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.

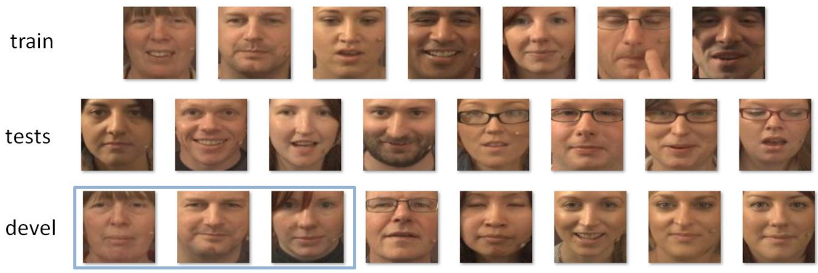

Figure 1: All the subjects in AVEC2012 dataset, where only the three highlighted subjects appear in both the training and testing phases.

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.

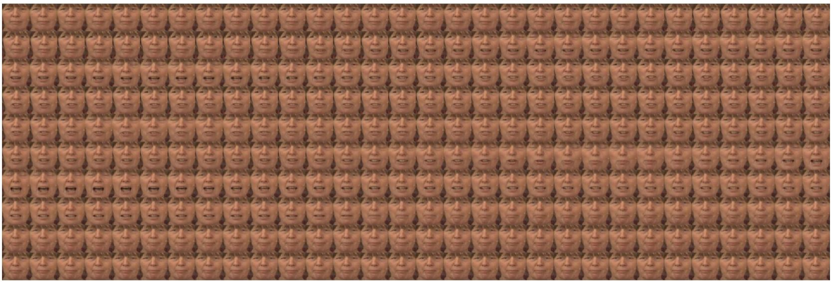

Figure 2: A sequence of 300 continuous frames in the original video.

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.

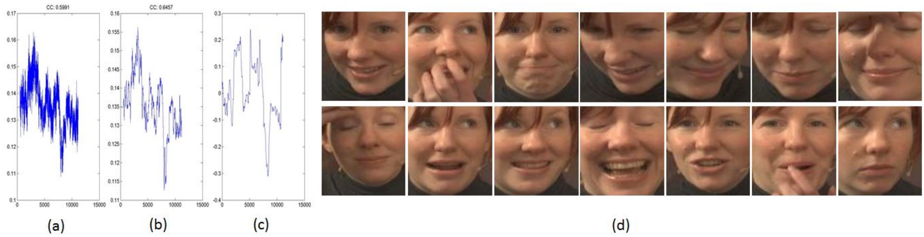

Figure 4: Video011 in AVEC2012 development set. (a) and (b): The prediction of valence before and after median filtering; (c): Label of valence; and (d): Some sample frames.

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.

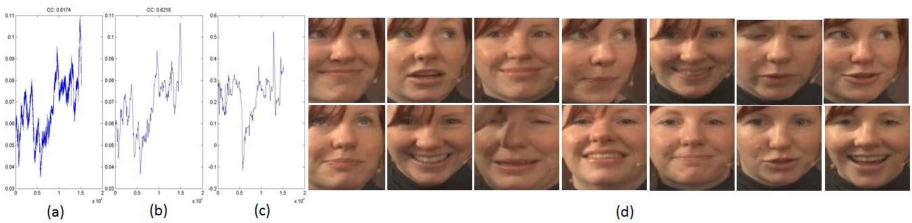

Figure 7: Video012in AVEC2012 development set. (a) and (b): The prediction of arousal before and after median filtering; (c): Label of arousal; and (d): Some sample frames.

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.

Figure 11: Video026 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.

2.2 More results

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

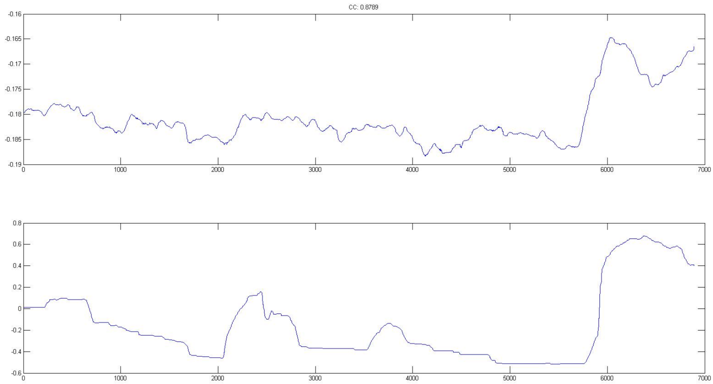

Figure 12: Predictions of Valence to Video001 in AVEC2012 development set.

