A Spatio-Temporal Probabilistic Framework for Dividing and Predicting Facial Action Units

A.K.M. Mahbubur Rahman, Md. Iftekhar Tanveer, and Mohammed Yeasin

Electrical and Computer Engineering, The University of Memphis

Abstract. This paper proposed a probabilistic approach to divide the Facial Action Units (AUs) based on the physiological relations and their strengths among the facial muscle groups. The physiological relations and their strengths were captured using a Static Bayesian Network (SBN) from given databases. A data driven spatio-temporal probabilistic scoring function was introduced to divide the AUs into : (i) frequently occurred and strongly connected AUs (FSAUs) and (ii) infrequently occurred and weakly connected AUs (IWAUs). In addition, a Dynamic Bayesian Network (DBN) based predictive mechanism was implemented to predict the IWAUs from FSAUs. The combined spatio-temporal modeling enabled a framework to predict a full set of AUs in real-time. Empirical analyses were performed to illustrate the efficacy and utility of the proposed approach. Four different datasets of varying degrees of complexity and diversity were used for performance validation and perturbation analysis. Empirical results suggest that the IWAUs can be robustly predicted from the FSAUs in real-time and was found to be robust against noise.

Keywords: Affective computing, Spatio-Temporal AU relations.

1 Introduction

Autonomous analysis and synthesis of facial expressions and emotions are emerging issues in affective computing and agent-human communication. Facial action units (AUs) defined in Facial Action Coding System (FACS) [2] has been widely used in recognizing facial expressions (i.e., [12]), emotions (i.e., [14]), and affective states [8] to compute description of facial behavior. Despite the recent surge of computational methods, robust and real-time recognition of AUs remains challenging due to inaccuracies in measurements of subtle facial deformation and pose.

State-of-the art methods in recognition of AUs are limited to subset of *posed expressions* that are inadequate for recognition of spontaneous facial expressions, modeling blended emotions and also unsuitable for real-time applications. For example, the number of AUs recognized by Tong *et al.* [12], Lucey *et al.* [5], Zhang *et al.* [15], and Bartlett *et al.* [1] were 14, 17, 18, and 20, respectively. The choices of subset of AUs in the reported literatures were done mostly by *adhoc*-principles for a variety of reasons (that include but are not limited to): (i) skewed and non-uniform distribution of "representative examples" in existing

S. D´Mello et al. (Eds.): ACII 2011, Part II, LNCS 6975, pp. 598–607, 2011.

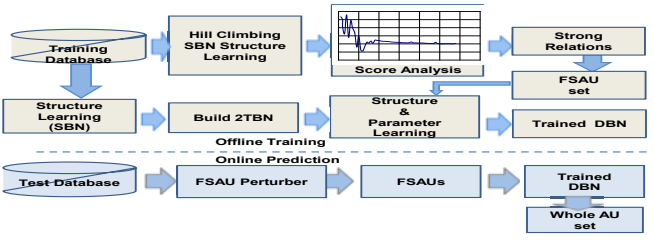

Fig. 1. Proposed Approach

emotion databases [5], (ii) lack of systematic approaches to determine significant subset of AUs even in the context of a niche application to characterize facial behavior, and (iii) lack of framework for real-time processing of full set of AUs. It is easy to note that methods based on *heuristics* and *adhoc* rules preclude important relationships between AUs.

The problems mentioned above can be addressed by logically dividing the AUs into two subsets: **f**requently occurred and **s**trongly connected **AU**s (FSAUs) and **i**nfrequently occurred and **w**eakly connected **AU**s (IWAUs). By exploiting the physiological constraints, the IWAUs can be predicted from the FSAUs in real time. In addition to physiological constraints, AUs are evolved over time when facial expressions evolves from onset to apex to offset. By modeling both the spatial and temporal relations, it is possible to rely on a smaller significant subset of AUs to infer the occurrences of other AUs.

1.1 Proposed Approach

This paper introduces a probabilistic mechanism that captures the spatio-temporal evolution of AUs to logically divide them into FSAUs and IWAUs and to predict the IWAUs from the FSAUs in real-time. The full set containing (m) AUs is divided into two subsets: (i) subset $P = \{p_1, p_2, \ldots, p_n\}$ containing FSAUs and (ii) subset $S = \{s_1, s_2, \ldots, s_{m-n}\}$ containing IWAUs. The key objective is to keep the size of P as small as possible while maintaining robustness in inferring the set S. The spatial relations and physiological constraints among the AUs were modeled and synthesized with SBN while their temporal evolution was modeled using DBN. The scoring mechanism was defined using an optimized SBN that captures the strength of AU relations learned from the database. The figure 1 shows the conceptual framework for proposed solution.

