

# AST-Net: An Attribute-based Siamese Temporal Network for Real-Time Emotion Recognition (Supplementary Material)

BMVC 2017 Submission # 299

## 1 AVEC2012 Dataset

### 1.1 Subjects

Figure 1 shows all the subjects in AVEC2012 dataset. As shown in Fig. 1, except the three subjects (highlighted by a blue rectangle), the other subjects in the development and test sets are different from those in the training set. Because the training set contains only 7 subjects, it is difficult to learn a discriminative representation and an effective prediction model from this small-scaled dataset without suffering the over-fitting problem.

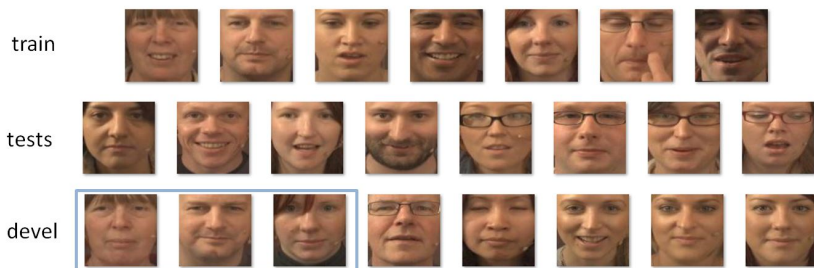


Figure 1: All the subjects in AVEC2012 dataset, where only the three highlighted subjects appear in both the training and testing phases.

### 1.2 Subtle emotional changes in short duration

Figure 2 shows that, there is usually very little emotional change between successive frames in the original video. Therefore, we will not be able to learn informative temporal dependency from the original videos. By contrast, in Fig. 3, the sampled video captures subtle emotion change among successive frames and should better serve as the training set to learn the temporal model.

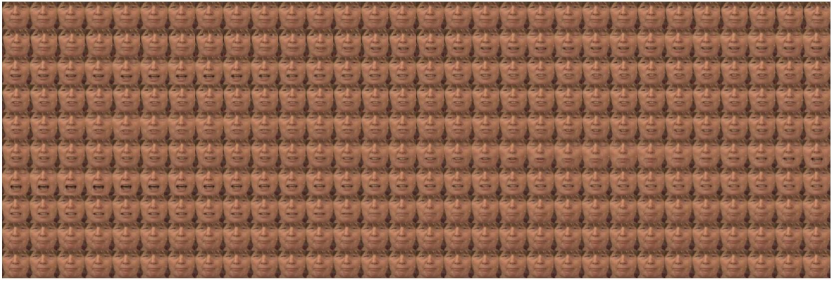


Figure 2: A sequence of 300 continuous frames in the original video.



Figure 3: The set of sampled frames that covers the same temporal duration of 300 frames in the original video.

## 2 Experimental Results

### 2.1 Evaluation of feature invariance

There exist various facial variations that are unrelated to the emotional changes (e.g. individual characteristics, ethnic, illumination changes, and poses) in AVEC2012 dataset. From Fig. 1, we have seen that there is little overlap of subjects between the training and test sets. Thus, the proposed representation indeed captures identity-invariant features. Below we will show more examples to demonstrate that the learned features are also invariant to other variations, such as poses and wearing glasses.

#### 2.1.1 Poses

Figures 4, 5, 6, and 7 show that, even after face alignment, there still exist large pose variations in these videos. Some of the frames also suffer from the partial occlusion issue (e.g., occluded by moving hands). Even under these challenging cases, the proposed AST-Net still successfully predicts the dimensional emotion change with high accuracy.

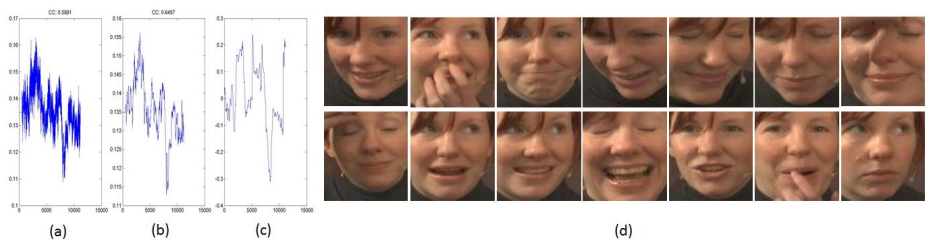


Figure 4: Video011 in AVEC2012 development set. (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.

