Exemplar Hidden Markov Models for Classification of Facial Expressions in Videos

Karan Sikka
Univ. of California San Diego
California, USA
ksikka@ucsd.edu

Abhinav Dhall
Univ. of Canberra, Australian National University
Australia
abhinav.dhall@anu.edu

Marian Bartlett
Univ. of California San Diego
California, USA
mbartlett@ucsd.edu

Abstract

Facial expressions are dynamic events comprised of meaningful temporal segments. A common approach to facial expression recognition in video is to first convert variable-length expression sequences into a vector representation by computing summary statistics of image-level features or of spatio-temporal features. These representations are then passed to a discriminative classifier such as a support vector machines (SVM). However, these approaches don't fully exploit the temporal dynamics of facial expressions. Hidden Markov Models (HMMs), provide a method for modeling variable-length expression time-series. Although HMMs have been explored in the past for expression classification, they are rarely used since classification performance is often lower than discriminative approaches, which may be attributed to the challenges of estimating generative models.

This paper explores an approach for combining the modeling strength of HMMs with the discriminative power of SVMs via a model-based similarity framework. Each example is first instantiated into an Exemplar-HMM model. A probabilistic kernel is then used to compute a kernel matrix, to be used along with an SVM classifier. This paper proposes that dynamical models such as HMMs are advantageous for the facial expression problem space, when employed in a discriminative, exemplar-based classification framework. The approach yields state-of-the-art results on both posed (CK+ and OULU-CASIA) and spontaneous (FEEDTUM and AM-FED) expression datasets highlighting the performance advantages of the approach.

1. Introduction

Automatic facial expression recognition (AFER) enables machines to understand a form of human behavior, and can be used towards building intelligent systems [4, 11, 12]. Research efforts in the last decade have led to significant improvements in AFER and have also opened new research avenues [4]. In particular AFER research is transitioning from expression recognition in images [21, 15] to recognition in videos [20, 29]. Image based methods, sometimes referred to as static approaches, utilize visual features from a snapshot (such as the apex expression in a sequence [29, 24]) for predicting facial expressions [20, 15]. In image-based approaches, expression dynamics are not explicitly incorporated in the features or in the classifier, and instead are analyzed in the time series of the output. In addition, image-based approaches often assume temporal segmentation of facial expressions for training. Since facial expressions are dynamic events, video based approaches that incorporate dynamics earlier, in the features or in the model, may have an advantage over image-based approaches, and have been shown to outperform their image based counterparts on several AFER problems [9, 26].

Video based approaches can be roughly categorized into space-time and sequential approaches. In space-time techniques, localized features (in space and time) such as Bag of Words (BoW) [21], fiducial point positions [18, 8], and LBPTOP [29], are first extracted across the entire video. This is followed by applying spatio-temporal pooling operations (such as average or maximum [20]) over the entire video [8] or fixed grids [29] to obtain a fixed length vector representation [20]. These fixed-length vectors are then passed to a classifier. Owing to the use of discriminative classifiers such as Support Vector Machines (SVM), we shall refer to these approaches as Discriminative space-time (Disc-ST) methods. Disc-ST methods can extend image-based features to video based methods by summarizing the per-frame features over the entire video using summary statistics [20]. This is known as pooling across the temporal dimension. Although Disc-ST methods such as LBPTOP are used frequently for AFER, we highlight two inherent issues that can result in a performance loss. These are:

1. Discriminative power: Pooling features across the entire video (or pre-defined grids) works well for pre-segmented video clips. However, the approach can lack discriminative power for unsegmented video, which can contain more than one expression as well as neutral periods. For instance the video clip shown in Figure. 1, is composed of angry expression segment as well as neutral expression. Since the final features summarize the entire video, the performance might degrade when some sub-segments are non-informative. For the problem of pain classification in unsegmented videos, Sikka et al. [20] showed that Disc S-T approaches yielded a lower performance than methods utilizing information from task specific sub-segments.

2. Temporal alignment: Facial expression in a video can be characterized as a dynamic event that passes through several states. A recognition pipeline would ideally match corresponding states, which requires temporal alignment between sequences [9]. A fixed feature pooling strategy ignores this correspondence, and doesn't capture temporal relationships between these states.

Sequential approaches [3, 24, 19] present an alternate strategy for analyzing facial expressions in videos. Such models first convert a video into a sequence of observations at regular intervals and then analyze the sequence for presence of action-specific features and their dynamics. This work focuses on Hidden Markov Models (HMMs) that can describe a facial expression as a dynamic event comprising of several sub-events (apex, onset and offset) with specific temporal relationships between them [22]. Since a latent state variable is associated with each observation, HMMs naturally define a temporal segmentation of the video [22, 3]. The training routine of HMM consists of estimating class-conditional probability distributions for each class. A test video is then classified into the class that corresponds to maximum posterior probability.