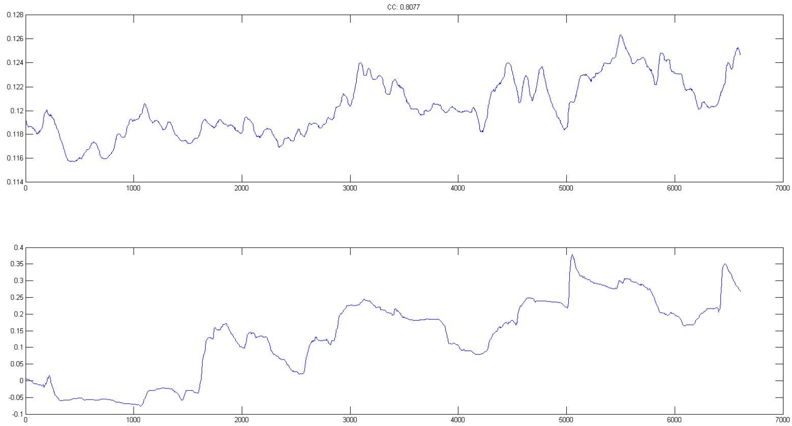

Figure 13: Predictions of Valence to Video005 in AVEC2012 development set.

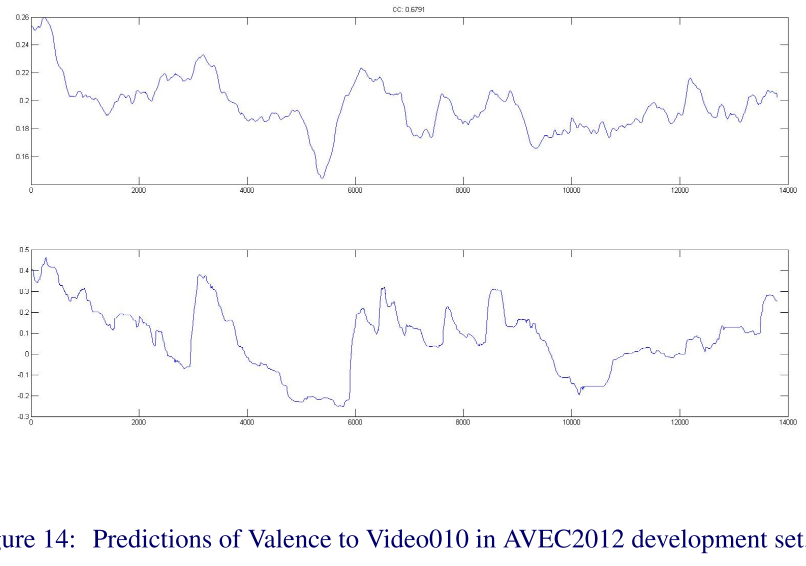

Figure 14: Predictions of Valence to Video010 in AVEC2012 development set.

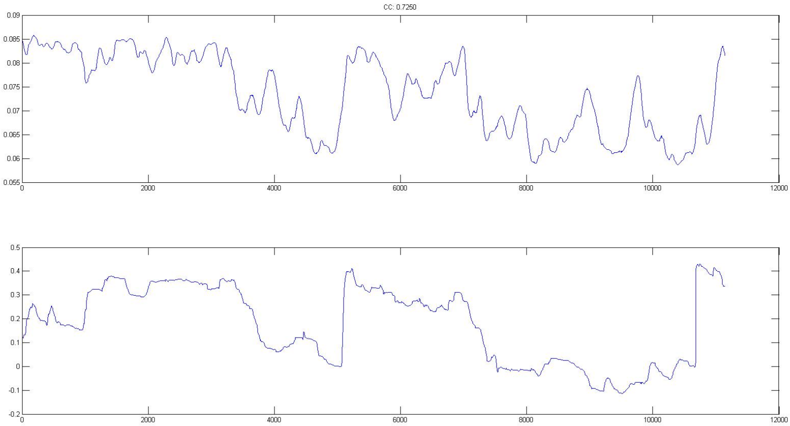

Figure 15: Predictions of Valence to Video011 in AVEC2012 development set.

