Facial Action Unit Detection using Active Learning and an Efficient Non-Linear Kernel Approximation

Thibaud Senechal, Daniel McDuff and Rana el Kaliouby
Affectiva
Waltham, MA, 02452, USA
thibaud.senechal@gmail.com, {daniel.mcduff,kaliouby}@affectiva.com

Abstract

This paper presents large-scale naturalistic and spontaneous facial expression classification on uncontrolled webcam data. We describe an active learning approach that helped us efficiently acquire and hand-label hundreds of thousands of non-neutral spontaneous and natural expressions from thousands of different individuals. With the increased numbers of training samples a classic RBF SVM classifier, widely used in facial expression recognition, starts to become computationally limiting for training and real-time performance. We propose combining two techniques: 1) smart selection of a subset of the training data and 2) the Nyström kernel approximation method to train a classifier that performs at high-speed (300fps). We compare performance (accuracy and classification time) with respect to the size of the training dataset and the SVM kernel, using either an RBF kernel, a linear kernel or the Nyström approximation method. We present facial action unit classifiers that perform extremely well on spontaneous and naturalistic webcam videos from around the world recorded over the Internet. When evaluated on a large public dataset (AM-FED) our method performed better than the previously published baseline. Our approach generalizes to many problems that exhibit large individual variability.

1. Introduction

Facial expressions contain rich non-verbal information. In recent years many compelling applications for the automated measurement of facial expressions have been presented including: detection of depression [4] and psychological disorders [15], pain measurement [6, 8] and understanding consumer preferences [12]. The facial action coding system (FACS) [1] is the most widely used and comprehensive taxonomy of facial behavior. FACS is a catalog of 27 unique upper and lower facial action units (AUs) that correspond to each of the face's muscles. Manual coding of

Figure 1. Map showing the distribution of the 1.8 million face videos in our dataset. We manually coded 27,000 videos for our experiments. The actions are sparse and there is large individual variability, in addition to challenging lighting and pose.

FACS from video is laborious and requires special training.

Automated facial action detection systems have made a lot of progress over recent years [20]. For many of the applications it is critical that subtle expressions can be detected in real-life settings and in real-time. One of the main limiting factors is that it has been challenging to get training datasets of spontaneous and naturalistic expressions. The Cohn-Kanade dataset (in its extended form called CK+) [7] played a key role in extending the state-of-the-art in facial expression analysis. The CK+ database contains 593 recordings of posed and non-posed sequences. The sequences are recorded under controlled conditions of light and head motion, and range between 9-60 frames per sequence. A number of other public databases have also contributed significantly to the progress of the field: MMI [21],

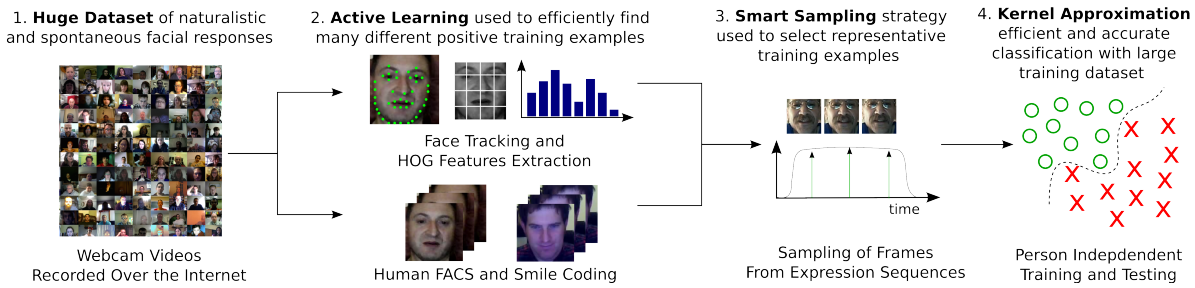

Figure 2. An overview of our approach for facial expression classification. 1) We collected hundreds of thousands of spontaneous and naturalistic facial responses from around the world over the Internet (27,000 of these videos were labeled for facial action units (eyebrow raiser and eyebrow lowered) and smiles). 2) We used an active learning approach to efficiently find positive examples and increase labeling speed. 3) We selected a subset of the frames using a smart sampling technique. 4) We use an SVM with efficient kernel approach for classification to achieve high accuracy and real-time performance.

UNBC-McMaster Shoulder Pain Archive [8], the Bosphorous database [14] and the DISFA dataset [9]. However, many of these datasets were collected under controlled conditions and/or contain posed expressions. Furthermore, these dataset only contain examples of expressions from a few hundred different individuals. We present state-of-the-art AU classification and analyze the impact of training data using the largest facial expression dataset in the world (containing 27,000 expertly labeled recordings). In this work we show results for detection of three important actions: AU02 (outer eyebrow raiser), AU04 (eyebrow lowerer) and smiles. Figure 3 shows examples of these actions. But our approach will generalize to all AUs and to many other computer vision and affective computing problems.