092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137



Figure 5: Video001 in AVEC2012 test set. (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.

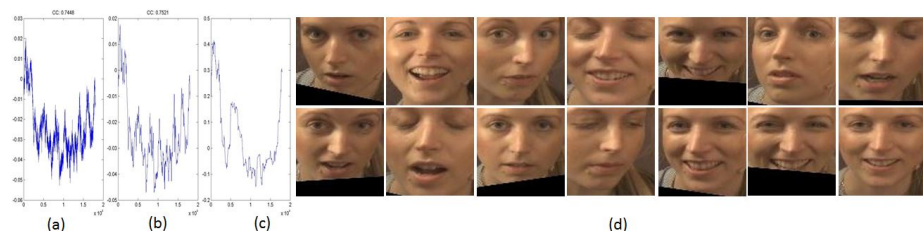


Figure 6: Video022 in AVEC2012 development set. (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.

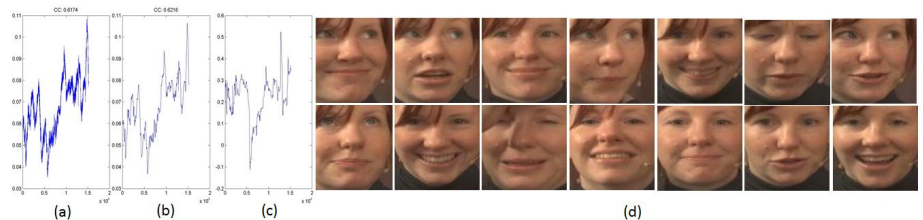


Figure 7: Video012 in AVEC2012 development set. (a) and (b): The prediction of arousal before and after median filtering; (c): Label of arousal; and (d): Some sample frames.

### 2.1.2 Glasses

In Figs. 8, 9, 10, 11, we show some examples with subjects wearing glasses. The videos in Figs. 8, 9, and 11 also contain large pose variation and partial occlusion. Our prediction results in these cases again verify that the proposed model is invariant to these variations.



Figure 8: Video013in AVEC2012 development set. (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.



Figure 9: Video017in AVEC2012 test set. (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.



Figure 10: Video025 in AVEC2012 test set. (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.

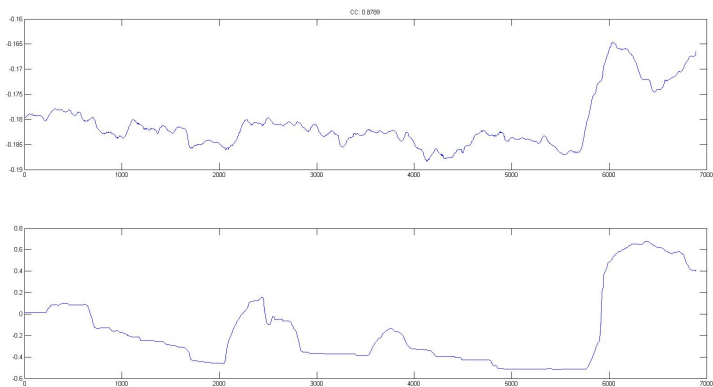
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
\*/183



192 Figure 11: Video026 in AVEC2012 test set. (a) and (b): The prediction of valence before  
193 and after median filtering; (c): Label of valence; and (d): Some sample frames.

## 195 2.2 More results

196 More experimental results are given below. In each of the figures, the first row shows the  
197 predictions and the second row is ground-truth labels.



