

Dynamic Cascades with Bidirectional Bootstrapping for Spontaneous Facial Action Unit Detection

Yunfeng Zhu^{1,2} Fernando De la Torre² Jeffrey F. Cohn^{2,3} Yu-Jin Zhang¹

¹ Department of Electronic Engineering, Tsinghua University, Beijing 100084.

² Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

³ Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

Email: zhu-yf06@mails.tsinghua.edu.cn ftorre@cs.cmu.edu jeffcohn@pitt.edu zhang-yj@mail.tsinghua.edu.cn

Abstract

A relatively unexplored problem in facial expression analysis is how to select the positive and negative samples with which to train classifiers for expression recognition. Typically, for each action unit (AU) or other expression, the peak frames are selected as positive class and the negative samples are selected from other AUs. This approach suffers from at least two drawbacks. One, because many state of the art classifiers, such as Support Vector Machines (SVMs), fail to scale well with increases in the number of training samples (e.g. for the worse case in SVM), it may be infeasible to use all potential training data. Two, it often is unclear how best to choose the positive and negative samples. If we only label the peaks as positive samples, a large imbalance will result between positive and negative samples, especially for infrequent AU. On the other hand, if all frames from onset to offset are labeled as positive, many may differ minimally or not at all from the negative class. Frames near onsets and offsets often differ little from those that precede them. In this paper, we propose Dynamic Cascades with Bidirectional Bootstrapping (DCBB) to address these issues. DCBB optimally selects positive and negative class samples in training sets. In experimental evaluations in non-posed video from the RU-FACS Database, DCBB yielded improved performance for action unit recognition relative to alternative approaches.

1. Introduction

The face is one of the most powerful channels of nonverbal communication. Facial expression provides cues about emotional response, regulates interpersonal behavior, and communicates aspects of psychopathology. While people have believed for centuries that facial expressions can reveal what people are thinking and feeling, it is only recently that the face has been studied scientifically for what it can tell us

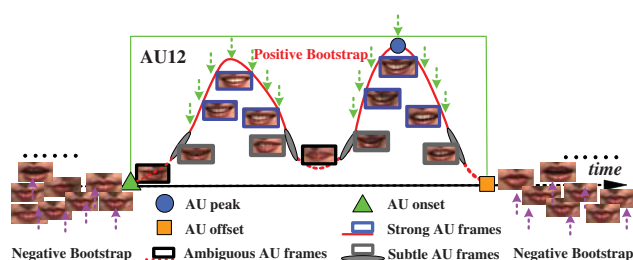


Figure 1. Examples of AU12. Frames between the onset and offset as differentiated as strong, subtle, and ambiguous AUs. The strong frames typically correspond to the peak of the AU and the ambiguous ones to the onset and offset. Our approach iteratively selects the positive and negative frames that optimize classification performance.

about internal states, social behavior and psychopathology.

Faces possess their own language. To represent the elemental units of this language, Ekman and Friesen [14] proposed the Facial Action Coding System (FACS). FACS segments the visible effects of facial muscle activation into "action units." Each action unit is related to one or more facial muscles. The FACS taxonomy was defined by manually observing graylevel variation between expressions in images and to a lesser extent by studying the electrical activity of underlying facial muscles [10]. Because of its descriptive power, FACS has become the state of the art in manual measurement of facial expression and is widely used in studies of spontaneous facial behavior [15]. Much effort in automatic facial image analysis seeks to automatically recognize FACS action units [23, 32, 26]. This task is challenging for several reasons: More than 7000 AU and AU combinations have been observed [24], non-frontal pose and moderate out-of-plane head motion are common and the temporal scale of facial actions is highly variable.

