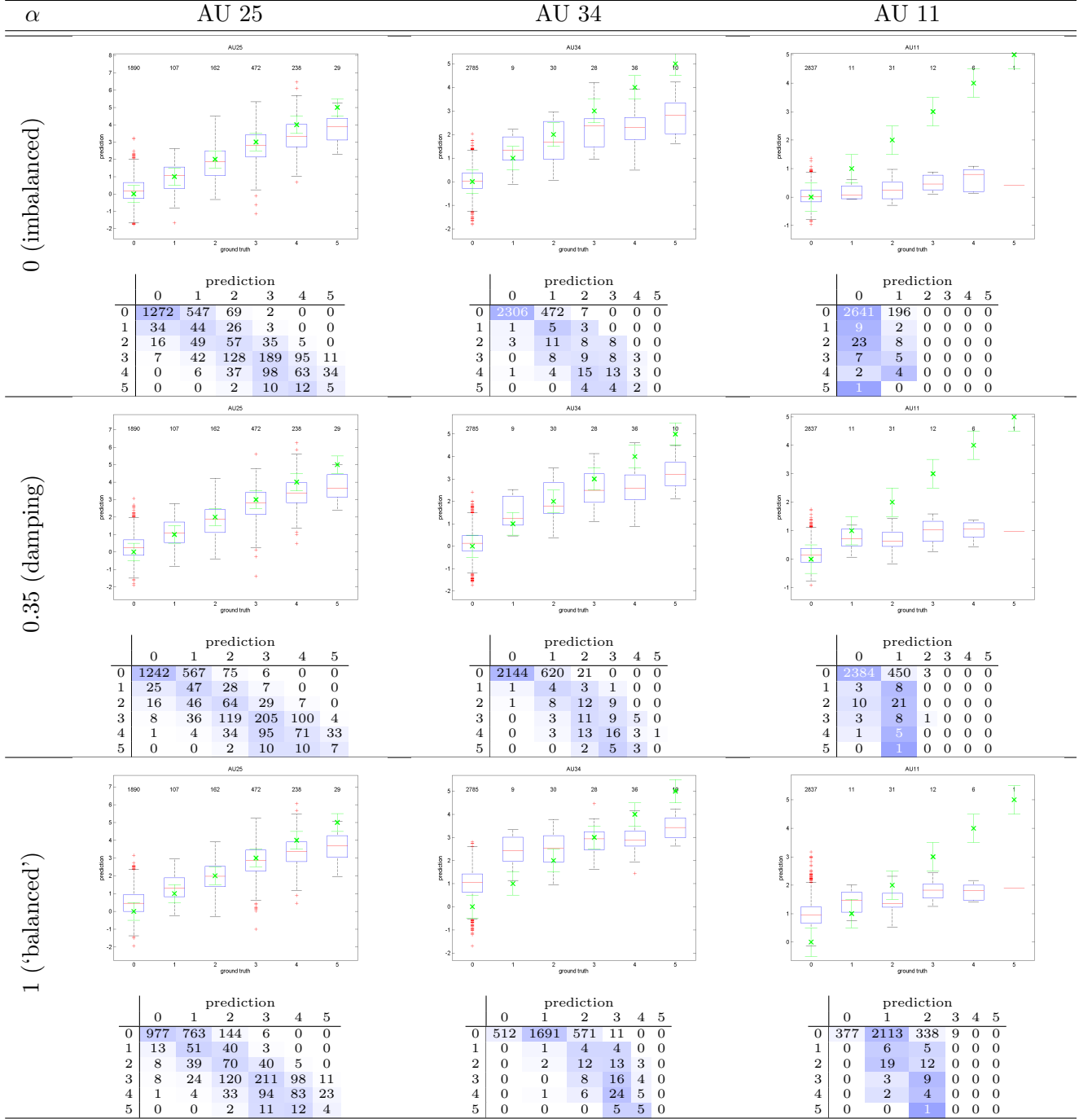