214 Figure 12: Predictions of Valence to Video001 in AVEC2012 development set.

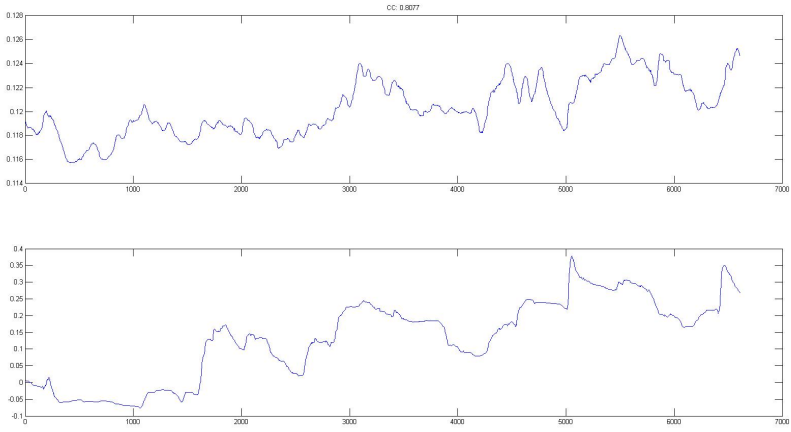


Figure 13: Predictions of Valence to Video005 in AVEC2012 development set.

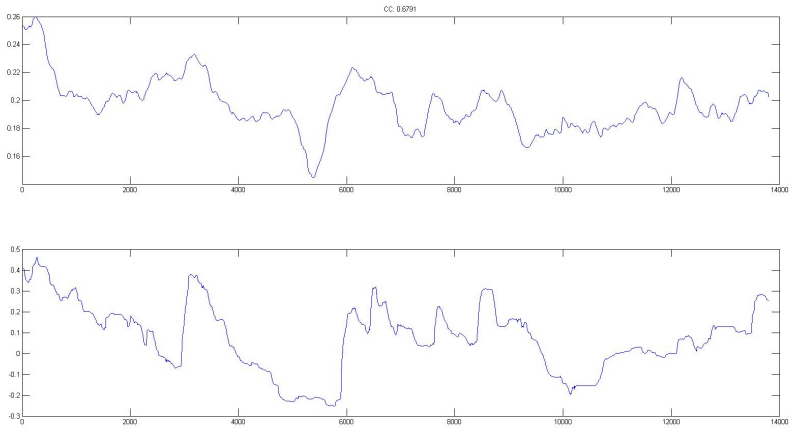


Figure 14: Predictions of Valence to Video010 in AVEC2012 development set.

230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
\*/275

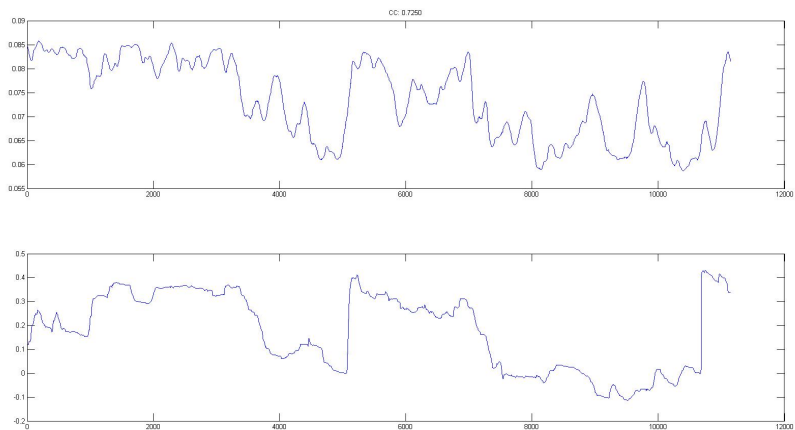


Figure 15: Predictions of Valence to Video011 in AVEC2012 development set.

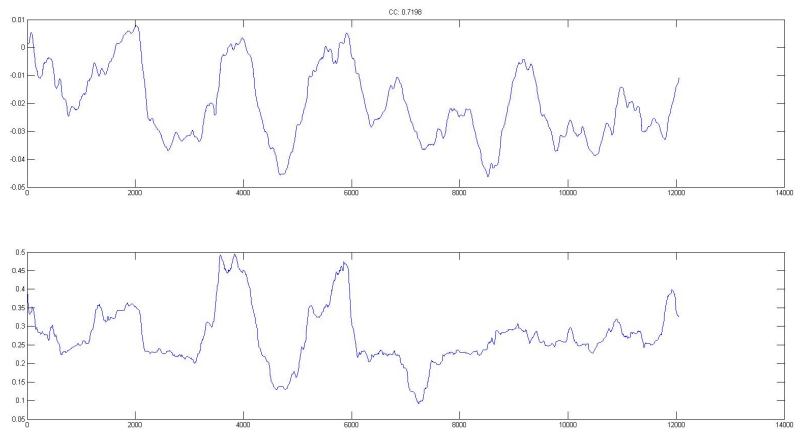


Figure 16: Predictions of Valence to Video021 in AVEC2012 development set.