Selection of training samples presents an additional challenge given the complexity of facial action units and the

large data sets encountered in working with non-posed, real-world data. Most approaches to AU recognition pose the task as a binary classification. For a given AU, video frames annotated as peaks are used as positive examples in training, and those that have been annotated as other AU or neutral (i.e. AU 0) are randomly selected for inclusion in the negative class. This approach presents at least two problems. One is that the number of samples in the positive and negative class typically is unbalanced, with a small set of positive examples and a very large set of negative ones. Another is that the number of potential training samples easily may exceed the limits of the classifier. Support Vector Machines (SVMs), for instance, fail to scale well with increases in the number of training samples (e.g. $O(n^3)$ for the worst case in SVM). Yet, how best to choose the positive and negative samples is problematic. If we only choose the peaks as positive samples, there will be a large imbalance between positive and negative samples, especially for infrequent AU, and many examples of moderate or lower intensity may be neglected. On the other hand, if all the frames from onset to offset are labeled as positive, there is risk of significant error in training the classifier, as samples close to the onsets and offsets may differ imperceptibly from the negative cases.

To address these issues, we propose Dynamic Cascades with Bidirectional Bootstrapping (DCBB), an extension of AdaBoost typically used in face detection [30]. Manual FACS annotation labels the onset, peak, offset of AUs [10], but training with all the samples is computationally expensive and the use of subtle or ambiguous frames from near the onset and offset impairs learning. As illustrated in figure 1, action units near the onset and offset (dotted red line) may be subtle and difficult to distinguish from non-AU frames. To optimize the sampling of positive cases, DCBB uses an iterative approach to sample AU, beginning with the peak and near-by frames and then extending iteratively toward the onset and offset. Selection continues until maximum recognition occurs. Specifically, DCBB starts by selecting the peak and two adjacent frames as samples in the positive class and uses a cascade AdaBoost classifier as first approximation. Next, a bootstrapping approach is used to select frames that belong to the positive and negative class until training convergence. Figure 1 illustrates the idea of spreading from the AU peaks to the subtle AUs.

2. Previous Work

This section describes previous work on FACS and prior work on automatic recognition of AUs from video.

2.1. FACS

The Facial Action Coding System (FACS) [14] is a comprehensive, anatomically-based system for measuring nearly all visually discernible facial movement. FACS de-

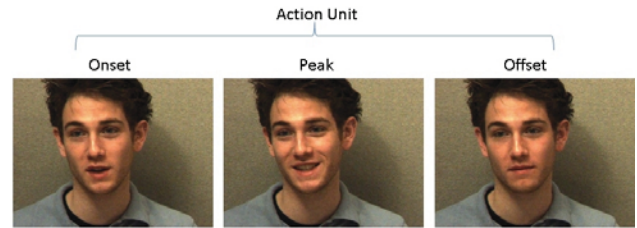


Figure 2. FACS coding typically involves frame-by-frame inspection of the video, paying close attention to transient cues such as wrinkles, bulges, and furrows to determine which facial action units have occurred and their intensity. Full labeling requires marking onset, peak and offset and may include annotating changes in intensity as well. Left to right, evolution of an AU 12 (involved in smiling), from onset, peak, to offset.

scribes facial activity on the basis of 44 unique action units (AUs), as well as several categories of head and eye positions and movements. Facial movement is thus described in terms of constituent components, or AUs. Any facial expression may be represented as a single AU or a combination of AUs. For example, the felt, or Duchenne smile is indicated by movement of the zygomatic major (AU12) and orbicularis oculi, pars lateralis (AU6). FACS is recognized as the most comprehensive and objective means for measuring facial movement currently available, and it has become the standard for facial measurement in behavioral research in psychology and related fields. FACS coding procedures allow for coding of the intensity of each facial action on a 5-point intensity scale (which provides a metric for the degree of muscular contraction) and for measurement of the timing of facial actions. FACS scoring produces a list of AU-based descriptions of each facial event in a video record. Fig. 2 shows an example for AU12. Comprehensive reviews of automatic facial coding may be found in [23, 32, 26].