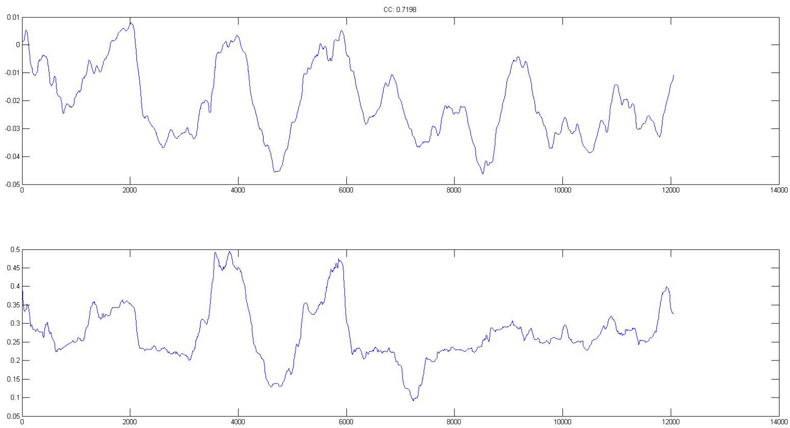

Figure 16: Predictions of Valence to Video021 in AVEC2012 development set.

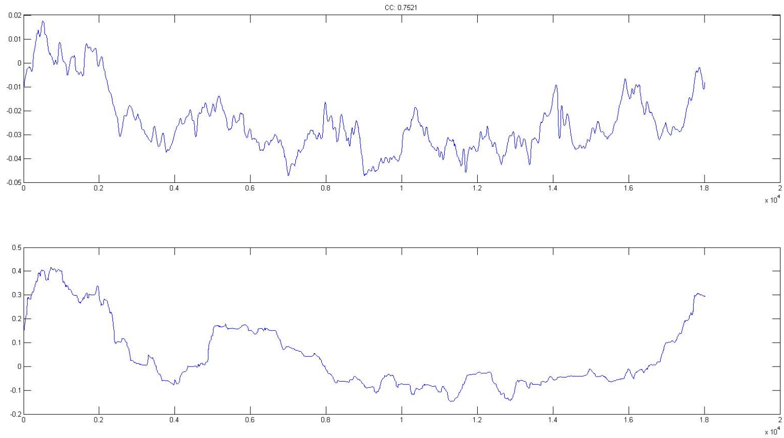

Figure 17: Predictions of Valence to Video022 in AVEC2012 development set.

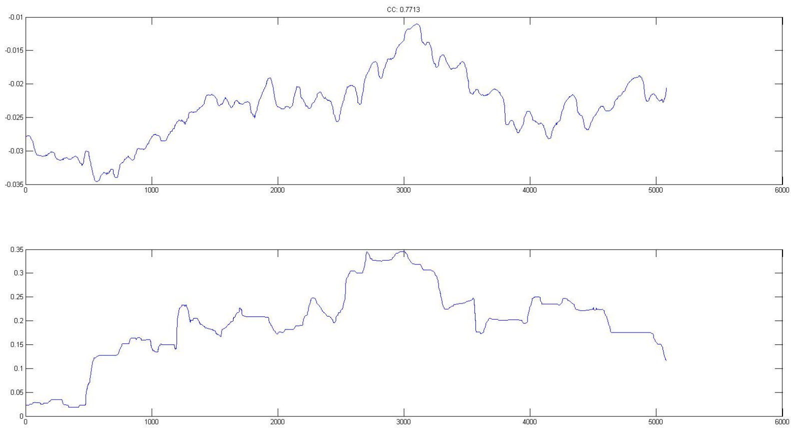

Figure 18: Predictions of Valence to Video028 in AVEC2012 development set.

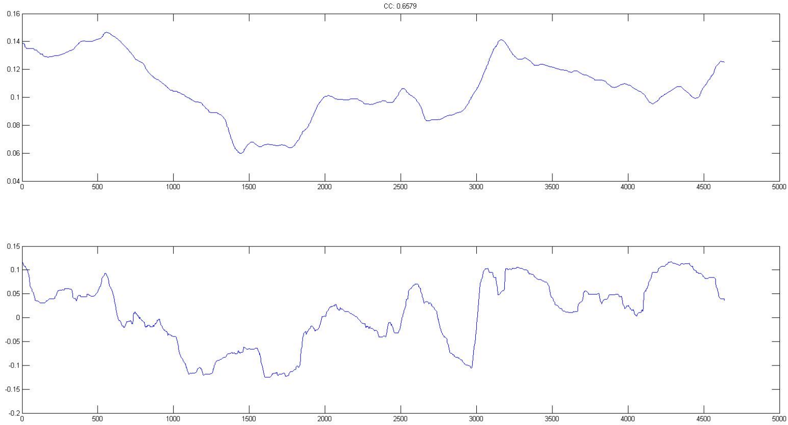

Figure 19: Predictions of Valence to Video001 in AVEC2012 test set.