1.2 How Is It Different?

Adhoc selection of AUs [12], [6],[15] bypass the problem of finding of significant AUs. A number of closely related works (i.e., [15,12,13]) use Bayesian analysis with frequent AUs. The proposed approach is significantly different from these techniques in a number of ways. The key difference is the use of AU relations

obtained using the "scoring mechanism" as opposed to "frequency" that clearly separates this work from reported related works. Since the combination of AUs and their temporal evolutions are mostly responsible for spontaneous facial behavior, the proposed solution incorporates a spatio-temporal statistical approach to capture the *AU relations* from the evidences. The scoring process defined using the SBN is novel (to the best of our knowledge) though it uses existing tools and was found to be very robust. Moreover, the prediction performance of IWAUs outperformed the contemporary related works.

Additionally, perturbation (noise) analysis was performed at AU level as well as at the level of categorical emotion recognition. At first, robustness in predicting IWAUs against perturbation in the FSAUs were analyzed. It was observed that the proposed approach is stable in the presence of varying degrees of perturbation in the FSAUs while maintaining reasonable IWAU recognition performance. However, different IWAUs have different level of noise tolerance. Also, empirical analysis was performed to characterize the effect of perturbation at FSAUs on the robust prediction of facial expressions.

2 Analysis of AU Relations

The relations between AUs are functions of physiological constraints as well as facial expressions. Physiological constraints that are resulted from the anatomy of the human face are critical for analyzing relations between AUs. A number of examples of physiological constraints are described in brief for the sake of clarity. AU 15 pulls the corners of the lips down that have been affected by AU 17 (Chin Raiser). Both AUs are connected through *Orbicularis oculi* and *Mentalis* resulting in a stronger relation. Conversely, *Zygomaticus Minor* and *Risoricus* are not connected to each other but the muscles *Zygomaticus*, *Orbicularis Oculi*, and *Masseter* are responsible for weaker relation between AU 11 and AU 20. Relations between AUs are also functions of facial expressions. Though AU 6 and AU 12 are related to the different facial regions, they are involved in happy expressions frequently resulting strong relation among themselves. Relations between AU 2 (Outer brow raiser) and AU 27 (mouth stretch) possess a mixture of the both kinds of relations.

Using a linguistic analogy, one can define the AU tree and the root based on frequently occurred strong relations that are present in the expressions database(s). The main objective here is to separate the strong relations between AUs from weaker ones.

3 Identify Strong AU Relations

We developed a proposition to identify strong relations between AUs based on a scoring mechanism while empty SBN is used as the initial network in the hill climbing algorithm. For detail proof and derivation of the proposition, please refer to [9].

Proposition: Strong relations are modeled in the earlier iterations while weak relations are modeled at the later iterations of the SBN structure learning process using hill climbing algorithm.

3.1 Strong Relations vs. Weak Relations

Following the proposition, the entire structure learning process is divided into two parts, where the first part ("Buildup Area") contains iterations which are involved to model stronger edges and the later part "Tuning Area" is responsible for modeling weak relations. In the "Buildup Area", the stronger AU relation increases the score at a very high rate of change albeit one at a time. Therefore, the main portion of score buildup has occurred in "Buildup Area". In "Tuning Area", the structure is being tuned with weaker edges.

Initially, Extended Cohn-Kanade dataset (CK+) [5] has been used to build the SBN structure to identify the strong relations. 23 AUs are used in the experiments. In figure 2, before the 8^{th} iteration, rates of change of network score are high. Alternately, after the 8^{th} iteration, rates of change of score are very low. A sharp transition is observed at iteration 8. Using the proposition, we would say that up to iteration 8, all strong relations are modeled in SBN structure. Weaker relations are added in later iterations. Therefore, the SBN structure shown in figure 3 obtained at iteration 8 gives us all strong or frequent edges.