2.2. Automatic FACS recognition from video

Two main streams in the current research on automatic analysis of facial expressions consider emotion-specified expressions (e.g., happy or sad) and anatomically based facial actions (e.g., FACS). The pioneering work of Black and Yacoob [5] recognizes facial expressions by fitting local parametric motion models to regions of the face and then feeding the resulting parameters to a nearest neighbor classifier for expression recognition. De la Torre et al. [13] use condensation and appearance models to simultaneously track and recognize facial expression. Chang et al. [8] use a low dimensional Leipschitz embedding to build a manifold of shape variation across several people and then use I-condensation to simultaneously track and recognize expressions. Lee and Elgammal [17] use multi-linear models to construct a non-linear manifold that factorizes identity from expression. Recently there has been an emergence of

efforts toward explicit automatic analysis of facial expressions into elementary AUs [29, 21] as they are very suitable to be used as mid-level parameters in automatic facial behavior analysis [9]. Several promising prototype systems were reported that can recognize deliberately produced AUs in either near frontal view face images (Bartlett et al., [2]; Tian et al., [26]; Pantic & Rothkrantz, [22]) or profile view face images (Pantic & Patras, [21]). These systems employ different machine learning methods and different image representations as they are the key stages for automatic AU recognition.

Most work in automatic analysis of facial expressions differs in choice of features and/or classifiers. Bartlett et al. [3] investigate machine learning techniques including SVMs, Linear Discriminant Analysis, and AdaBoost, concluding that the best recognition performance is obtained through SVM classification on a set of Gabor wavelet coefficients selected by AdaBoost. However, the computational complexity of Gabor and SVMs are considerable. To develop and evaluate facial action detector, large collections of training and test data are necessary. Although high scores have been achieved on posed facial action data [28, 31, 25], only a small number of studies being conducted on non-posed spontaneous data [7, 3, 19]. The latter are preferable to posed as they are representative of real world facial actions. In our paper, we focus on a problem common to almost all approaches to facial expression analysis; that is, how best to exploit the training data to improve classification performance. We evaluate our approach by detecting FACS action units (AU) in a relatively large data set of non-posed, spontaneous facial behavior.

3. Facial appearance features

This section describes the process for facial feature alignment using Active Appearance Models and the construction of appearance features.

3.1. Facial alignment

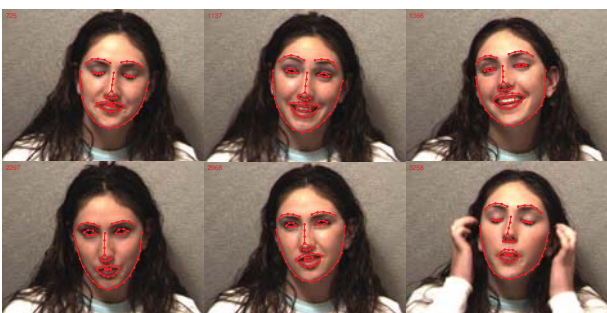


Figure 3. AAM tracking across several frames

Over the last decade, appearance models have become

increasingly important in computer vision and graphics. Parameterized Appearance Models (PAMs) (e.g. Active Appearance Models, Morphable Models, Eigentracking) have been proven useful for detection, facial feature alignment, and face synthesis [6, 12, 11, 20]. In particular, Active Appearance Models (AAMs) [11] have proven an excellent tool for aligning facial features with respect to a shape and appearance model. In our case, the AAM is composed of 66 landmarks that deform to fit perturbations in facial features. Person-specific models were trained on approximately 5% of the video. Fig. 3 shows an example of AAM [20] tracking of facial features in several subjects from the RU-FACS [4] video data-set. Once the tracking is done, facial alignment can be achieved using the registration parameters, and several alignment methods are possible.