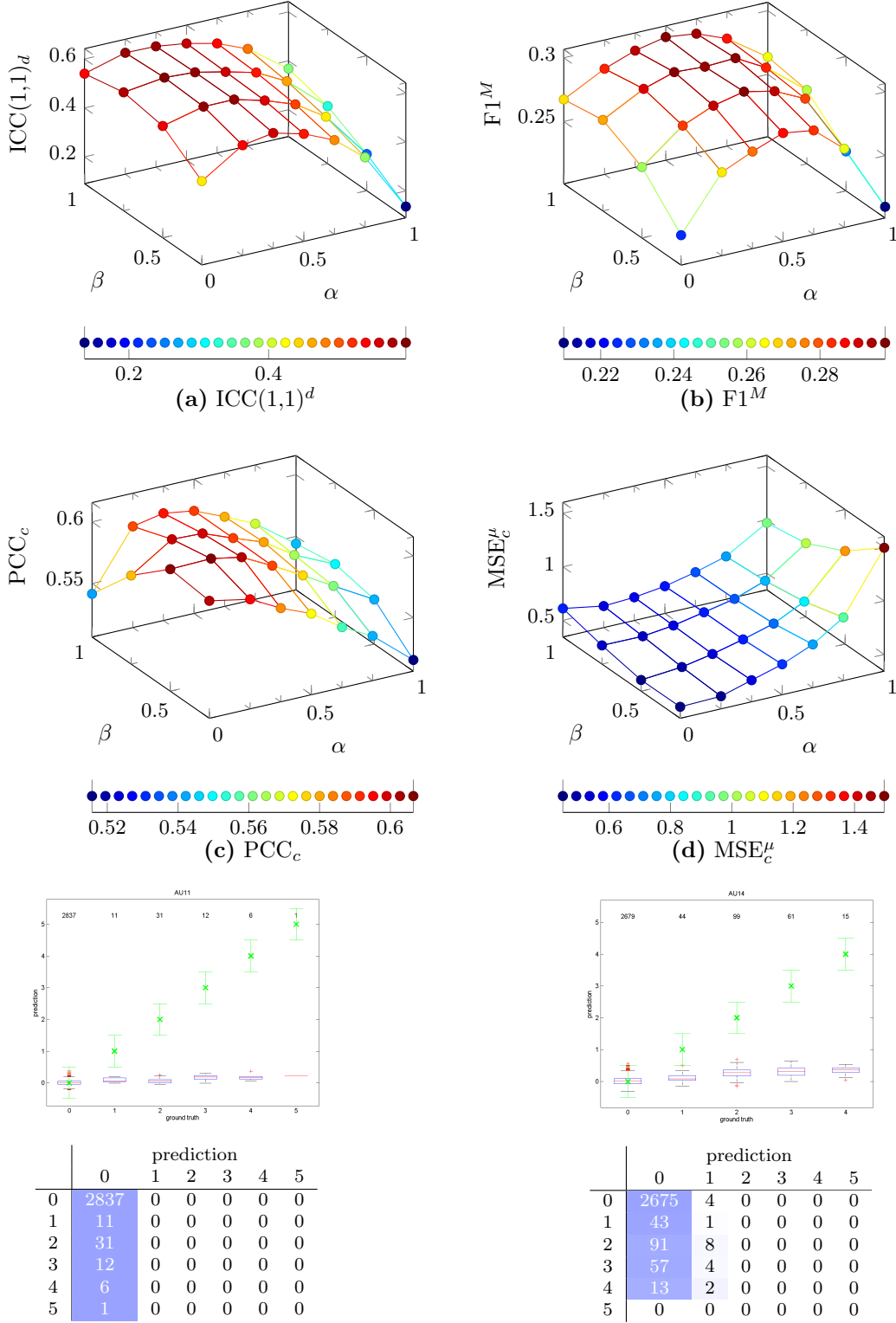