Because HMMs offer a form of temporal segmentation and alignment, HMMs provide modeling advantages that may lead to better classification performance than Disc-ST methods for AFER. However HMMs, being generative models, are often weaker classifiers than discriminative models since estimating probability distributions is a harder problem than solving the classification problem directly [14]. As a result, generative HMMs generally have a lower performance compared to Disc S-T approaches [9, 24], and are seldom used despite their modeling capabilities. A possible solution towards overcoming the disadvantages of HMMs is to estimate the model parameters within a discriminative learning framework [1, 16]. Approaches based on this idea, such as Hidden Conditional Random Fields, have previously been used for AFER [16]. This paper follows an alternative solution that focuses on estimating similarity kernels for probability distributions [7, 6]. Such approaches allow the possibility to measure meaningful distances between probabilistic models describing non-vectorial data.

We argue that when embedded in a discriminative classification framework based on probabilistic kernels, dynamical models such as HMMs may be effective for the facial expression problem space. An approach from the machine learning literature, generative kernels, combines the modeling advantages of HMMs with the discriminative advantages of SVMs in a principled fashion in an exemplar (or similarity) based classifier [2, 7]. Each example is first abstracted into an individual HMM, which models the spatio-temporal characteristics specific to that example. A distance metric, referred to as a probabilistic kernel [6, 7], is then used for computing distances between two Exemplar-HMMs. The probabilistic kernel can also be visualized as a distance between two data points lying on some non-linear manifold of HMM models. These distances are then used as an input to a kernel SVM. In this work we employed the Probabilistic Product Kernel (PPK) [7] since it provides a closed-form solution for HMMs and can also be interpreted as estimating distances between the temporal segments (or states) of two videos while taking into account the transition probabilities. Exemplar-HMMs with probabilistic product kernels have been shown to be effective for clustering motion capture data [7] and recognizing handwritten words [17]. Here we explore this class of models in the problem space of facial expression recognition in video. We shall refer to this approach as Exemplar-HMMs.

The performance advantage of our approach was exhibited through an evaluation on both posed and spontaneous facial expression datasets. The AFER problems in these datasets ranged from predicting basic emotions to predicting whether a commercial presented over the internet was liked. On each of these AFER problems, our approach achieved state-of-the-art results compared to its Disc-ST counterparts and also in comparison to recently proposed AFER algorithms that exploit facial expression dynamics.

2. Related Work

This work is motivated by the idea of using model-based similarity for measuring distances between non-vectorial data such as time-series or sets of vectors [7, 2, 6, 17]. The distances are computed by (i) mapping each vector set into a probability distribution, and (ii) using a probabilistic measure to compute distances between example specific models. Availability of distances between examples allows one to define a kernel and use discriminative classifiers such as SVMs. The primary benefit underlying these approaches is the possibility of including example-specific structural information during classification. This hybrid discriminative-

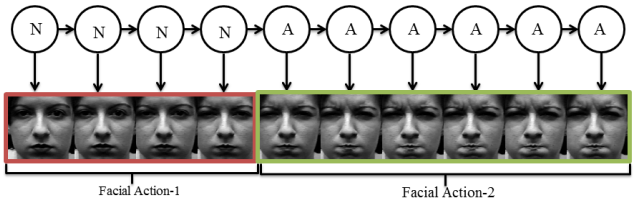

Figure 1: Shows a facial expression video modeled via an HMM, where hidden states (shown on top) are assigned to each observation, while forming a Markov Chain. It depicts that the HMM modeling is able to represent the video as comprised of two distinct sub-events (neutral and apex).

generative framework has shown its advantages over holistic feature based approaches on several problem such as handwriting recognition [17], gene classification [7], and shape recognition [2], among others.

Previous work has identified the importance of exploiting spatio-temporal structure for classifying facial expressions. In a recent paper by Liu et.al [9], the authors proposed mid-level representations, referred to as Expressionlets, that were obtained by aligning localized spatio-temporal features from an input video with a universal Gaussian Mixture Model (GMM) model. The authors argued that via localized alignment, their approach (STM-ExpLet) allows flexible spatio-temporal range among low-level features and is able to achieve improvements over Disc-ST approaches. Our approach is motivated on a similar argument as [9], however it differs on two points. Firstly, we incorporate temporal information via the model (instead of low-level features) and secondly the temporal flexibility is derived from the application of a probabilistic kernel.