3.2. Appearance Features

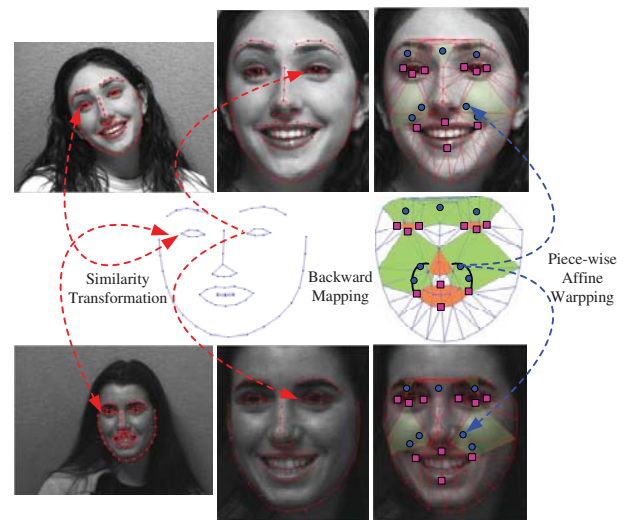


Figure 4. Two-step alignment

Appearance-based representation have been widely used in the literature on AU recognition [3, 27]. For many AU, appearance, appearance features have been shown to outperform shape features (See [1] for comparison of shape and appearance features). In this section, we explore the use of local SIFT descriptors [18] as appearance features. After the face is tracked using AAMs, similarity transform is used to register the face with respect to an average face while the difference of scale, in-plane-rotation and transformation among the images are removed (see middle column in Fig. 4). The features are computed using SIFT descriptors [18] around the points of interest which are tracked in AAMs. Moreover, we also use some areas that have not been explicitly tracked (e.g. nasolabial furrow). To obtain accurate positions of these areas that have not been tracked, we use a backward piece-wise affine warp with the same topology

3. $F_{i+1} = F_t \times f_t$; $D_{t+1} = D_t \times d_{min}$; keep in Q_0 the negative samples that the current strong classifier H_t misclassified, record its size as K_{fq} .
4. Repeat using detector H_t to bootstrap false positive samples from negative Q randomly until the negative working set has N_q samples.

END While

Output:

A t-levels cascade where each level has a strong boosted classifier with a set of rejection thresholds for each weak classifier. The final training accuracy figures are F_t and D_t .

Table 1: Initial Learning

4.2. Dynamic Learning

Once a cascade of peak frame detectors is obtained in the Initial Learning stage, we are able to enlarge the positive set to increase the discriminative performance of the whole classifier. The AU frames detector will become stronger as new AU positive samples are added during the training step, and the distribution of positive and negative samples will be more representative of the whole training data. A constraint scheme is designed in dynamic learning to avoid add ambiguous AU frames to the dynamic positive set. The algorithm has been summarized in Table 2.

Input

- Cascade detector H_0 , from the Initial Learning step;
- Dynamic working set $S_D = \{P_D, Q_D\}$;
- All the frames in this action unit as potential positive samples $P = \{P_s, P_v\}$. P_s contains the strong positive samples, P_0 contains peak related samples described above, $P_0 \in P_s$. P_v contains obscure positive samples;
- A large negative data set Q , which contains all the other AUs and non-AUs. Its size is N_{qtotal} .

Update positive working set by spreading in P and update negative working set by bootstrap in Q Dynamic cascade learning:

Initialize: We set the value of N_p as the size of P_0 . The size of the old positive data set is $N_{p,old} = 0$. Current diffusing stage is $t = 1$.

While $(N_p - N_{p,old})/N_p > 0.1$

1. **AU Positive Spreading:** $N_{p,old} = N_p$. Using current detector on the potential positive data set P to pick up more positive samples, P_{spread} are all the positive examples that decided by the cascade classifier H_{t-1} .

2. **Constrain the spreading:** k is the index of current AU event, i is the index of current frame in this event. Calculate the similarity values (Eq. 1) between the peak frame in event k and all peak frames with the lowest intensity value 'A', the average similarity value is S_k . Calculate the similarity value between frame i and peak frame in event k , its value is S_{ki} , if $S_{ki} < 0.5 \times S_k$, frame i will exclude from P_{spread} .
3. After above step, the remained positive work set is $P_w = P_{spread}$, $N_p = \text{size of } P_{spread}$. Using H_{t-1} detector to bootstrap false positive samples from the negative set Q until the negative working set Q_w has $N_q = N_p \times R_t$ samples, the ratio R_t will become smaller while N_p the become larger.
4. Train the Cascade Detector H_t with the dynamic working set $\{P_w, Q_w\}$. As R_t varies, the maximum acceptable false positive ratio per cascade stage f_{max_t} also becomes smaller (Eq. 2).
5. $t = t + 1$; empty P_w and Q_w .