The performance of machine learning algorithms is a factor of the choice of features, classifier and the training data used [16]. One of the difficulties with collecting a large number of examples of spontaneous and naturalistic expressions is that facial expression data are very sparse. In our dataset of reactions to online video content we have found that typically we need to label 30 60-second video responses to obtain a positive expression segment for an action like AU02 or AU04, finding a large number of examples of action units can be a very resource intensive task. Active learning is a semi-supervised learning approach which uses an algorithm to identify data samples more likely to represent an expression of interest. These samples are then labeled by an expert human coder. Active learning is particularly effective in cases where unlabeled data is abundant but only a small portion is worth being labeled. Specifically, we use an active learning approach that allows for much more efficient discovery of positive examples of action units within sparse facial expression data.

With a larger number of training examples it is more challenging to train accurate classifiers that are fast enough to run in real-time. The Radial Basis Function (RBF)-kernel Support Vector Machine (SVM) classifier, which has been widely used in facial recognition [27], starts to become computationally limiting for both training and real-time performance. The classification time of a trained RBF-kernel classifier is proportional to the number of support vectors selected, which depends on the size of the training dataset and parameters of the classifier. We propose two ideas to adapt the classifier in this case. The first is to train the classifier by using a smart subset of the training dataset (where we try to maximize the number of examples from different individuals). Previous work has shown that smart training example selection can be beneficial for performance in both face detection and facial expression detection [17, 5]. Our approach is well suited to our application as samples can be consecutive frames of a video and are similar. The second is to use the Nyström kernel approximation method to find a feature embedding. We compare performance (accuracy and classification time) with respect to the size of the training dataset and the SVM kernel, using either an RBF kernel, a linear kernel or the Nyström approximation method.

In summary, the main contributions of this paper are to: 1) collect a huge video dataset of naturalistic facial actions from across the globe using the Internet, 2) present an active learning approach for efficient labeling of very large numbers of positive AU samples, 3) analyze the effect of the training data size and subject diversity on classifier performance, 4) use a Nyström approximation method for classification to achieve high accuracy and real-time performance. This is, to the best of our knowledge, the first time a real-time facial recognition system has been trained on 80,000 of examples from 4,000 different individuals, and the first time the Nyström approximation method has been studied to improve the trade-off between accuracy and classification time. Figure 2 shows an overview of the main contributions of this paper.

2. Related work

Volume of Training Data: Previous work has shown that more training data is beneficial for reducing computer vision detection error (examples include object detec-

tion [29] and smile detection [23]). However, in the field of human behavior,the "amount of data" problem is not solved, especially given the sparsity with which natural and spontaneous behaviors are observed in real-life settings. We show how to collect huge amounts of positive labeled data with the help of active learning and show results using 4x the number of training samples used by Whitehill *et al*. [23]. Furthermore, Whitehill *et al*. only presented results for one action (smiles) which occur more frequently in everyday life than many other facial actions (e.g. AU02 and AU04).

Active Learning: Active learning has been a topic of recent interest within the machine learning and computer vision communities. Tong and Chang [18] proposed the use of SVMs for providing relevance feedback in image retrieval. Similarly, our method prioritizes sequences for FACS labeling based on SVM classifier outputs. A key difference between the approach in [18] and ours is that we rank image segments based on combinations of actions, since facial action units can occur in different combinations, these add diversity to both the positive and negative training sets. Tong and Koller [19] proposed a method of choosing unlabeled samples by minimizing the version space within the SVM formulation. Zhang and Schueller [28] found active learning to be beneficial in acoustic emotion recognition - a problem that has many similarities (sparsity, unbalanced classes) to visual emotion recognition. Yan *et al*. [25] proposed a multi-class active learning approach for automatically labeling video data. We use an active learning approach to prioritize video segments for labeling by expert coders. FAST-FACS [3] is the closest example of active learning being applied for efficiently FACS coding video sequences. The method uses automated detection to help identify onsets and offsets of actions.

Kernel Approximations: Non-linear kernel methods have been shown to be effective at building discriminative models. However, with larger training datasets the computational cost of non-linear kernels - such as an RBF can become intolerably high. Rahimi and Recht [13] propose mapping data to a randomized low-dimensional feature space in such a way as the inner products of features are similar to those obtained using a kernel (such as an RBF). Using this approach the application of fast linear methods is possible while still obtaining similar performance to more complex kernels. Yang *et al*. [26] compared the generalization performance using random Fourier features and the Nyström method for kernel learning. In this work we compare the Nyström method against an RBF and linear kernel and show that it provides a much better trade-off between accuracy and classification time.