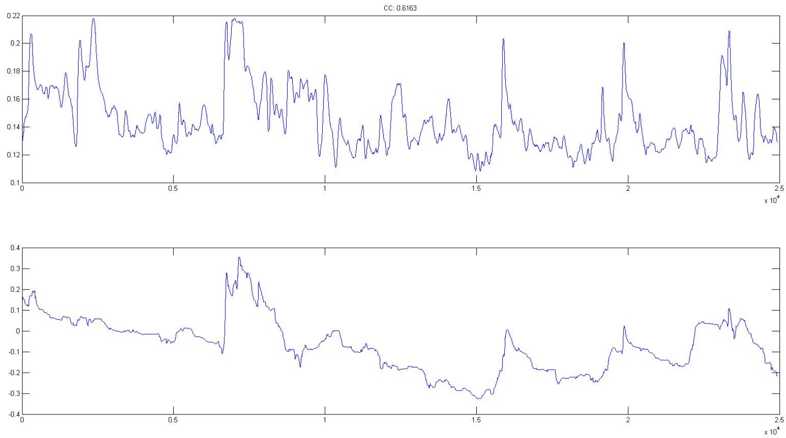

Figure 20: Predictions of Valence to Video007 in AVEC2012 test set.

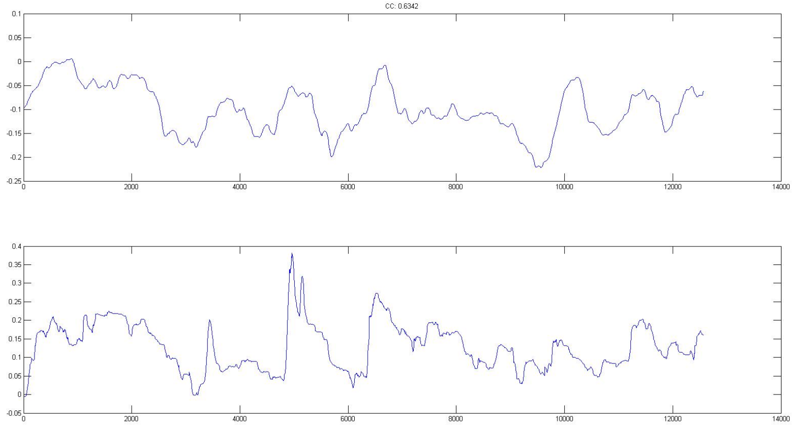

Figure 21: Predictions of Valence to Video017 in AVEC2012 test set.

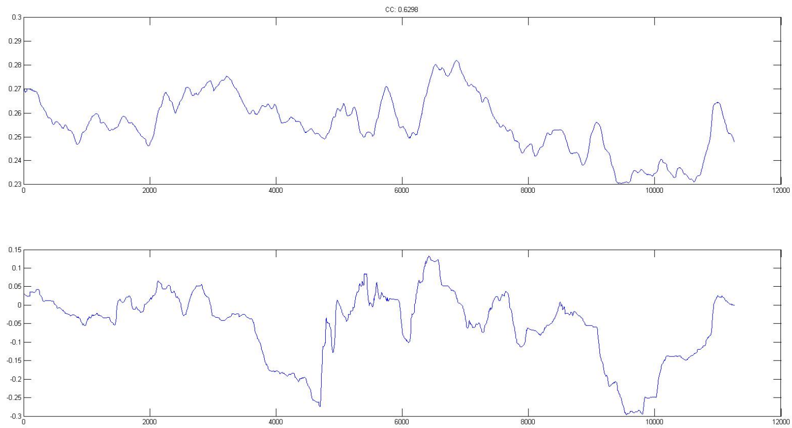

Figure 22: Predictions of Valence to Video019 in AVEC2012 test set.