END While

Table 2: Dynamic Learning

In eq.1, n is the total number of AU sections with intensity 'A', and m is the length of the AU features. The similarity description used in eq.1 is the Radial Basis Function between the appearance representation of two frames.

$$S_k = \frac{1}{n} \sum_{j=1}^n Sim(\mathbf{f}_k, \mathbf{f}_j), \quad j \in [1 : n]$$

$$S_{ik} = Sim(\mathbf{f}_i, \mathbf{f}_k) = e^{-(Dist(i,k)/\max(Dist(:,k)))^2}$$

$$Dist(i, k) = \left[\sum_{j=1}^m (f_{kj} - f_{ij})^2 \right]^{1/2}, \quad j \in [1 : m] \quad (1)$$

The dynamic positive work set becomes larger but the negative samples pool is finite, so R_t and f_{max_t} need to be changed dynamically. Also, some AUs, like AU12, are more frequent than others. After the spreading stage, the ratio between positive and negative samples becomes balanced, except for some rare AUs (e.g., AU4, AU10) which keep unbalanced because of the scarceness of positive frames in the database. Instead of tuning these thresholds one by one, we assume that the false positive rate f_{max_t} changes exponentially in each stage t , which means

$$f_{max_t} = f_{max} \times (1 - e^{-\alpha R_t})$$

$$R_t = \beta \times R_0 \times N_{qtotal}/N_p \quad (2)$$

In our experiment, we set α as 0.2 and β as 0.04 respectively because those values are suitable for all the AUs to avoid lacking of useful negative samples in RU-FACS database.

5. Experiments

This section reports experimental results for AU recognition on the RU-FACS database [4]. RU-FACS consists of video-recorded interviews with 34 men and women of varying ethnicity. Interviews were approximately minutes in duration. Video from four subjects could not be processed for technical reasons (e.g., noisy video), which resulted in data from 29 participants. Meta-data included manual FACS codes. The FACS codes include the peak frame as well as the onset and offset frame for each action unit. Because some AUs occurred too infrequently, we focus our experiments on ten AUs: AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17. For all the AUs, the SIFT descriptor is built using a square of 48×48 pixels and the face is normalized to have 212×212 pixels. We trained 10 dynamic cascade classifiers as described in section 4, using one versus all scheme for each AU. 19 subjects were randomly selected as training, and the remaining 10 subjects were used as testing subjects.

5.1. Positive Samples Spreading in Training Step

This section illustrates the bootstrapping approach to select positive and negative samples and the improvement in ROC (Receiver-Operator Characteristic) curves at successive iterations. ROC curves are obtained by plotting true positives ratio against false positives ratio for different decision threshold values.

In our method, the forced constrain is introduced by setting the lowest boundary of similarity description as described in eq.1. The value of the weight (0.5) in the stage "Constrain the spreading" of Dynamic Learning can be varied between 0.3 to 0.6, the results are insensitive to the value in this range. While the newly adopted positive samples are becoming closer to the optimal hyperplane which decided by the previous cascade classifier, the less samples will be picked up during the Bootstrap stage. So the spreading speed defined in the stage "While" of Dynamic Learning are used to prevent the spreading from subtle AU frames to ambiguous AU frames. Empirical analysis shows that when it reaches a low level (0.1 in our experiment), the boundary between strong AU frames and ambiguous AU frames can be considered reached.