Figure 1: SVR predictions (box plots and confusion matrices) on test set for some representative AUs and selected values of α on Bosphorus dataset. Also see Table 2.

3 MIDRUS Parameters: SVR Ensemble

In this Section we supplement the MIDRUS SVR Ensemble results presented in Sec. 4.1 and Fig. 3b of the paper. In Fig. 2a-d we show the results of varying α and β for several additional measures. $\text{ICC}(1,1)_d$ (see Fig. 2a) and F1^M behave similarly to $\text{ICC}(3,1)_d$. The benefit of damping ($0 < \alpha < 1$) is apparent. Further, you see that it may be beneficial to select $\beta < 1$ with an ensemble, not only in terms of training time, but also regarding predictive performance. Sub-figures 2c-d depict the plots for two other widely used measures: PCC_c and MSE_c^μ . The plots differ qualitatively from the other measures, but mainly due to issues of the measures that we already discussed. In the sub-figures 2e-f we illustrate with two examples that PCC_c and MSE_c^μ can be misleading and can lead to wrong conclusions. The issue in (e) is related to the performance level achievable with the trivial classifier. In (f) the PCC_c yields a quite high performance estimate due to the invariance of the measure to linear transformations.



(e) AU 11 with $\alpha = 0$ and $\beta = 0.25$: $MSE_c^\mu = 0.12$. The learned model classifies all samples to class 0, but gets the best MSE_c^μ among all AUs, which illustrates a problem with the widely used measure.

$ICC(3,1)_d = 0$, $ICC(1,1)_d = -0.01$, $F1^M = 0.16$.

(f) AU 14 with $\alpha = 0$ and $\beta = 0.25$: $PCC_c = 0.45$ is too high for the poor performance, which illustrates a problem (invariance to linear transformation) with the widely used measure.

$ICC(3,1)_d = 0.06$, $ICC(1,1)_d = 0.03$, $F1^M = 0.20$.

Figure 2: Cross-validated performances on Bosphorus dataset (mean of 26 AUs) with MIDRUS SVR Ensemble (varying α and β , $T = 10$). (a) $ICC(1,1)_d$ and (b) $F1^M$ yield similar results to $ICC(3,1)_d$ (see Fig. 3a of the paper). The widely used measures (c) PCC_c and (d) MSE_c^μ differ qualitatively, mainly due to problems with the measures. We consider $\alpha = 0$ and $\beta = 0.25$, where we see good performance in (c) and (d) but bad performance in (a) and (b). E.g. AU 11 (e) and AU 14 (f) illustrate issues of the measures with under-estimation in imbalanced problems.

4 AU Performance Comparison

Table 3, 4, and 5 show the comparison of our proposed approach (SVR Ensemble with MIDRUS: $0 < \alpha < 1$) for the Bosphorus, DISFA, and UNBC dataset. To keep the training time technically feasible, we restricted the sample count per each ensemble model to 2,500 (which is similar to a small β for large datasets like DISFA and UNBC). We provide the results for several performance measures for future comparisons.