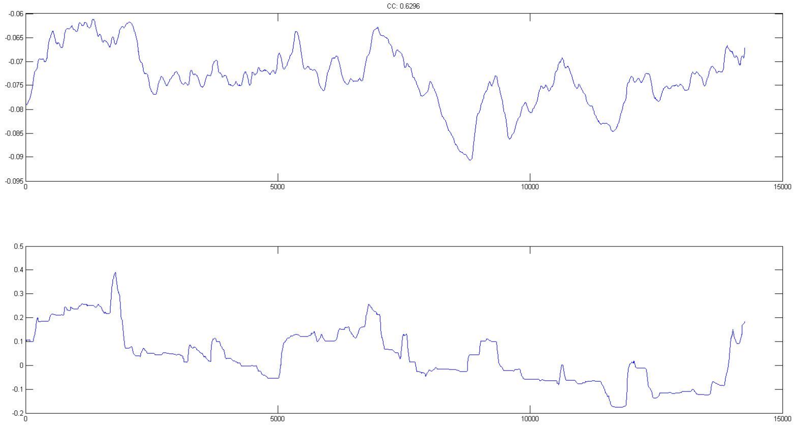

Figure 23: Predictions of Valence to Video026 in AVEC2012 test set.

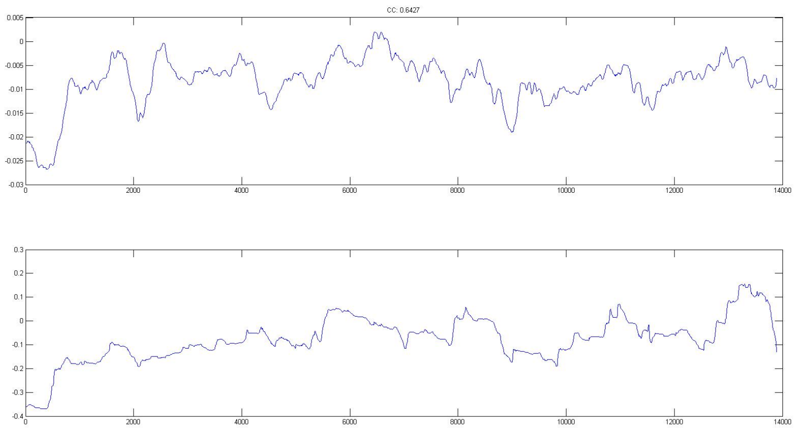

Figure 24: Predictions of Arousal to Video007 in AVEC2012 development set.

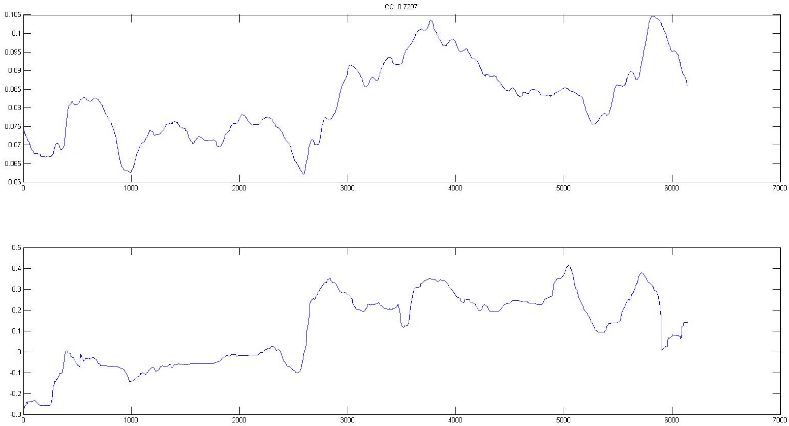

Figure 25: Predictions of Arousal to Video013 in AVEC2012 development set.

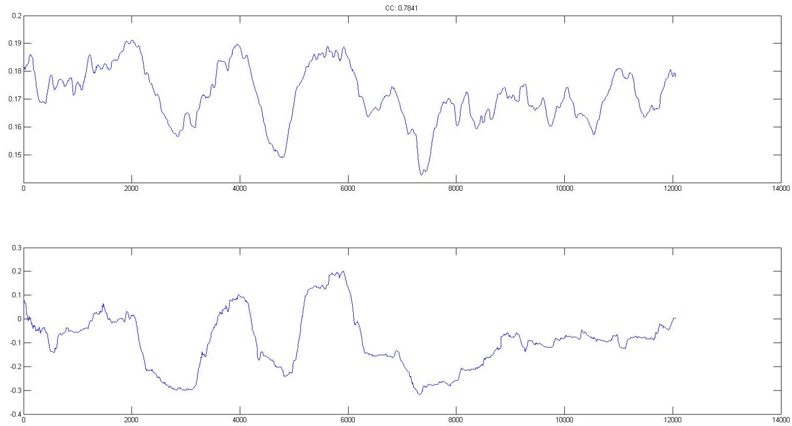

Figure 26: Predictions of Arousal to Video021 in AVEC2012 development set.