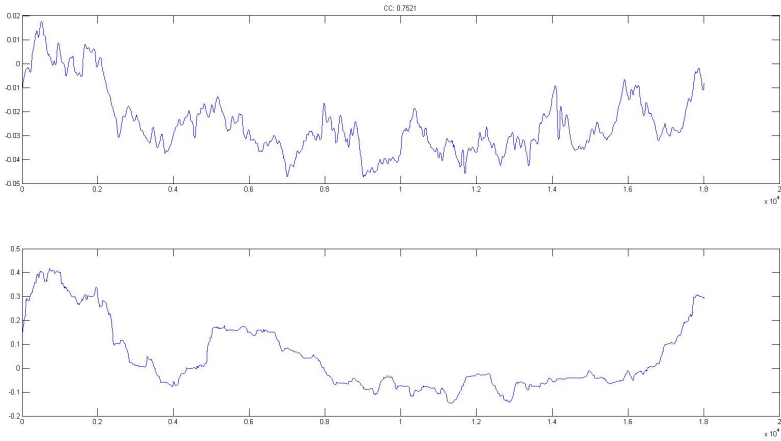


Figure 17: Predictions of Valence to Video022 in AVEC2012 development set.

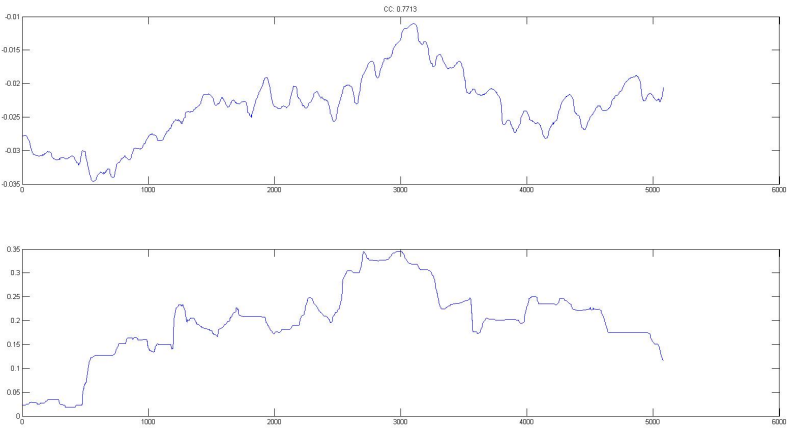


Figure 18: Predictions of Valence to Video028 in AVEC2012 development set.

322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
\*/367



368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413

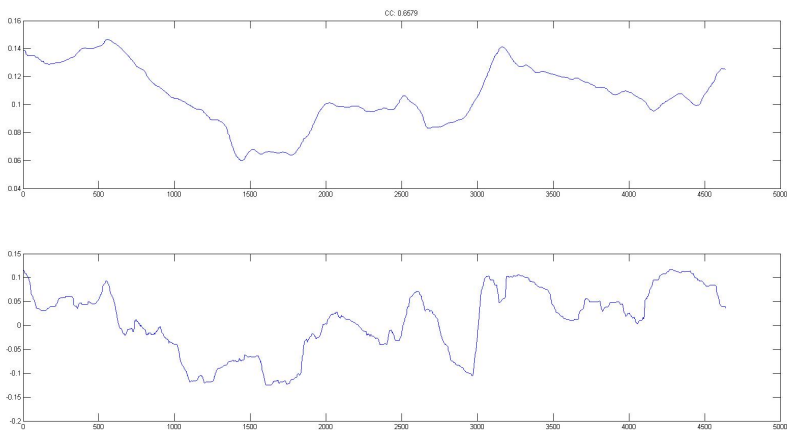


Figure 19: Predictions of Valence to Video001 in AVEC2012 test set.