3. Data

Collection: The data we use in our analysis was collected using a web-based framework, similar to that used in [10]. Our framework was deployed to capture facial responses from individuals over the Internet using their webcam. At the time of writing we have collected approximately 1.8 million face videos of individuals responding spontaneously to video content around the world. Figure 1 shows the number of face videos collected in each country. The individuals were responding to a variety of video content including: many types of advertisements, political debate clips, movie trailers and entertainment clips. The different stimuli and demographics gives us a broad range of expressions with different emotional significance. Due to the laborious nature of FACS coding it was only possible for human coders to hand-label a subset of this data (27,000 videos) as described below.

Figure 3. Positive and negative examples of facial actions from our dataset. In many cases the expressions are subtle.

The participants were recruited through market research panels with subjects being contacted via email. During the data collection participants were asked to opt-in to each study and allow their webcam feed to be recorded. The consent forms and instructions were electronic and translated into the appropriate local language. We believe that providing an example of data collection on this scale across the globe is a considerable feat and one of the significant contributions of this work. The data is naturalistic (not induced) and consequently the expressions are sparse and there is large individual variability. In addition to containing subtle spontaneous expressions the resulting webcam videos vary in quality due to lighting conditions and Internet bandwidth. Figure 1 shows examples of the data.

Labeling: For training and testing the action unit classifiers, we had a subset of the webcam videos labeled for each action (Smile, AU02, AU04). A minimum of three FACS trained labelers coded the data. For AU02, AU04 and smiles the free marginal kappa coefficient calculated for 1,100,000 frames were 0.89, 0.79, 0.74 respectively.

Frames used as positive examples of an action had to be labeled as positive by at least 50% of the labelers. Negative examples of an action had to be labeled as negative by 100% of the labelers. All other frames were neither used for training nor testing.

In naturalistic data the distribution of specific action units can be sparse. Therefore it can take a large amount of video coding to collect a small set of positive examples. We used an active learning strategy to efficiently prioritize

the labeling of the data in order to efficiently find positive training examples. We will describe the methodology used below. The approach increased the labeling efficiency significantly. In total 27,000 expression segments were labeled (labeled examples featured at least 4,000 of individuals per action). The data used for training were collected in different studies than those used for testing in order to increase the generalizability. The training and testing data were participant independent.

4. Approach

4.1. Active Learning

In order to collect a large number of positive examples of each action unit, especially for AU02 and AU04 which are sparser than smile expressions, we used an active learning strategy. An overview of the procedure is shown in Figure 4. An initial set of AU02 and AU04 classifiers, RBF-kernel SVMs, trained using much less data than those described in this paper, were used to generate predictions of the presence of each action across a set videos. The number of images in each class in the initial classifier training sets were 6275, 3,771 and 1,858 for Smile, AU02 and AU04 respectively. These classifiers, calibrated between 0 and 100, were applied on all our data. The calibration was performed by applying a sigmoid function to classifier outputs and multiplying the result by 100. The center of the sigmoid was selected to achieve an operating point at 2% false positive rate. Using the classifiers, we selected a set of video segments for labeling as follows:

1) We looked for segments of video in which the output starts below 10, increases above 10 for at least two seconds and then goes back below 10. The threshold of 10 (10% of the maximum value) was found to work well and produce a balance of true positives and false positives for labeling.

2) We ranked these segments using the average value of the classifier output over this segment. Therefore, the segments with a high output but also a sharp onset (start of action) and offset (end of action) have the highest rank. Ranking was performed using the outputs of all three classifiers together. Therefore, we generate labeled positive (expressive) examples for one action that can be used in the negative set for a different action, thus increasing the diversity of the training pool.

3) We selected 13,500 of the highest ranked segments from each of the AU02 and AU04 classifiers to yield a total of 27,000 segments. We labeled them for AU02, AU04 and smile using the labeling proceeding described above.

Without active learning, less than 2% of the data we labeled contained an AU02 or AU04, 20% of the data contained a smile. Using active learning, around 30% of the segments found using the AU02 classifier had at least one frame with an AU02 and the same ratio was found for AU04. Around 20% of the segments had at least one frame labeled as smile. Active learning helps us to label more positive samples for the training set, but also help us find expressions that are likely to generate a false alarm, expressions that we can use as negative samples in the training dataset.