Fig. 6 shows the labeling for AU12 for subject *S015*. There are eight AU12 labeled units of varying intensity from A (trace level) to D (close to maximum). The curves in the lower panel represent the similarity (eq. 1) between each peak and the neighboring frames. This graphic shows the complex temporal patterns and the positive samples spreading in each step. The positive samples in each step are represented by Green Asterisk, Red Plus sign, Blue Cross, Black Circle. The later adopted frames (Black Circle, Blue Cross) are mainly crowd around the low value areas of similar-

ity curve in high intensity AU, conversely the frames with Black circle and Blue cross are scattered around the crest of similarity curve in low intensity AU, see it in subfigure number 3, 8 and 7. It is interesting to observe subfigure number 2 and number 8, the action is shrinking among the unit while the wave of the similarity curves are adopted at last or not adopted as positive samples. Subfigure number 7 shows that for low intensity AU, only the frames around the peak frame are adapted as positive samples. The ellipses in the bottom curves with different gray values (from black to gray) correspond to the strong AU frames, subtle AU frames and ambiguous AU frames which are illustrated in Figure 1.

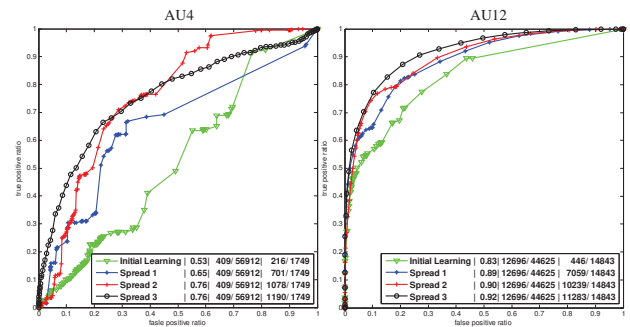


Figure 7. The ROCs improve with the spreading of positive samples

Fig. 7 shows the improvement in the ROCs using our approach. The first number between lines | denotes the area under the ROC, the second number the size of positive samples in the testing dataset and separated by / is the size of negative samples in the testing dataset, the third number denotes the size of positive samples in training working sets and separated by / the total frames of current AU in training data sets. We illustrate the method with the AU4 and AU12, where AU4 has a minimum number of examples and AU12 has the largest number of examples. We can observe that the area under the ROC for frame-by-frame classification is greatly improved after applying our method. The area improves faster for the case of AU4 than AU12, because the peak frames of AU12 with different intensity in the initial step for learning represent the maximum number of AUs while for AU4, the new adopting positive samples can improve the representative ability a lot, as very few positive samples can be used in the initial training set.

5.2. Improving recognition accuracy

This section reports experiments results for AU recognition and compares with previous approaches that use shape or appearance and Support Vector Machines (SVMs) [19] for classification.

We have trained the classifiers using 19 subjects from the RU-FACS dataset and we have selected the remaining 10

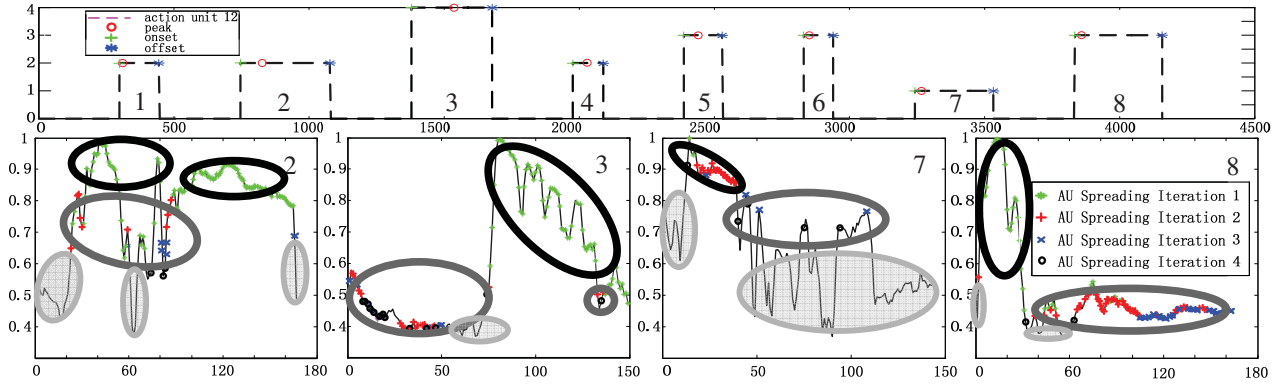


Figure 6. The spreading of positive samples during each dynamic training step for AU12. See text for the explanation of the number between bars.