ICC(3,1)_d	Mean	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU11	AU12	AU14	AU15	AU16
Easy Ensemble	0.340	0.478	0.519	0.241	0.355	0.306	0.584	0.534	0.148	0.005	0.769	0.267	0.118	0.299
SVRe $\alpha = 0$	0.553	0.675	0.698	0.752	0.551	0.618	0.660	0.795	0.519	0.169	0.802	0.321	0.405	0.393
SVRe $\alpha = 0.5$	0.603	0.709	0.725	0.766	0.597	0.648	0.675	0.794	0.591	0.202	0.817	0.453	0.485	0.499
SVRe $\alpha = 1$	0.533	0.648	0.671	0.720	0.549	0.589	0.676	0.681	0.522	0.152	0.792	0.403	0.382	0.388
		AU17	AU18	AU20	AU22	AU23	AU24	AU25	AU26	AU27	AU28	AU34	AU38	AU43
Easy Ensemble		0.239	0.477	0.189	0.162	0.068	0.343	0.870	0.297	0.608	0.150	0.230	0.204	0.391
SVRe $\alpha = 0$		0.403	0.672	0.455	0.514	0.131	0.552	0.831	0.409	0.758	0.558	0.656	0.306	0.788
SVRe $\alpha = 0.5$		0.520	0.712	0.489	0.529	0.307	0.593	0.840	0.488	0.785	0.606	0.668	0.389	0.793
SVRe $\alpha = 1$		0.490	0.634	0.322	0.344	0.274	0.509	0.831	0.471	0.724	0.457	0.520	0.389	0.714
ICC(1,1)_d	Mean	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU11	AU12	AU14	AU15	AU16
Easy Ensemble	0.226	0.391	0.450	0.100	0.259	0.219	0.563	0.494	-0.004	-0.514	0.754	0.155	-0.070	0.204
SVRe $\alpha = 0$	0.540	0.658	0.684	0.745	0.529	0.606	0.658	0.786	0.506	0.166	0.797	0.307	0.390	0.376
SVRe $\alpha = 0.5$	0.567	0.676	0.689	0.749	0.558	0.620	0.672	0.775	0.558	0.160	0.804	0.402	0.438	0.442
SVRe $\alpha = 1$	0.368	0.506	0.540	0.658	0.419	0.499	0.668	0.575	0.408	-0.195	0.757	0.232	0.088	0.110
		AU17	AU18	AU20	AU22	AU23	AU24	AU25	AU26	AU27	AU28	AU34	AU38	AU43
Easy Ensemble		0.082	0.434	0.055	0.063	-0.170	0.237	0.869	0.171	0.575	0.104	0.155	0.066	0.243
SVRe $\alpha = 0$		0.382	0.656	0.442	0.495	0.112	0.534	0.829	0.394	0.745	0.544	0.637	0.284	0.781
SVRe $\alpha = 0.5$		0.477	0.679	0.441	0.471	0.240	0.553	0.836	0.452	0.755	0.567	0.624	0.340	0.769
SVRe $\alpha = 1$		0.347	0.498	0.014	0.048	-0.035	0.292	0.821	0.350	0.622	0.286	0.276	0.225	0.554
F1^M	Mean	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU11	AU12	AU14	AU15	AU16
Easy Ensemble	0.251	0.280	0.298	0.262	0.229	0.249	0.349	0.240	0.240	0.074	0.404	0.260	0.185	0.207
SVRe $\alpha = 0$	0.267	0.304	0.379	0.375	0.261	0.390	0.341	0.368	0.227	0.161	0.406	0.220	0.188	0.164
SVRe $\alpha = 0.5$	0.298	0.343	0.410	0.419	0.262	0.401	0.384	0.427	0.268	0.154	0.426	0.269	0.205	0.202
SVRe $\alpha = 1$	0.264	0.306	0.372	0.384	0.218	0.366	0.375	0.384	0.248	0.108	0.415	0.230	0.181	0.159
		AU17	AU18	AU20	AU22	AU23	AU24	AU25	AU26	AU27	AU28	AU34	AU38	AU43
Easy Ensemble		0.217	0.265	0.217	0.184	0.160	0.219	0.416	0.275	0.366	0.229	0.216	0.176	0.305
SVRe $\alpha = 0$		0.222	0.266	0.246	0.219	0.150	0.222	0.343	0.210	0.322	0.230	0.245	0.190	0.300
SVRe $\alpha = 0.5$		0.274	0.308	0.269	0.249	0.173	0.252	0.362	0.226	0.367	0.297	0.262	0.203	0.331
SVRe $\alpha = 1$		0.247	0.294	0.198	0.168	0.131	0.215	0.343	0.230	0.364	0.213	0.211	0.185	0.311

Table 3: Performance comparison for each AU in the Bosphorus dataset (similar to Table 4a in the paper). SVRe $\alpha = 0$ corresponds to imbalanced Support Vector Regression Ensemble (SVRe), SVRe $\alpha = 1$ corresponds to balanced SVR Ensemble, and SVRe $\alpha = 0.5$ corresponds to MIDRUS SVR Ensemble. $\beta = 1$ for all SVRe experiments.