4.2. Tracking and Features Extraction

In order to obtain image features the video sequences were analyzed on a frame-by-frame basis. The OpenCV face detector [22] was used to identify the largest face within the image. We then applied a custom facial feature point tracker (similar to that proposed by Xiong and De la Torre [24]) within this region of interest to identify 34 landmark points on the subject's face. The image region of interest (ROI) was defined using the outer eye corners and mouth points with the resulting ROI containing the whole of the eyebrows and mouth. The face ROI was normalized by performing a rotation, to align the eyes horizontally, and scaling, to a uniform 96x96 pixel scale. Histogram of oriented gradient (HOG) [2] features were extracted from the resulting image ROI. The HOG features were extracted from 32 x 32 pixel blocks (cell-size 8 x 8 pixels) with a stride of 16 pixels. A histogram with 6 bins was used for each block. This results in a feature vector of length 2,400 (25*16*6). To find these parameters we performed intensive tests on a separate cross-validation data set using different combinations of HOG block, stride, cell and bin size.

4.3. Data Sampling and Kernel Approximation using the Nyström method

Kernelized SVMs work well on complicated non-linear separable classification tasks, but this does not scale well to many training samples, as the training time is $O(N^3)$ and the classification time is often linear to the number of training samples. Our goal was to develop a facial expression detector that can be trained on an increasing number of training samples and still be able to be used in real-time. To achieve this we combine two techniques.

First, sequential frames within a video can be very similar and may not yield significant additional information. We propose to select a subset of the frames in our videos which maximizes the number of examples from different individuals. The results of our experiments show that not only the number of training samples is important, but also the variability of these samples, and having several examples of an expression from the same subject does not increase the result as much as having examples of an expression from different subjects. We only select a subset of frames; if we were to use all the frames from our dataset in our analysis the training and validation would be extremely time consuming.

Second, considering the RBF kernel function for two

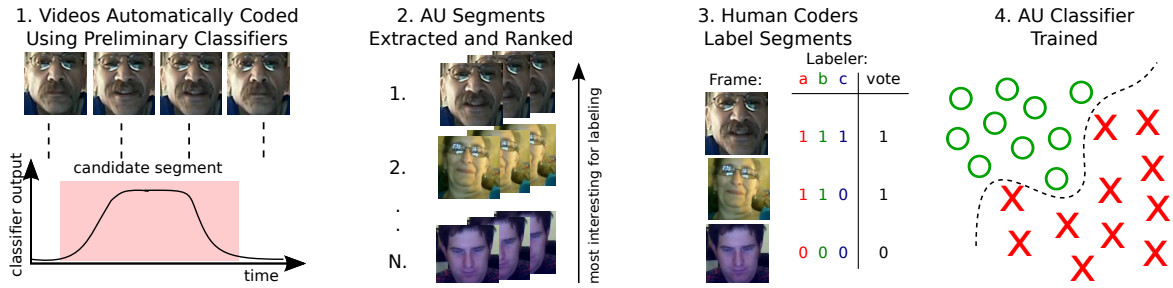

Figure 4. An overview of the active learning strategy used to efficiently generate positive examples of each action unit. 1) Facial videos automatically coded using preliminary classifiers. 2) AU segments extracted and prioritized for labeling. 3) FACS trained coders label the segments. 4) AU classifier trained using new examples.

samples or features vector $\mathbf{x_i}$ and $\mathbf{x_j}$, $k_{RBF}(\mathbf{x_i}, \mathbf{x_j}) =< \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) >= \exp(-\gamma(\mathbf{x_i} - \mathbf{x_j})^T(\mathbf{x_i} - \mathbf{x_j}))$, where γ is an hyper parameter, we try to find an approximation $\widetilde{\phi}$ of the mapping function ϕ, which is of infinite dimensionality, that can be applied directly to each sample. This can be done as only a finite subspace of that infinite space is needed to solve the SVM problem, the one spanned by the samples of the training data.

As using all training samples would lead to a projection to an $(\mathbb{R}^N)$ dimensional space and have the same scaling problems as the RBF-kernel SVM, we can find an approximate embedding by selecting a random subset of the training samples. This is called the Nyström method. Concretely, if we consider *Ns* samples $(\mathbf{x_i})_{i=1\cdots Ns}$ randomly selected from the training dataset, the mapping $\widetilde{\phi}$ for any sample $\mathbf{x}$ is:

$$(\widetilde{\phi}(\mathbf{x}))_i = \exp(-\gamma(\mathbf{x} - \mathbf{x_i})^T(\mathbf{x} - \mathbf{x_i}))/\sqrt{(\mathbf{s})_i} \quad (\text{for } i = 1 \cdots Ns) \tag{1}$$