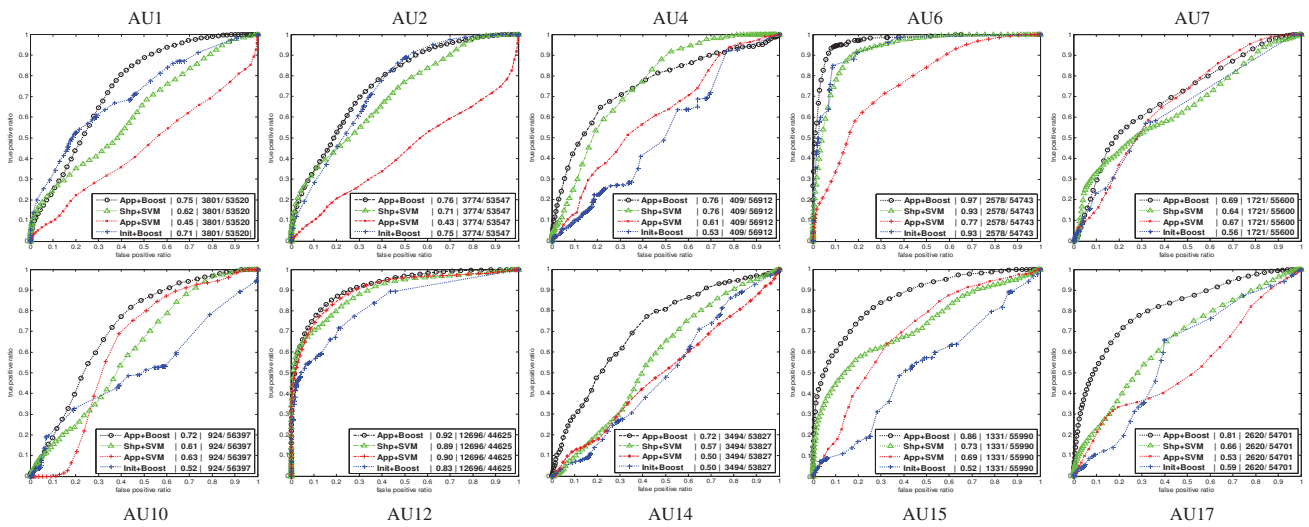


Figure 8. ROC curve for 10 AUs using three different methods: SVM and appearance features (App+SVM), SVM and shape features (shape+SVM), our method (App+Boost) and Initial learning in our method(Init+Boost).

subjects for testing. We report results on 10 AUs using our dynamic bidirectional cascade classifiers on the appearance features (as explained in section 3), and using one versus all strategy. We measure the performance using a frame-by-frame ROC curve. The ROC curves for the 10 AUs can be seen in Figure 8. There are four curves with different labels, 'App+Boost' the proposed method, 'Shp+SVM' is shape features with SVM [19], 'App+SVM' [19] is appearance features with SVM, 'Init+Boost' is initial learning stage in the proposed method. As we can observe, our method outperforms in most AUs to the SVM with shape or appearance features. The SVM is trained using as positive samples the peaks of the current AUs and two adjacent frames. The negative samples are selected randomly (but the same for shape and appearance methods). The ratio between positive and negative samples is fixed to 30. The method particularly

boost performance in AU2, AU 15 and AU 17. Moreover, Compared with the initial learning stage, the dynamic learning stage improve the performance in each of the AUs.

6. Conclusions

This paper proposes an automatic method to automatically select the set of positive and negative samples from the training set that improves recognition performance on AU. Our approach is able to detect subtle AUs and provides a good segmentation for the training data. We compare the performance with existing method using appearance and shape features with Support Vector Machines (SVMs) and AdaBoost, and show how our approach achieves better performance. In future work, we plan to model the dynamic patterns around the onset and offset of AU events and ex-

cluding false positive samples for all AUs.