3. Methods

3.1. Problem Statement

The training data from a facial expression dataset D with N samples is represented as $D = \{X_i, Y_i\}_{i=1}^{N}$, where X_i is a facial expression video and $Y_i \in Y$ is its class label. Each video is a sequence of images $\{x_{it}\}_{t=1}^{T_i}$, where each $x_{it} \in \mathcal{R}^d$ is represented either in native pixel intensity space or feature space. The final goal of AFER is to predict the class label for an unseen test sequence. We first briefly describe the HMM framework being used for spatio-temporal modeling of facial expression sequences, followed by a description of probabilistic kernels. Thereafter we propose our method, referred to as Exemplar-HMMs, for the AFER task.

3.2. HMMs for AFER

HMM is a parametric model that is used to statistically describe a time-series under Markov assumption. In particular, the HMM models the joint probability of a time-series X_i as a chain of observations x_{it} and corresponding discrete (unobserved) hidden state z_{it}. For example, in Figure. 1, the dicrete hidden states are N and A, and the observations are the individual snapshots shown at the bottom.

In the context of AFER, an HMM can be visualized as a spatio-temporal model describing a video as a chain of discrete sub-events. Each hidden state in an HMM model is essentially an abstraction for a sub-event and describes a distinct expression state such as neutral or a particular cluster in the space of facial actions (see Figure. 1). Thus an HMM model consists of two parts (1) dynamic (temporal): describes the transitions between distinct facial states, and (2) appearance (spatial): describes the observation space that characterizes each sub-event. The observation space was modeled using multivariate Gaussian distribution with diagonal Covariance matrix. We found this modeling assumption to work well since the Mahanalobis distance used in Gaussian distribution serves as a good metric for the facial features used in our experiments and have also been shown to work well in previous AFER approaches [3].

HMM modeling consists of estimating parameters $\theta = \{\pi_k, A_k, \mu_k, \Sigma_k\}_{k=1}^{N_s}$, where $\pi_k = Pr(z_1 = k)$ is the initial probability of being in state k, $A_{k,j} = Pr(z_t = k|z_{t-1} = j)$ is the transition probability (stationary) of transitioning from state j to state k, μ_k and Σ_k are the mean vector and diagonal covariance matix respectively of the Gaussian distribution corresponding to state k, and N_s is the number of hidden states. The parameters for an HMM are estimated using maximum likelihood paradigm via the expectation maximization (EM) algorithm. For effectively learning the model parameters in current setting we made use of the Bayesian formulation for HMM as described below.

Bayesian HMM: In contrast to learning an HMM from a set of examples, this work involved learning an HMM model from a single example. Since HMMs involve many parameters, learning parameters using only one example could result in severe over-fitting. Thus to obtain a robust solution, we employed a Bayesian solution to this problem by incorporating (conjugate) prior distributions for different parameters, which can effectively regularize the EM solution and avoid over-fitting [14, 5]. Normal Inverse-Wishart distribution was used as a prior for Gaussian parameters, and the Dirichlet distribution was used for both the parameters of initial and transition probabilities. The parameters are then estimated by using maximum a posteriori (MAP) estimation along with the EM algorithm (MAP-EM). Interested readers are refered to [14, 5] for more details. We found during our experiments that this formulation not only improved the reliability of model estimation but also improved results by allowing to learn Gaussian observation states with diagonal covariance structure.

3.3. Probabilistic Kernels

Parametric modeling approaches, such as Gaussian classifiers and logistic regression represent each object as a fixed-length feature vector followed by learning the parameters of the model. However, a fixed-length representation might not be the natural choice for objects such as variable length sequences, probability functions or high-dimensional objects [17, 7]. An alternate strategy is to design algorithms that are based on measuring the similarity between pairs of data-points, without ever requiring the explicit feature representation of each data point. Examples of similarity functions include distances or inner products between X and X', where X and X' lie in a space S. We denote $\mathcal{K}(X, X') \geq 0$ a kernel function. Learning approaches based on computing kernel functions between data points are examples of non-parametric methods and are generally referred to as kernel methods.

In the present work, we represent each datum as a probability distribution and thereafter use a kernel, referred to as probabilistic kernel, to compare distances between two distributions. For the task of AFER, we employ the Probability Product Kernel (PPK), proposed by Jebara et al., for computing distances between two distributions. PPK is computed between two distributions from the same family with parameters θ and θ' as:

$$\mathcal{K}_{PPK}(\theta, \theta') = \int_{X_{1:T}} \Pr(X_{1:T}|\theta)^{\rho} \Pr(X_{1:T}|\theta')^{\rho} dX_{1:T} \tag{1}$$

$X_{1:T}$ refers to a sequence of length T. Here parameter ρ controls the non-linearity of these kernels, while parameter T determines the temporal extent to which two models shall be compared. These parameters are important for calculating the distance and are tuned using cross-validation for each dataset (see Section. 4). PPK can be interpreted as an inner-product between two probability distributions and is a positive definite kernel. Alternate probabilistic kernels, such as Fischer kernel or heat kernels, also exist. However in this work we concentrate on PPK owing to its computational feasibility and nonlinear flexibility [7]. Also in contrast to the popular kernel based on Kullback-Leibler Divergence, PPK is better able to provide a closed form solution for HMM models.