where $\mathbf{s}$ are the eigenvalues of the *Ns* samples kernel matrix. This normalization is done so $k_{RBF}(\mathbf{x_i}, \mathbf{x_j}) = < \widetilde{\phi}(\mathbf{x_i}), \widetilde{\phi}(\mathbf{x_j}) >$ for all the samples (i, j) belonging to the subset of *Ns* samples. This mapping is applied on all our data, and then a linear SVM is learned in this $\mathbb{R}^{Ns}$ dimensional space. This way, the classification time comes mostly from computing the feature vector $\widetilde{\phi}(\mathbf{x})$ which is proportional to *Ns*. This allows us adapt the classifier to our specific application: we can make the system faster by reducing *Ns*, or get an approximation closer to the RBF kernel by increasing it.

5. Experiments

Training: For each experiment, we built the training dataset by taking as many positive samples as negative samples from our labeled data using the data sampling explained above. In the experiments below we compare the performance with and without the active learning data as part of the training set. When the active learning data is not used, the positive samples selected for each expression are from 1800, 1800 and 1400 subjects for Smile, AU04 and AU02 respectively. With the active learning dataset, they are from 4000, 4800 and 5200 subjects respectively. Our sampling method selectively chooses samples to maximize the number of different subjects in the training set. As an example, for training AU02 using the active learning dataset, when using 1000 training examples the examples would be taken from 1000 of the 4000 videos, each from a different subject, and when using 10000 training examples the examples would be taken from all 4000 videos. As we discuss below, part of the reason for the performance leveling off when training with 10000s of positive examples might be the diversity of the training set not growing significantly when the same individuals appear more than once. In all the experiments the training and testing data were from different participants.

In the training process, we tested the following SVM parameters on an independent validation dataset. In Section 6 we report the results on the test dataset after tuning on the validation dataset.

SVM with Linear Kernel - The penalty parameter, *C*.

SVM with RBF Kernel - The penalty parameter, *C*, and the RBF kernel parameter, γ.

SVM with Approximated RBF Kernel - We test classifiers with the RBF kernel approximated using *Ns*=200, 500, 1000 and 2000 samples. For each case, we try several values for *C* and γ.

Testing: We built our test dataset using 10,000 samples from videos of 2,500 different individuals. To have the test dataset completely independently of the active learning process, these videos were fully labeled not filtered by the active learning algorithm, and were of different participants than the videos used in training. The 10,000 samples were chosen to have an equal representation of the following group of expressions: Smile, AU02, AU04, neutral and others (some containing other expressions like disgust or AU15 (lip corner depressor)). Samples within each group were chosen randomly.

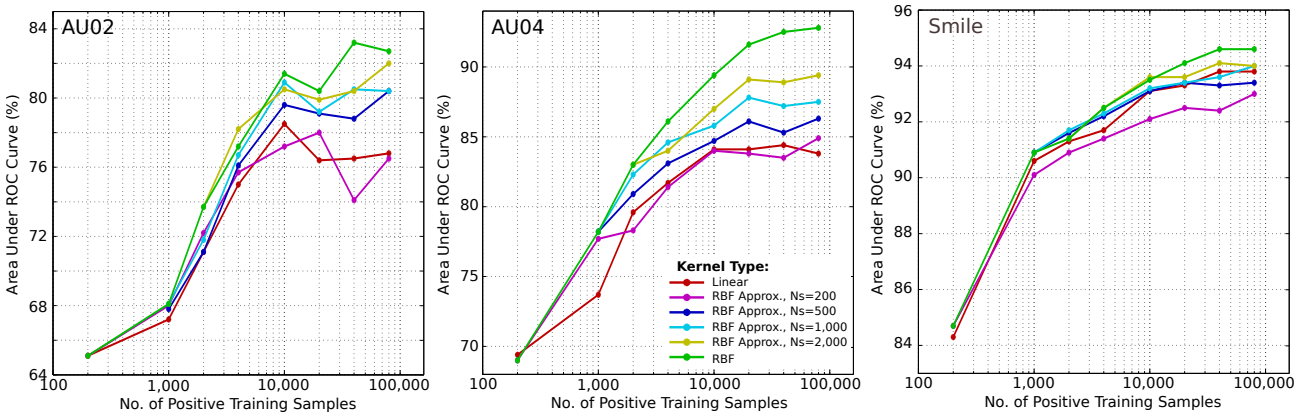

Figure 5. Area under the ROC curve for SVM classifiers with linear, approximated RBF and RBF kernels. The results with 200, 1,000, 2,000, 4,000, 10,000, 20,000, 40,000 and 80,000 training samples are shown. Left) Results for AU02, middle) results for AU04, right) results for smile. The accuracy increases with the number of training examples used. The RBF gives the best performance in almost all cases but the approximations perform well and considerably better than a linear kernel model.