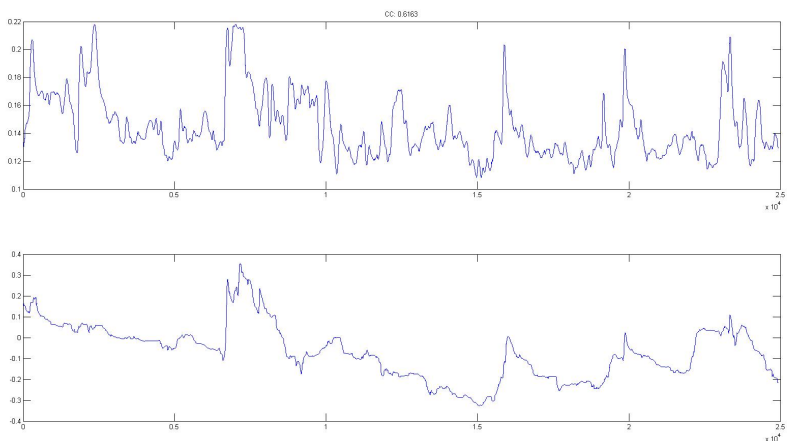


Figure 20: Predictions of Valence to Video007 in AVEC2012 test set.

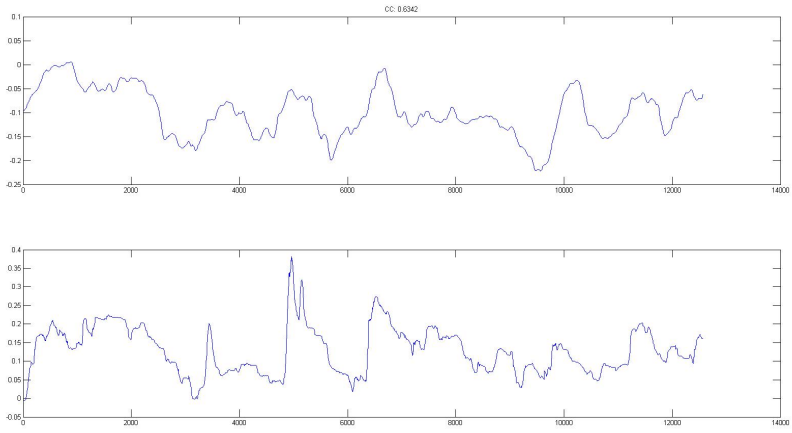


Figure 21: Predictions of Valence to Video017 in AVEC2012 test set.

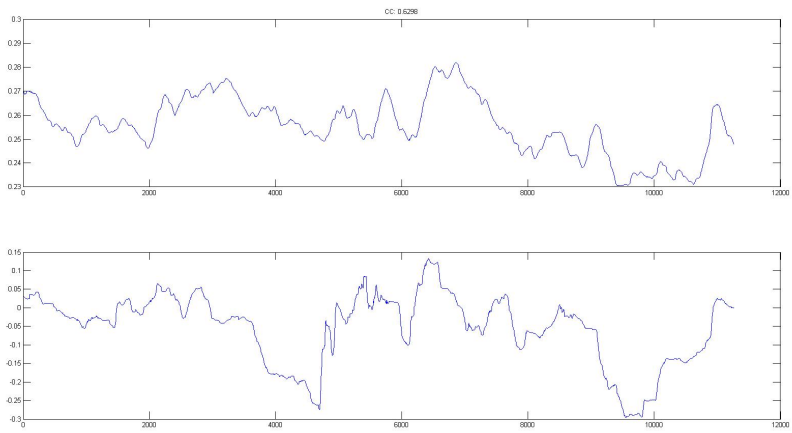


Figure 22: Predictions of Valence to Video019 in AVEC2012 test set.