ICC(3,1)_d	Mean	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
Mavadati 2014	0.235	0.250	0.220	0.280	0.080	0.170	0.160	0.570	0.080	0.110	0.040	0.630	0.230
Easy Ensemble	0.362	0.199	0.257	0.532	0.112	0.470	0.344	0.769	0.212	0.075	0.064	0.890	0.421
SVRe $\alpha = 0$	0.346	0.194	0.199	0.671	0.000	0.524	0.247	0.785	0.112	0.155	0.021	0.853	0.394
SVRe $\alpha = 0.5$	0.439	0.284	0.321	0.681	0.206	0.547	0.446	0.798	0.199	0.296	0.187	0.856	0.450
SVRe $\alpha = 1$	0.412	0.216	0.283	0.665	0.221	0.525	0.384	0.783	0.226	0.239	0.139	0.842	0.425
ICC(1,1)_d	Mean	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
Easy Ensemble	0.334	0.167	0.219	0.515	0.073	0.449	0.332	0.758	0.197	0.005	0.029	0.889	0.379
SVRe $\alpha = 0$	0.337	0.192	0.198	0.655	-0.008	0.523	0.228	0.780	0.097	0.128	0.006	0.852	0.390
SVRe $\alpha = 0.5$	0.392	0.200	0.239	0.636	0.160	0.525	0.371	0.783	0.150	0.240	0.115	0.853	0.426
SVRe $\alpha = 1$	0.322	0.118	0.169	0.599	0.082	0.465	0.272	0.751	0.114	0.119	-0.026	0.836	0.367
F1^M	Mean	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
Easy Ensemble	0.264	0.221	0.225	0.275	0.177	0.261	0.243	0.367	0.240	0.198	0.211	0.497	0.254
SVRe $\alpha = 0$	0.236	0.162	0.161	0.296	0.165	0.233	0.172	0.386	0.207	0.182	0.201	0.446	0.217
SVRe $\alpha = 0.5$	0.258	0.190	0.182	0.355	0.179	0.242	0.249	0.403	0.202	0.190	0.202	0.443	0.261
SVRe $\alpha = 1$	0.247	0.182	0.192	0.327	0.200	0.241	0.210	0.407	0.214	0.174	0.172	0.395	0.251

Table 4: Performance comparison for each AU in the DISFA dataset (similar to Table 4a in the paper). SVRe $\alpha = 0$ corresponds to imbalanced Support Vector Regression Ensemble (SVRe), SVRe $\alpha = 1$ corresponds to balanced SVR Ensemble, and SVRe $\alpha = 0.5$ corresponds to MIDRUS SVR Ensemble.

ICC(3,1)_d	Mean	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU43
Easy Ensemble	0,273	0,132	0,384	0,367	0,122	0,338	0,438	0,029	0,386	-0,012	0,551
SVRe $\alpha = 0$	0,152	0,051	0,432	0,233	0,024	0,078	0,513	0,000	0,053	-0,003	0,139
SVRe $\alpha = 0.9$	0,307	0,190	0,520	0,359	0,187	0,255	0,518	0,091	0,311	0,085	0,556
SVRe $\alpha = 1$	0,298	0,191	0,491	0,380	0,110	0,243	0,509	0,103	0,286	0,071	0,595
ICC(1,1)_d	Mean	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU43
Easy Ensemble	0,261	0,130	0,379	0,367	0,097	0,328	0,437	0,012	0,371	-0,058	0,543
SVRe $\alpha = 0$	0,143	0,047	0,432	0,214	0,021	0,077	0,512	-0,006	0,038	-0,020	0,119
SVRe $\alpha = 0.9$	0,228	0,099	0,427	0,281	0,105	0,209	0,457	-0,033	0,217	-0,036	0,554
SVRe $\alpha = 1$	0,203	0,080	0,378	0,290	-0,003	0,189	0,436	-0,036	0,176	-0,066	0,591
PCC_c	Mean	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU43
Kaltwang 2012 (LBP+LM)	0,306	0,260	0,508	0,276	0,339	0,312	0,545	0,095	0,213	0,118	0,396
Easy Ensemble	0,301	0,149	0,396	0,368	0,219	0,422	0,442	0,053	0,415	-0,019	0,569
SVRe $\alpha = 0$	0,286	0,163	0,494	0,395	0,158	0,226	0,542	0,089	0,250	0,107	0,432
SVRe $\alpha = 0.9$	0,311	0,200	0,509	0,344	0,210	0,285	0,524	0,128	0,326	0,089	0,492
SVRe $\alpha = 1$	0,301	0,202	0,472	0,363	0,154	0,263	0,513	0,141	0,305	0,084	0,515
MSE_c	Mean	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU43
Kaltwang 2012 (LBP+LM)	0,325	0,201	0,544	0,429	0,091	0,070	0,625	0,179	0,449	0,482	0,184
EasyEnsemble	0,827	0,459	1,052	0,475	0,760	0,213	0,981	0,731	0,558	2,979	0,058
SVRe $\alpha = 0$	0,231	0,155	0,492	0,305	0,071	0,065	0,559	0,064	0,265	0,291	0,041
SVRe $\alpha = 0.9$	0,467	0,379	0,824	0,600	0,252	0,154	0,742	0,300	0,541	0,824	0,052
SVRe $\alpha = 1$	0,508	0,391	0,949	0,604	0,345	0,171	0,790	0,312	0,597	0,868	0,052