Acknowledgments This research was supported in part by NIMH grant MH 51435. The first author was also partially supported by the scholarship from China Scholarship Council. The work was performed when the first author was at Robotics Institute, Carnegie Mellon University. Thanks to Tomas Simon, Feng Zhou, Zengyin Zhang for their valuable suggestions.

References

- [1] A. Ashraf, S. Lucey, J. Cohn, T. Chen, K. M. Prkachin, and P. Solomon. The painful face: Pain expression recognition using active appearance models. *Image and Vision Computing.*, 2009.
- [2] M. Bartlett, G. Littlewort, I. Fasel, J. Chenu, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *AFGR*, pages 223–228, 2006.
- [3] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *CVPR05*, pages II: 568–573, 2005.
- [4] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 2006.
- [5] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *IJCV*, 25(1):23–48, 1997.
- [6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999.
- [7] B. Braathen, M. S. Bartlett, G. Littlewort, and J. R. Movellan. First steps towards automatic recognition of spontaneous facial action units. In *Proceedings of the ACM Conference on Perceptual User Interfaces*, 2001.
- [8] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold based analysis of facial expression. In *CVPR '04 Workshops*, page 81, 2004.
- [9] J. Cohn and P. Ekman. Measuring facial action by manual coding, facial emg, and automatic facial image analysis. *Handbook of nonverbal behavior research methods in the affective sciences.*, 2005.
- [10] J. F. Cohn, Z. Ambadar, and P. Ekman. *Observer-based measurement of facial expression with the Facial Action Coding System*. The handbook of emotion elicitation and assessment. Oxford University Press Series in Affective Science., New York: Oxford., 2007.
- [11] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, pages 484–498, 1998.
- [12] F. de la Torre and M. Nguyen. Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment. In *CVPR*, 2008.
- [13] F. de la Torre, Y. Yacoob, and L. Davis. A probabilistic framework for rigid and non-rigid appearance based tracking and recognition. In *AFGR*, pages 491–498, 2000.
- [14] P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press.*, 1978.
- [15] P. Ekman and J. Hager. *What the Face Reveals(2nd ed.)*. Oxford University Press, ISBN 0-19-517964-1, 2005.
- [16] K. kay Sung and T. Poggio. Example-based learning for view-based human face detection. *TPAMI*, pages 39–51, 1998.
- [17] C. Lee and A. Elgammal. Facial expression analysis using nonlinear decomposable generative models. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 17–31, 2005.
- [18] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.
- [19] S. Lucey, A. B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. In K. Kurihara, editor, *Face Recognition Book*. Pro Literatur Verlag, 2007.
- [20] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, pages 135–164, 2004.
- [21] M. Pantic and I. Patras. Dynamics of Facial Expression: Recognition of Facial Actions and their Temporal Segments from Face Profile Image Sequences. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 36:433–449, 2006.
- [22] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 1449–1461., 2004.
- [23] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *ACM International Conference on Multimedia*, pages 669–676, 2005.
- [24] K. Scherer and P. Ekman. *Handbook of Methods in Nonverbal Behavior Research*. 1982. Cambridge Univ. Press.
- [25] Y. Sun and L. Yin. Facial expression recognition based on 3d dynamic range model sequences. In *ECCV08*, pages II: 58–71, 2008.
- [26] Y. Tian, J. F. Cohn, and T. Kanade. *Facial expression analysis*. In S. Z. Li and A. K. Jain (Eds.). *Handbook of face recognition*. New York, New York: Springer., 2005.
- [27] Y. Tian, T. Kanade, and J. F. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *AFGR*, pages 229–234. Springer, 2002.
- [28] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *TPAMI*, pages 1683–1699, 2007.
- [29] M. F. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *CVPR '05 Workshops*, page 76, 2005.
- [30] R. Xiao, H. Zhu, H. Sun, and X. Tang. Dynamic cascades for face detection. In *ICCV07*, pages 1–8, 2007.
- [31] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas. Facial expression recognition using encoded dynamic features. In *CVPR*, 2008.
- [32] W. Zhao and R. Chellappa. (Editors). *Face Processing: Advanced Modeling and Methods*. Elsevier, 2006.