414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
\*/459

460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505

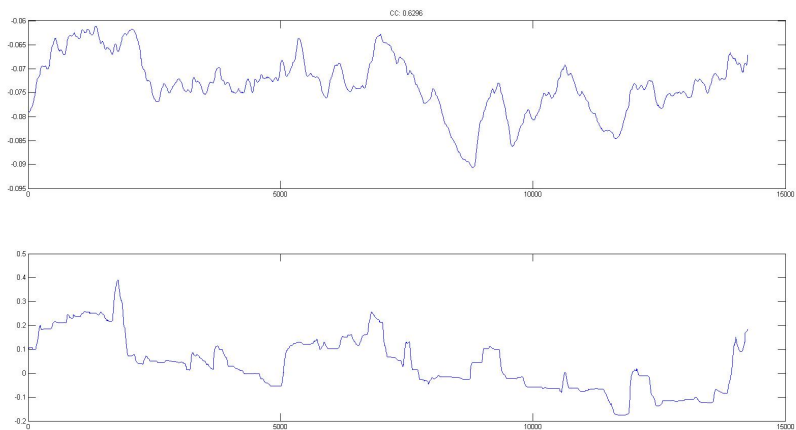


Figure 23: Predictions of Valence to Video026 in AVEC2012 test set.

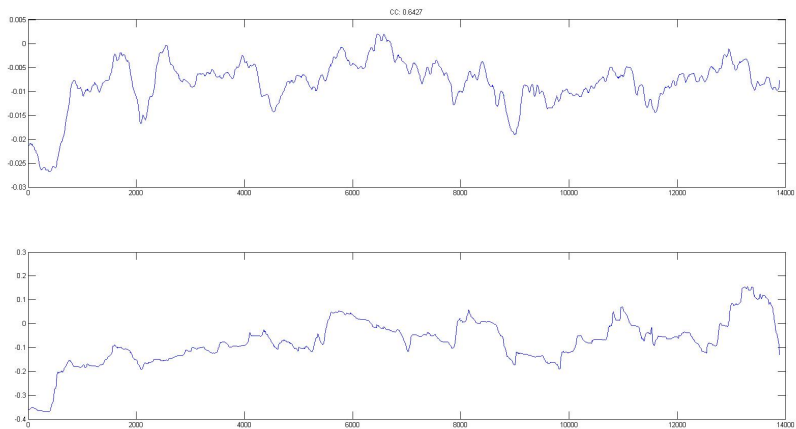


Figure 24: Predictions of Arousal to Video007 in AVEC2012 development set.

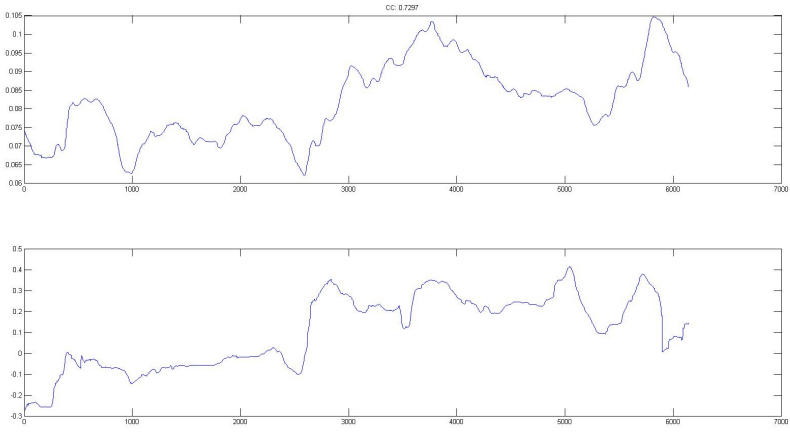


Figure 25: Predictions of Arousal to Video013 in AVEC2012 development set.

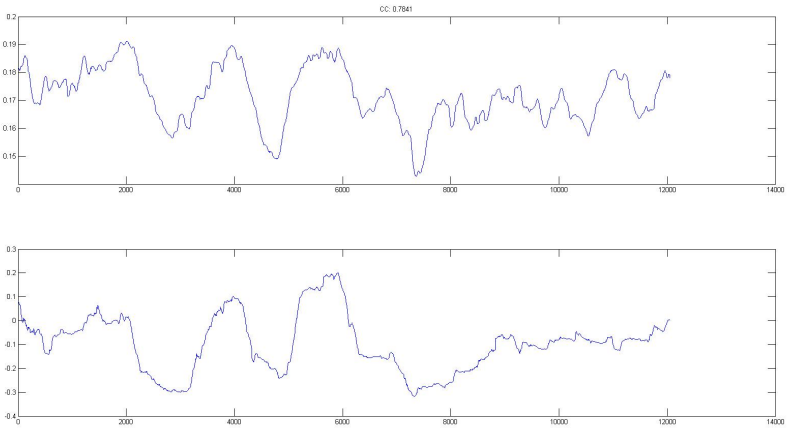


Figure 26: Predictions of Arousal to Video021 in AVEC2012 development set.