Table 5: Performance comparison for each AU in the UNBC dataset (similar to Table 4a in the paper). SVRe $\alpha = 0$ corresponds to imbalanced Support Vector Regression Ensemble (SVRe), SVRe $\alpha = 1$ corresponds to balanced SVR Ensemble, and SVRe $\alpha = 0.5$ corresponds to MIDRUS SVR Ensemble.

5 Qualitative Observations

Action Unit intensity estimation is far away from being perfect. In this supplemental material we provide some challenging examples from which our approach still suffers. We identify the main challenges along with our interpretation about the cause of the problems.

5.1 Underestimation

Underestimation occurs much more often than overestimation. Especially for action unit intensities that appear infrequently underestimation is apparent. The intensity estimation for AU 15 and AU 20 are most difficult because they often appear very subtle. Most methods perform worst for these action units. Fig. 3 shows a typical example:

1. Almost the entire sequence is underestimated for both action units. Even though the first two frames show no appearance of any action unit nor head rotation or occlusion, the estimated intensity is still lower than the ground truth. Be belief that this is due to person specific attributes, e.g. natural low mouth corner positions etc.
2. Even though image 5 (at frame no. 1758) show strong lip corner depression (AU 15) and lip stretching (AU 20) the estimation barely reaches level 1. Because the different intensity levels of both action units appear rarely in the dataset, the underestimation is probably due to regression model limitations. We also observe that the corner detector often fails during strong facial actions. This can also be observed in image 5 and might contribute to the strong underestimation.

The estimation of action unit intensities with sufficient data examples in the training set (e.g. AU 25) is already very sophisticated. Fig. 4 shows the good estimation for AU 25 (lips part) with two incidents for which the estimation still struggles:

1. At the first image (at frame no. 2656) the lips are pressed together and the corresponding estimation drops into negative values. Our linear regression model underestimates the ground truth due to the very close landmarks.
2. Image 5 (at frame no. 2826) the estimation clearly outperforms the ground truth, since the mouth is slightly closing.

5.2 Overestimation

Even though underestimation is a more dominant problem, overestimation still occurs. E.g. our proposed approach overestimates the ground truth for subject 26, AU 25 as can be seen in Fig. 5. The example images with the detected landmarks provide several explanations:

1. Even though the mouth is completely closed between frame no. 730 and 973 the intensity is constantly overestimated. Especially image 3, 5, and 6 (at frame no. 834, 917, and 958, respectively) have no head rotation, misalignments, or occlusions but are still erroneous. Since this observation is apparent for almost all images for this subject, it is a person specific problem. We assume that this is caused by the slightly wider mouth and larger lips.
2. The first two images (at frame no. 752 and 793) show a smiling face (with closed mouth). Our approach (based on landmarks and LBP) predicts a slightly opened mouth (intensity reaching almost level 2). We think that because opened mouths often appear in the dataset along with smiles this conjunction is learned unfortunately.
3. Image 7 (at frame no. 999) is overestimated but image 8 (at frame no. 1040) is estimated perfectly. In comparison, image 7 has a slightly more opened mouth than image 8 but is also tilted a little. Although the labeling process was done carefully, we already experience its limitations.