6. Results

Below we show the performance of models isolating the impact of different parts of our approach. We use the area under the receiver operating characteristic curves as a metric for comparing the performance.

6.1. Kernel Approximation

Figure 5 shows the area under the receiver operating characteristic (ROC) curve for SVM classifiers using a linear kernel, RBF kernel and approximated RBF kernel (using for the approximation a no. of samples, Ns = [200, 500, 1000, 2000]). The results are shown for different numbers of training samples (200, 1000, 2000, 10000, 20000, 80000). Trends show that performance increases with greater numbers of training examples. The RBF kernel model performs the most accurately in almost all cases.

We also can notice that in the case of the approximated RBF SVM, increasing the number Ns of samples used to approximate the RBF kernel increases the accuracy, and get it closer to the RBF kernel.

6.2. Accuracy and Computational Cost Trade-off

Figure 6 shows the performance of the different kernels in terms of accuracy and classification time. To get a measure of the classification time, we measured the time spent to detect the AUs once the HOG were computed on 10,000 test frames. For the Approximated RBF kernel, this also includes the mapping of the HOG to a new feature space. The measure was computed using a C++ OpenCV implementation on a computer with 3.5GHz Intel Xeon processor and a 64-bit Ubuntu operating system. As one classifier has to be applied per each AU we try to detect, we need this part to be relatively fast. We found that each classifier has to have a classification time of at least 300 FPS for our server application, and 1000 FPS for our mobile application. For the linear SVM, we get a really fast classifier for which the process time is negligible compared to the HOG computation time (2500 fps). But for AU04 and AU02, the AUC score is much lower. For the RBF classifier, as expected, reducing the training dataset using our smart subsampling strategy allows us to get a classifier several order of magnitude faster, for a small hit in performance. For the approximated RBF, we can improve the AUC score by increasing the number of samples Ns used to approximate the RBF. But this is at a cost of a lower FPS, as when testing a new sample, we need to compute the value of the RBF function between the new sample and the Ns samples. But overall, the approximated RBF kernel offers a much better trade-off than the RBF kernel for FPS between 300 and 1000.

The classifier speeds reported are for the classification of each action from the HOG features (on a fast computer). Face detection, tracking and the detection of multiple AUs will slow the process down. For real-time applications, we found that we need 300FPS for server-side applications and 1000FPS for mobile applications. For a system that has to have 300 FPS, the approximated RBF kernel leads to ROC AUCs approximately increased from 77.5%, 87% and 93.5% to 82%, 89.5% and 94% for AU02, AU04 and Smile respectively. This shows the main advantage of the method, which offers good accuracy and flexibility in the choice of a classification time that suits the application.

6.3. Active Learning

Figure 7 shows the impact of including data from the active learning strategy for training the classifiers. Results for the linear kernel, the RBF kernel and the approximated RBF kernel using Ns = 1,000 samples are shown. The number of training samples is the same in both the case in which we

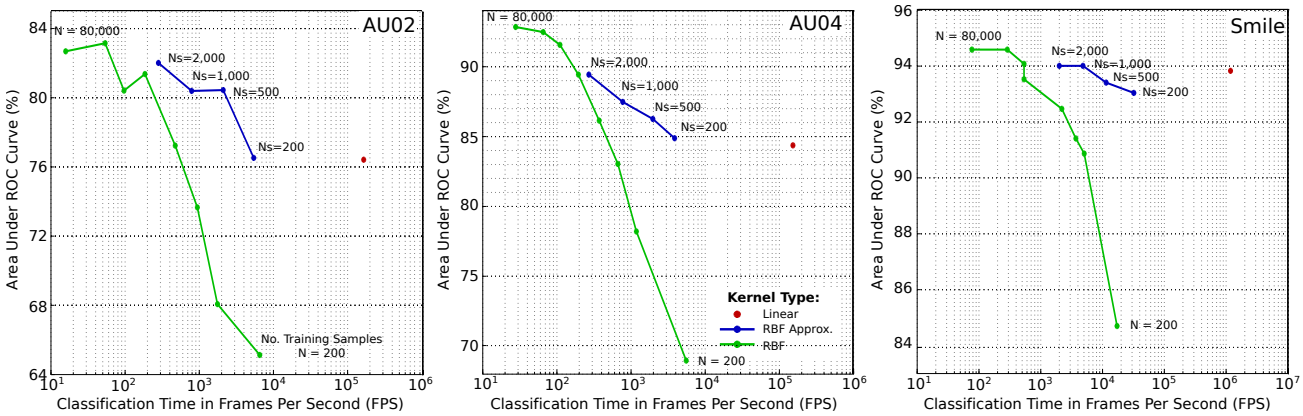

Figure 6. Trade-off between accuracy and classification time for different SVMs. For the RBF SVM, different values are from different size of training dataset (N = 200, 1,000, 2,000, 4,000, 10,000, 20,000, 40,000 and 80,000 samples). The approximated RBF SVM values are for a training dataset of 80,000 samples, but different values of Ns to approximate the RBF kernel (Ns = 200, 500, 1000, 2000).