506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
\*/551

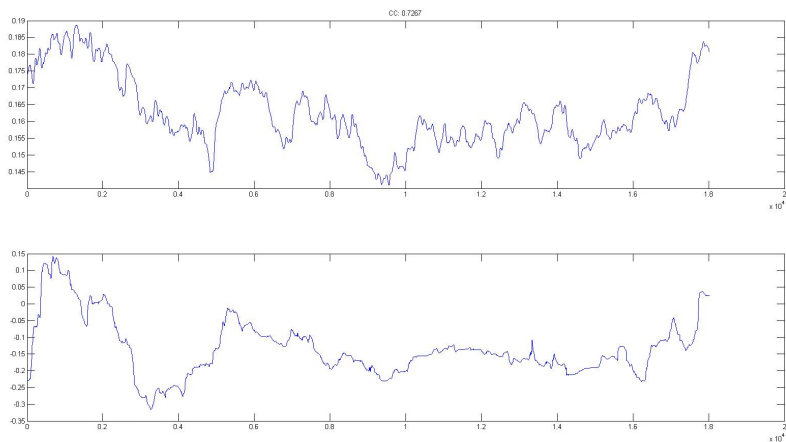


Figure 27: Predictions of Arousal to Video022 in AVEC2012 development set.

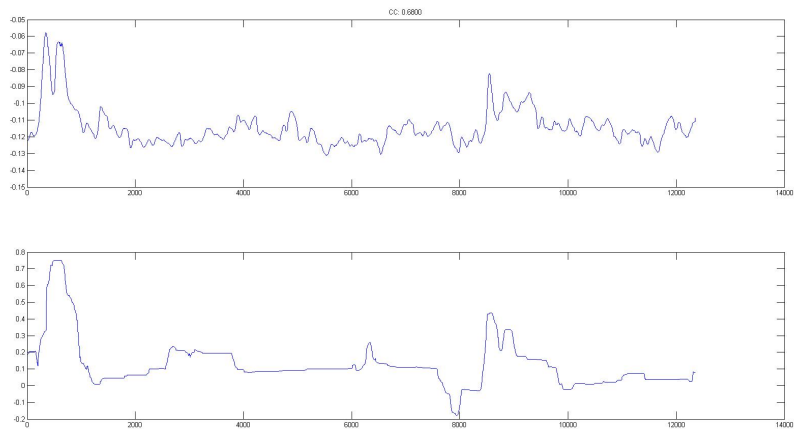


Figure 28: Predictions of Arousal to Video025 in AVEC2012 development set.

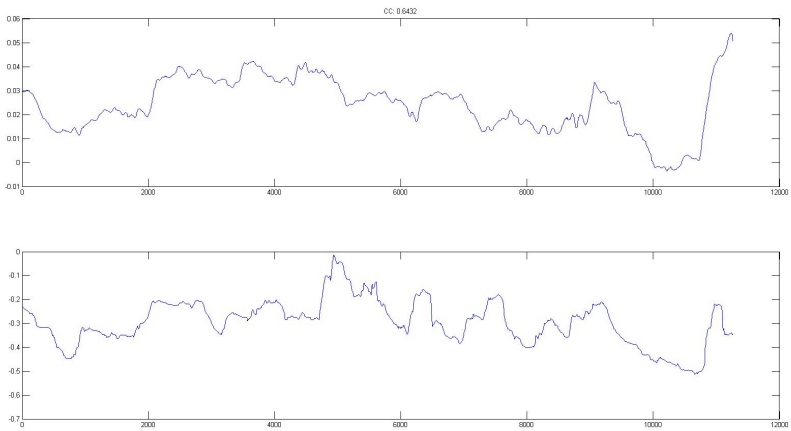


Figure 29: Predictions of Arousal to Video019 in AVEC2012 test set.

598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
\*/643