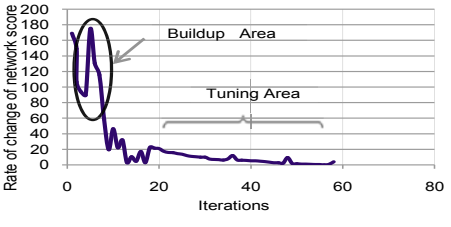

Fig. 2. Rate of Change of scores

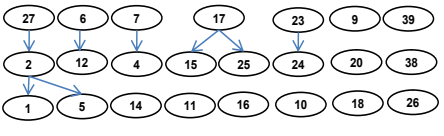

Fig. 3. Strong relations between AUs

3.2 FSAUs

Figure 3 shows that particular AUs are involved in building the strong relations. The responsible AUs to build strong relations are defined as FSAUs as they are the building blocks of root relations. Now, the right side in figure 4 shows the final structure of SBN where boxed nodes represent FSAUs. The status of node 1 and node 5 do not affect any other nodes. Consequently, AU 1 and AU 5 are removed from FSAU set and are considered as members of IWAUs.

3.3 Temporal Relationships among AUs

In spontaneous behavior, facial expressions involves a muscle group or a subset of action units that evolves over time.

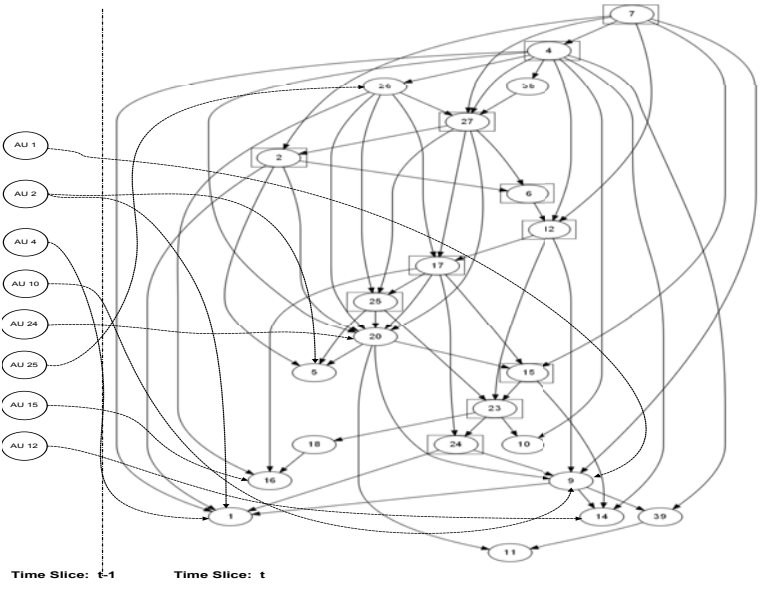

Fig. 4. Two Slice DBN

Modeling Temporal Evolution of AUs with DBN: A 2TBN is a very good framework with two connected SBN to represent the temporal evolution where a AU node in time slice t depends on AUs at time $t-1$ as well as other AUs at time t. Figure 4 shows the final 2TBN where temporal edges from time slice $t-1$ to time slice t represent the temporal evolution. Temporal edges between same IWAUs are not shown here. In inference procedure, nodes for corresponding FSAUs (boxed nodes) act as observed nodes while IWAUs are hidden.

4 Experimental Results

Empirical analyses using a number of different datasets consisting of varying degrees of variabilities were used to illustrate the utility of the proposed approach. In particular, four different datasets, **CK+** dataset (posed expressions [5]), Bosphorus dataset (posed 3D expressions [10]), M and M Initiative (**MMI**) dataset (mixture of posed and natural expressions [7]), and Emotion Elicitation (**EE**) dataset (natural expressions [14]) were used for empirical analyses.

4.1 Utility of Scoring Algorithm in Logical Division of AUs

It is widely acknowledged that the FACS provides a mechanism for analysis and synthesis of facial behavior that is consistent across culture, ethnicity, gender, and age groups. Hence, relations among the AUs are expected to be similar

across datasets. The scoring proposition described in the Section 3 was used to compute the relations among AUs. The trends in the rate of change of scores over iterations were used as an indicator to divide the AUs into FSAUs and IWAUs. Figure 5 shows the 2^{nd} derivative of the scores over iterations computed during the SBN structure learning process. From the figure 5 it is easy to note that the "Buildup Area" and "Tuning Area" were divided after 8^{th} iteration. All four datasets were used in the experimentation to check consistency in dividing the AUs into FSAUs and IWAUs. It was observed that the strong relations were found after 8^{th} iteration - these were **identical** across the datasets.