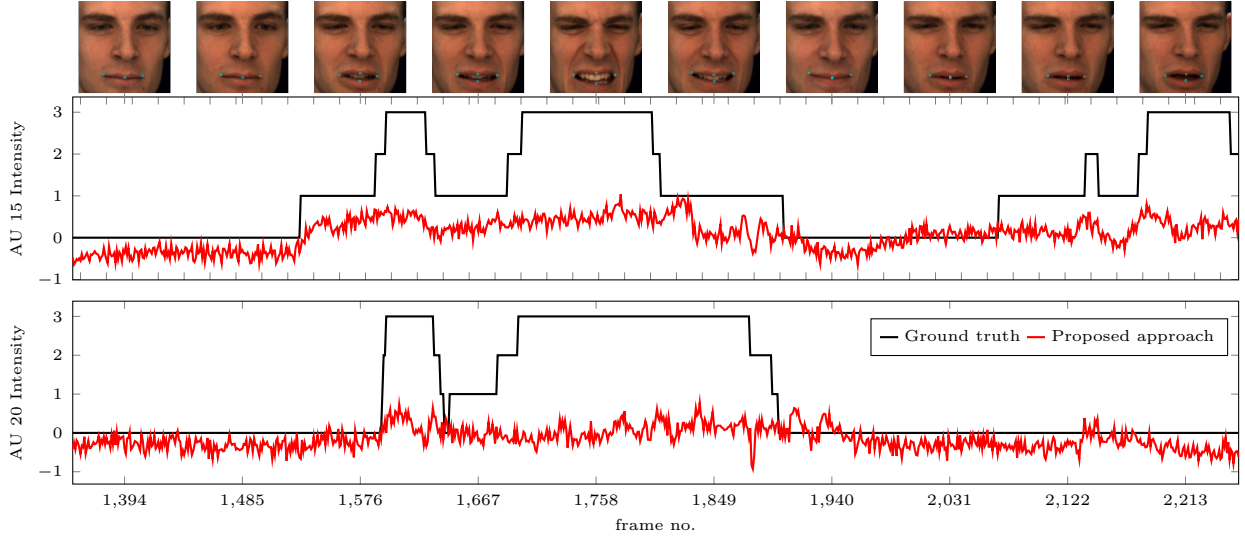


Figure 3: Intensity estimation on test set for AU 15 (lip corner depressor) and AU 20 (lip stretcher) shown for a sequence of frames from subject 28 (DISFA dataset). Our proposed approach (red) underestimates the ground truth (black) almost the entire sequence despite heavy facial activity.

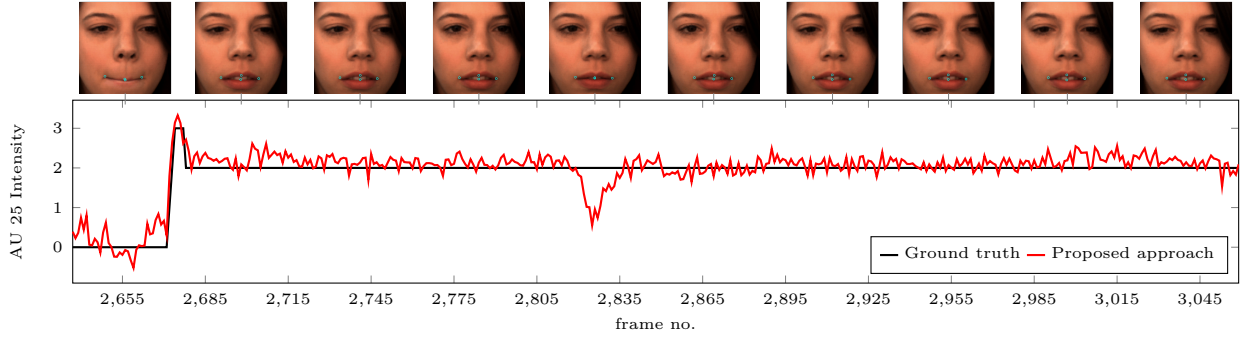


Figure 4: Intensity estimation on test set for AU 25 (lips part) shown for a sequence of frames from subject 25 (DISFA dataset). Our proposed approach, the MIDRUS SVR ensemble (red), underestimates the ground truth (black) for image 1 (at frame no. 2656) and image 5 (at frame no. 2826).