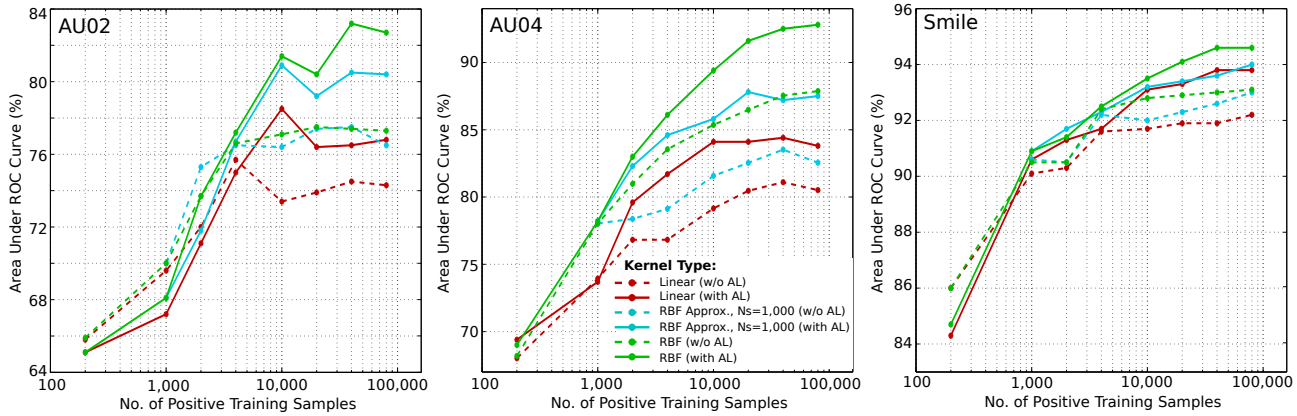

Figure 7. The impact of using active learning data for training the expression classifiers. The results above are shown for an SVM classifier (with RBF, approximated RBF kernel and linear kernels). The results with 200, 1,000, 2,000, 4,000, 10,000, 20,000, 40,000 and 80,000 training samples are shown. Including the data generated using active learning in our training dataset increased performance as this provided many more examples of actions from different people.

use the active learning data and the case we do not. However, with the active learning data, there are examples of the target expression from two to three times more individuals.

Clearly this is beneficial and there is a $\sim$5% increase in ROC AUC using this strategy rather than just sampling more examples from a smaller set of videos. It is not the number of training examples that matters, but the variability of the training data: the number of examples from different expression segments and different individuals. Even the smile classifier, which performs very well, is still improved using the active learning data.

Also, we notice in the case of the data without active-learning, accuracy does not increase significantly when we go from 10,000 to 80,000 training samples. Because several examples comes from the same expressive video segment (therefore the different frame contain similar information), we can aggressively subsample the training dataset as explained in 4.3, and get similar performance.

Overall, this shows that using the proposed active-learning approach, we can find and label naturalistic and spontaneous subtle expressions much faster, and get a high boost in accuracy with these samples. Although we show results for three actions in this paper, the framework could be used for all facial actions. Furthermore, it could apply to other computer vision problems involving sparse data and large individual variability. Active learning performed effectively for finding positive examples and is likely to be even more beneficial for actions that occur even less frequently in real-world data.

6.4. Performance on Public AMFED Dataset

In order to provide a baseline by which others can compare their results to ours we evaluated the performance of our algorithm on the AM-FED dataset [11]. The AM-FED

Method	AU02	AU04	Smile
McDuff *et al.* [11]	0.72	0.70	0.90
Ours (RBF)	0.87	0.70	0.94

Table 1. Area under the ROC curve on the AM-FED dataset for the AU02, AU04 and Smile classifiers with RBF kernel.

consists of 242 facial videos (168,359 frames) of facial responses collected via individuals webcams whilst they watched online videos. The number of positive AU02, AU04 and smile examples is 2,587 (1.5%), 2,274 (1.4%) and 37,623 (22.4%) respectively. We use the SVM RBF trained with 80,000 training examples (these examples are completely independent of the AM-FED data) selected using smart selection from the active learning data (independent from the AM-FED data). Table 1 shows the area under the ROC curve for our classifiers compared to the baselined presented in [11]. For the AU02 and Smile classifiers we observed improvements (17% and 4% respectively). The eyebrow furrow classifier performed similarly, perhaps due to the challenging nature of the AU04 examples in this dataset - some of the examples were very subtle.