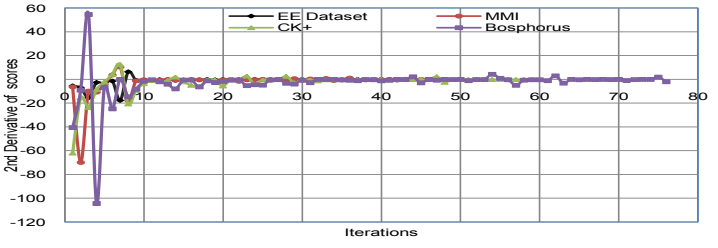

Fig. 5. 2^{nd} derivative of scores

AUs	ROCA
2	.975
4	.925
6	.975
7	.925
12	.975
15	.925
17	.95
23	.85
24	.92
25	.96
27	.99

Fig. 6. ROCA of FSAUs in [11]

4.2 IWAUs Predictions

Three different validation experiments were performed. Firstly, performance evaluation and comparison with recently reported literature were performed using the CK+ dataset. Secondly, generalization of the proposed approach was tested using MMI dataset while the DBN had been trained with the CK+. Furthermore, EE dataset [14] was used to illustrate the suitability of the proposed approach in dealing with spontaneous emotion in an uncontrolled environment.

Empirical Analysis using Posed Expressions: The CK+ was used to quantify the performance of the proposed system and also to perform comparative analysis with [11]. The area under the Receiver Operating Characteristic (ROC) Curve was used as a metric and was abbreviated as ROCA. FSAUs from CK+ were perturbed to add noise in such a way that their recognition performance followed the figure 6 that is reported in [11].

Figure 7 showed the comparative analysis of prediction accuracies of IWAUs. Numbers within the blue bars indicated the ratios between positive and negative samples. Figure 7 suggests that ROCAs for AU 1, 5, 10, 11, 16, 20, and 26 were increased compared to [11]. Particularly, performance of AU 10 was increased by 19%, AU 11 by 9%, AU 16 by 8.64%, and AU 26 by 12.5%. It indicates that the performance of the proposed approach is significantly better compared to concurrent methods even though IWAUs were predicted without any computer vision techniques. The key observation is that the particular AUs that are difficult AUs (AU 10, 11, 14, 16, 20, and 26) were predicted with significant ROCA. Another comparative analysis was performed with the work reported by Tong

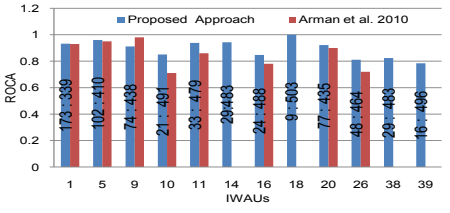

Fig. 7. ROCA for 3-fold cross validation with CK+

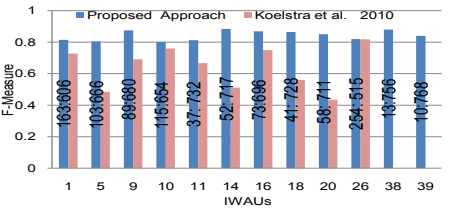

Fig. 8. Recognition Performance of IWAUs for MMI dataset

et al. [12] to provide additional perspectives in the light of dynamic modeling of AUs where the True Positive rate (TPR) of AU 1, 5 and 9 are 0.8, 0.75 and 0.9, respectively. The corresponding number for the proposed approach are 0.93, 0.97, and 0.98, respectively.

Generalization using Mixed Expressions: MMI dataset was used to test the generalization of the proposed approach. This experiment used CK+ dataset for training the DBN and the MMI dataset for testing. In this experiment, F-measure was used as performance metric. The input to the DBN were corrupted to produce FSAUs with F-measures reported in [3].

Comparative results between the proposed approach and [3] were shown in the figure 8. It was found that all IWAUs achieved significant increase in F-Measure. The average F-measure for the proposed approach was 0.8429 whereas [3] got 0.6404. The F-measure of the difficult AUs AU 11, 14, 18, and 20 were reasonable enough compared to the state-of-the-art techniques.

Analysis with Natural Expressions: To further illustrate the utility of the proposed method in the real life scenario, an experiment was performed using the EE dataset [14]. In this experiment, CK+ database was used for training while the EE dataset was used for testing. Comparison of the performance was made with spontaneous facial expression dataset (FACS-101) collected by Mark Frank [1] while FSAUs followed the ROCAs from [4]. Particularly, ROCAs of AU 14, 20, and 26 have been increased 14%, 33.8% and 21.45%, respectively although these AUs are considered error prone.