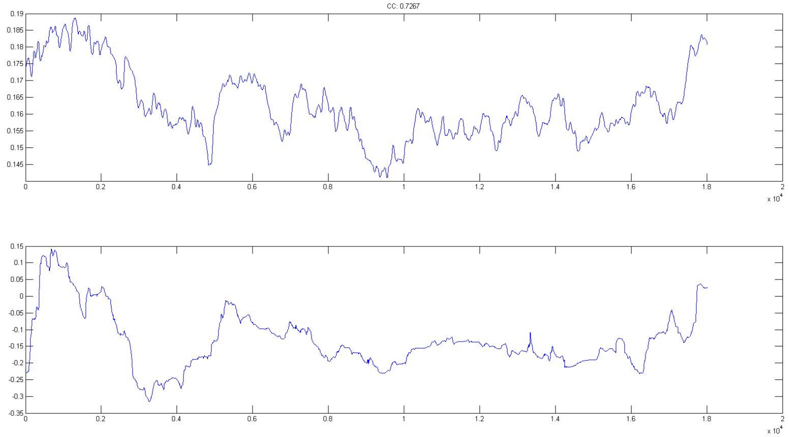

Figure 27: Predictions of Arousal to Video022 in AVEC2012 development set.

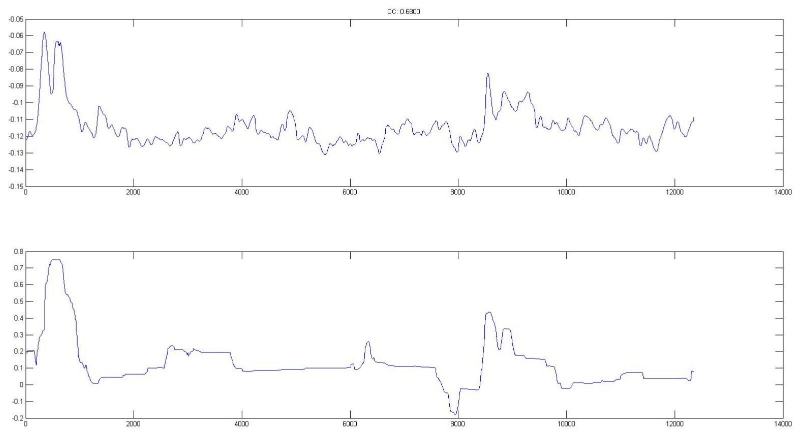

Figure 28: Predictions of Arousal to Video025 in AVEC2012 development set.

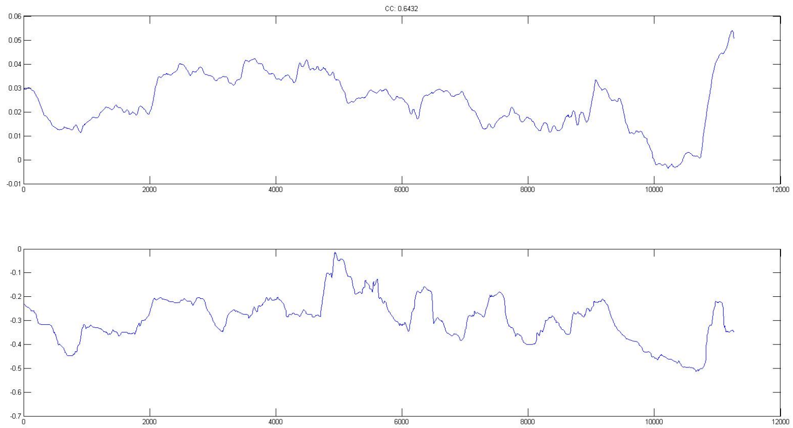

Figure 29: Predictions of Arousal to Video019 in AVEC2012 test set.