7. Conclusion and Future Work

Traditionally, facial expressions recognition systems have been trained on datasets of limited size due to the time and expense required to collected hand-labeled positive training examples. Using an Internet-based framework we have collected over 1.8 million videos of spontaneous facial responses to online media. Although occurrences of facial actions are sparse within this data, our active learning approach has allowed us to acquire hand-labeled positive examples from many different individuals up to 20x faster. In total we collected 27,000 sequences of expressions.

Using this data generated using active learning, we were able to significantly increase the accuracy of our action unit recognition system. We present state-of-art results on three actions AU02 (eyebrow raise), AU04 (eyebrow lowered) and smiles. We show the effect of training data size on classifier performance, both accuracy and classification time, using the largest dataset of naturalistic and spontaneous facial expressions in the world. On this challenging data the number of examples from different individuals has a large impact on the overall performance of each classifier.

We propose a novel data sampling strategy to extract samples from each expression sequence within a video, rather than using all the frames (many of which would be very similar) that would make training and validation very time consuming. We propose to apply the Nyström RBF kernel approximation that greatly improves the trade-off between accuracy and classification time. Experiments show that this is more accurate than both a linear SVM, or an RBF-kernel SVM that is trained on a small subset of the data in order to perform at high speeds ($\sim$ 300 fps). The classifier speeds reported are for the classification of one action from the HOG features (on a fast computer). Face detection, tracking and the detection of other AUs will slow the process down. For real-time applications, we found that we need 300FPS on the server and 1000FPS on mobile. This motivated our design of an efficient approximation.

We evaluate the model against a public dataset of challenging spontaneous and naturalistic expressions. We obtained a 17% performance improvement for eyebrow raise and a 4% improvement for smile over the previously published baseline on this spontaneous web-cam data.

We plan to extend the approaches described in this paper to a greater number of facial action unit classifiers. In particular, we anticipate that the active learning approach will be even more effective for action units that occur even more infrequently than smiles, eyebrow raises and brow furrows.

Using the approximated RBF kernel method has another advantage not exploited in this paper. As we first map our features to a new space and then learn a linear SVM, we could benefit from stochastic gradient descent (SGD) optimization which we would allow us to train our system faster and without being limited by memory. So far, using a smart selection of our data still allows us to train even an RBF-kernel SVM. However, as the amount of data is always increasing, our next step is to train the approximated RBF-kernel SVM using SGD optimization. We also plan to explore different approaches, like deep learning, that would leverage a large-scale action-unit coded database.

References

[1] 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*, pages 203–221, 2007.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] F. De la Torre, T. Simon, Z. Ambadar, and J. Cohn. Fast-facs: A computer-assisted system to increase speed and reliability of manual facs coding. *Affective Computing and Intelligent Interaction*, pages 57–66, 2011.

[4] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[5] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *Cybernetics, IEEE Transactions on*, 44(2):161–174, 2014.

[6] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and gen-

uine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009.

[7] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[8] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011.

[9] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013.

[10] D. McDuff, R. El Kaliouby, and R. Picard. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3(4):456–468, 2012.

[11] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. Affectiva-mit facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected" in-the-wild". In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 881–888. IEEE, 2013.

[12] D. McDuff, R. El Kaliouby, T. Senechal, D. Demirdjian, and R. Picard. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing*, 32(10):630–640, 2014.

[13] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

[14] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer, 2008.

[15] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[16] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, volume 2, page 3, 2011.

[17] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):39–51, 1998.

[18] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001.

[19] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[20] F. Torre and J. Cohn. Facial expression analysis. *Visual Analysis of Humans*, pages 377–409, 2011.

[21] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. Int'l Conf. Language Resources and Evaluation, W'shop on EMOTION*, pages 65–70, 2010.

[22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[23] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2106–2111, 2009.

[24] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.

[25] R. Yan, L. Yang, and A. Hauptmann. Automatically labeling video data using multi-class active learning. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 516–523. IEEE, 2003.

[26] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in neural information processing systems*, pages 476–484, 2012.

[27] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[28] Z. Zhang and B. Schuller. Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *INTERSPEECH*, 2012.

[29] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection?. In *BMVC*, volume 3, page 5. Citeseer, 2012.