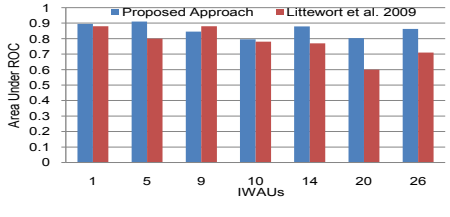

Fig. 9. Performance evaluation for IWAUs in spontaneous datasets

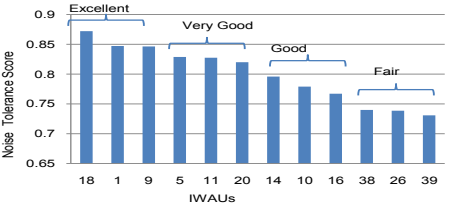

Fig. 10. Noise Tolerance Scores for IWAUs

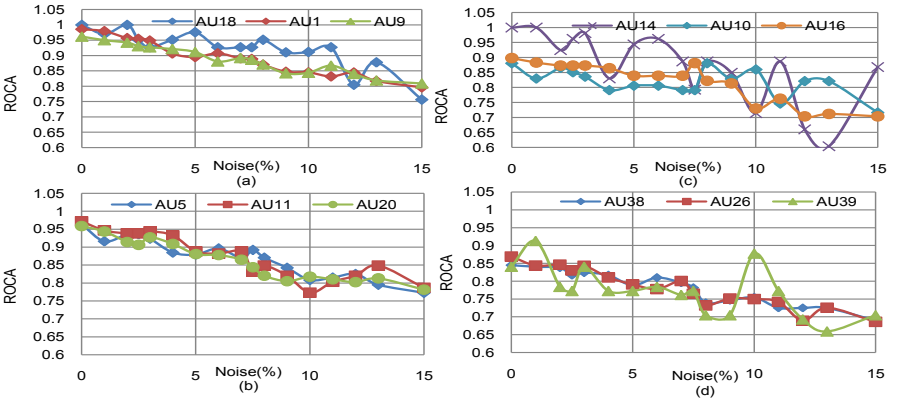

Fig. 11. Performance of IWAU Prediction against Noise

4.3 Performance Analysis with Perturbation

Though state of the art technologies have achieved significant improvement in the AU recognition, noise is added due to the inaccuracies in measurements of subtle facial deformation, pose, and out of plane head movements. A number of studies were performed to study the effect of perturbation both at the AU level as well as in predicting categorical emotion.

Noise Analysis in Predicting IWAUs: In the context of FSAUs, noises are propagated through the proposed DBN while inferring the IWAUs. However, the proposed method makes the IWAUs less susceptible to noise due to spatio-temporal relations between AUs. CK+ and MMI database were used for these experiments with three fold cross validations.

Noises from identical uniform distribution were added to each of the FSAUs and subsequently noisy FSAUs were used for IWAU prediction. Figure 11 depicts the prediction results of IWAUs with varying degrees of noise. To provide a better insight, a measure to calculate the "noise tolerance" was introduced. The approach was to find the RMS error of the performance graph from the baseline performance (with noise 0%). After that, RMS error was subtracted from theoretically maximum possible RMS error for each IWAU. Thus, the result can be interpreted as a measure of noise tolerance where ideal score should be 1.00. To provide better insight and visualization, IWAUs were grouped into four clusters (excellent, very good, good, and fair) based on the tolerance score (fig 10).

It was observed that AU 18, 1, and 9 had excellent noise tolerance in figure 11a. Among them, AU 18 seemed to be mostly tolerant against noise. It was also noted that noise had linear effect on the prediction of AU 1 and 9. The degradation of their performance was gradual as the ROCA remained above 0.90 and above 0.85 while the noise were increased to 7% and to 12%, respectively. AU 5, 11, and 20 showed very good noise tolerance as observed from the figure 11b. Average ROCA of this group remained above 0.9 for noise up to 4% while average ROCA went below 0.80 after 10% noise. The AU 14, 10, and 16 was found to have good tolerance (fig 11c) and the other AUs showed fair tolerance.

For fair tolerant IWAUs (fig 11d), more than 4% noise resulted inferior ROCA on average.

Noise Analysis in Facial Expression Recognition: An additional experiment was performed to analyze the effects of perturbation/noises in FSAUs on recognition of six categorial emotions. CK+ and MMI database were used for this experiment while emotion recognition was performed on 18 AUs (11 FSAUs and 7 IWAUs). FSAUs are perturbed accordingly to generate noisy FSAUs as earlier experiments. Then, IWAUs were inferred using proposed technique while another Bayesian net [15] had been used to predict the emotions.