5.3 Challenges with the UNBC dataset

The unbc is exceptionally challenging compared to the bosporus and the disfa dataset. Fig. 6 shows a high intensity estimation error:

1. This sequence shows a strong out of plane rotation which is typical for the unbc dataset. Since we use an affine transformation for the face alignment, only the roll angle is compensated but not the yaw and pitch angle. This example however shows high variations in the yaw angle of the head. The texture features suffer from the large background areas and the different perspectives due to the bad alignment. We assume that the performance for the unbc dataset can be improved by using a more appropriate transform for the face alignment.
2. This sequence also shows several false label examples. In the last two example images (between frame no. 1305 and 1357) the mouth is clearly opened. Such labeling errors are also apparent for other participants and action units.

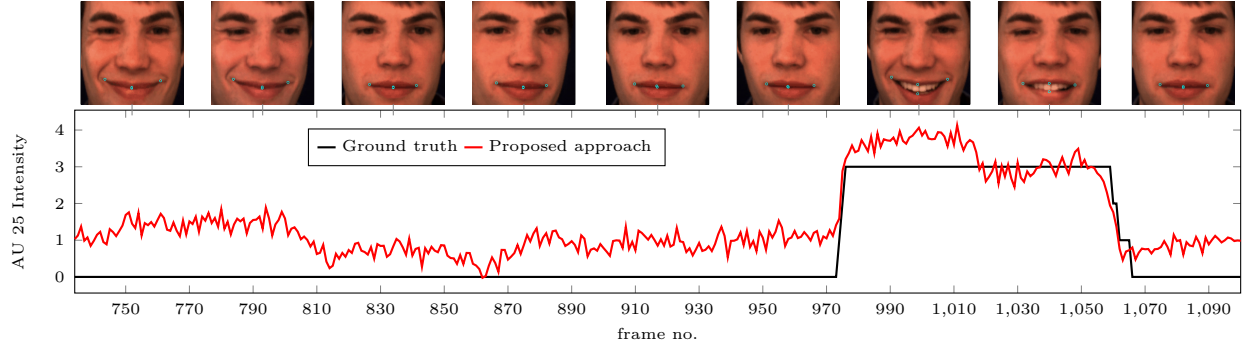


Figure 5: Intensity estimation on test set for AU 25 (lips part) shown for a sequence of frames from subject 26 (DISFA dataset). Our proposed approach (red) overestimates the ground truth (black) for this specific subject.

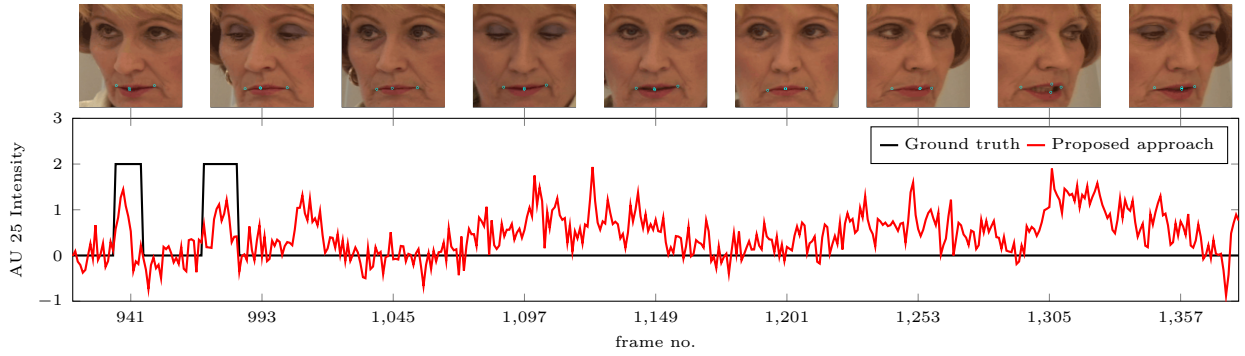


Figure 6: Intensity estimation on test set for AU 25 (lips part) shown for a sequence of frames from subject 47 (UNBC dataset). Our proposed approach (red) shows a high estimation error compared to the ground truth (black).