It was observed from figure 12 that '**Happy**' emotion had the best noise tolerance against noisy FSAUs. The performance was found to be almost consistent as the ROCA remained above 0.95 up to 29% noise in FSAUs with one exception. '**Surprise**' expression showed convincing error tolerance as well. It maintained ROCA more than 0.90 across the whole noise range. Though '**Disgust**' had inferior performance after 9% noise compared to Happy and Surprise expressions, its ROCA was more than 0.85 up to 23% noise level with two exceptions only. Performance in predicting '**Anger**' found to be somewhat better than 'Disgust'. The ROCA of Anger recognition was more than 0.90 where noise increased from 0% to 7% and ROCA remained more than 0.80 up to 28% noise. '**Sad**' has reasonable noise tolerance up to 24% noise while more than 0.85 ROCA has been maintained.

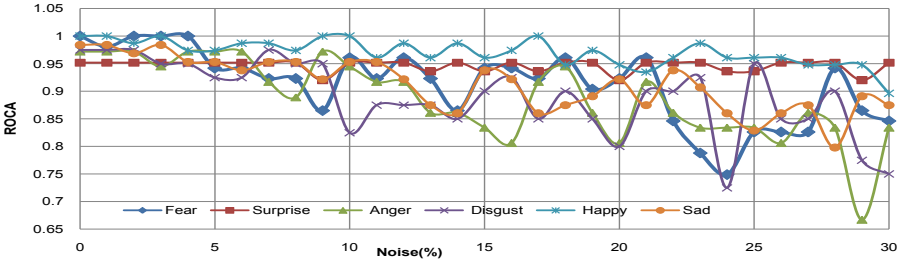

Fig. 12. Performance of Facial Expression Recognition against Noise

5 Conclusions

Reliable estimation of affective states and continuous categorical emotion spotting require all AUs or a majority of them. However, automated, robust, and real-time recognition of all AUs are computationally expensive and error prone. To address such issues, this paper proposed a spatio-temporal data driven probabilistic scoring function to divide the AUs into FSAUs and IWAUs. SBN was used to capture the relations among AUs and the DBN was used to capture their temporal evolution. A framework was implemented to predict the IWAUs on the fly from the FSAUs with very high accuracy. The proposed approach improved robustness in predicting IWAUs. In addition, perturbation anaylsis

was performed to understand the effect of noise at both the AU level as well as recognition of categorical emotion. These contributions will enable real-time analysis, synthesis, and tracking of complex and natural facial behaviors.

References

1. Bartlett, M.S., Littlewort, G., Frank, M.G., Lainscsek, C., Fasel, I.R., Movellan, J.R.: Automatic recognition of facial actions in spontaneous expressions. Jour. of Multimedia 1(6) (2006)
2. Ekman, P., Friesen, W.V.: Facial action coding system: A technique for the measurement of facial movement (1978)
3. Koelstra, S., Pantic, M., Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. IEEE Tran. on PAMI (2010)
4. Littlewort, G.C., Bartlett, M.S., Lee, K.: Automatic coding of facial expressions displayed during posed and genuine pain. Image Vi. Comput. 27, 1797–1803 (2009)
5. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: CVPR Workshops (June 2010)
6. Lucey, S., Ashraf, A., Cohn, J.: Investigating spontaneous facial action recognition through aam representations of the face. I-Tech Ed. and Pub. (2007)
7. Pantic, M., Valstar, M.F., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. IEEE, Los Alamitos (2005)
8. Picard, R.W.: Affective computing: challenges. Int. J. Hum.-Comp. Stud. 59 (2003)
9. Mahbubur Rahman, A.K.M.: Using probablistic graphical model in finding significant subset of facial action units. MS Thesis (2011)
10. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3D face analysis. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BIOID 2008. LNCS, vol. 5372, pp. 47–56. Springer, Heidelberg (2008)
11. Savran, S., Sankur, B., Bilge, M.T.: Facial action unit detection: 3d versus 2d modality. In: CVPR Workshop on Human Comm. Behavior Anal., USA (2010)
12. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans. on PAMI (29), 1699 (2007)
13. Tong, Y., Chen, J., Ji, Q.: A unified probabilistic framework for spontaneous facial action modeling and understanding. IEEE Trans. on PAMI 32(2), 258–273 (2010)
14. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. on Multimedia (2006)
15. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. on PAMI (27), 699–714 (2005)