3.4. Exemplar-HMMs for AFER

Revisiting the AFER problem discussed in Section. 3.2, the proposed approach begins by learning an HMM model (with parameters θ_i) for each example X_i. The next step computes a kernel matrix $\mathcal{K} \in \mathcal{R}^{N \times N}$ for the training data, whose elements are computed as $\mathcal{K}(i, j) = \mathcal{K}_{PPK}(\theta_i, \theta_j)$. The PPK kernel can be computed efficiently using factorization of HMM as mentioned in [7]. The kernel matrix is then normalized using a standard normalization procedure:

$$\mathcal{K}_{norm}(i, j) = \frac{\mathcal{K}(i, j)}{\sqrt{\mathcal{K}(i, i)}\sqrt{\mathcal{K}(j, j)}} \tag{2}$$

We then use the kernel matrix K_{norm} to train a support vector machine, using the libsvm library.[1]. The learned classifier can then be used to assign a classification score to a test sequence (X_t). This score is calculated based the similarity of its model $\Pr(X|\theta_t)$ to the models corresponding to the support vectors identified from the training sequences. During our experiments we keep N_s (the number of states in the HMM) constant while learning HMM models for a dataset. The value of N_s was estimated using double-cross validation (see Section 4) for each dataset.

Intuition behind PPK kernel: Before proceeding to the experiment section, it is important to have an intuitive understanding of the model similarity estimated by the PPK kernel. The PPK distance consists of two aspects (1) static: which computes a probabilistic similarity between different state-wise Gaussian distributions of the two HMM models, and (2) dynamic: incorporates transition probabilities and calculates similarity for joint state transitions. The PPK kernel principally combines both these aspects into a recursive formulation that finally calculates the probability of all possible state evolution undertaken by the two HMM models together. The parameter T is important since as T increases, the distance is dominated by the terminal states of the two HMMs as governed by the transition probabilities. For example, in the case of two expression sequences starting at neutral and ending at apex, a higher T will result in having more contribution from the distances between the apex states.

4. Experiments

Performance was evaluated was using two sets of experiments. The first set consisted of datasets containing basic emotions such as anger and sadness, and were captured under laboratory settings. To further highlight the performance advantages, we evaluated our algorithm on more a challenging AFER problem, where the facial expressions were spontaneous, and captured under more naturalistic settings.

PPK parameters: The PPK consists of two hyper-parameters ρ and T. In order to estimate the value of hyper-parameters without over-fitting on the test set, a double cross-validation (CV) protocol was employed. Double-cross validation simulates separate training, validation, and test sets, where the hyper-parameters are selected based on the validation set, and then evaluated on a separate test set. In this protocol, there are two nested cycles of CV. In the outer CV cycle, a set of test data is held out. Then an

[1] SVM is extended to multiclass classification using the one-vs-all strategy.

inner cycle of CV is conducted in order to select hyper-parameters. We then select the hyper-parameters that yield the highest average accuracy across all validation folds. System performance using those hyperparameters is then evaluated on the held out test set. The value of the two hyper-parameters was kept constant across a facial expression dataset.

4.1. Experiments on Posed and Spontaneous Basic Emotions

In this experiment a video was described as a **time-series** of facial landmark points [18, 8]. For every sequence, 49 landmarks points were obtained for each frame by using supervised gradient descent approach [25]. Displacement features for each frame were then obtained by subtracting x and y coordinates of the landmark points in that frame from the landmark coordinates in the first (neutral frame) in that video and concatenating these displacements into a single vector (dimensionality 98) [18]. A linear subspace was separately computed for each dataset by using these displacement features along with Principal Component Analysis (PCA) algorithm [14]. The features were then projected to a low dimensional subspace (of dimensionality d_{pca}) composed of principal components that preserved 99.5% variance. [2]

CK+ Dataset: CK+ [10] is a standard AFER benchmark dataset consisting of 593 sequences from 123 subjects. These subjects were asked to perform a series of 23 facial displays, of which 327 sequences (118 subjects) were categorized into one of the seven basic emotion- anger, disgust, fear, happiness, sadness, surprise and contempt. The length of the sequences in this dataset varies from 10 to 60 frames and the facial expression transitions from onset (neutral phase) to apex phase. d_{pca} in this case was 46 and the experiments were conducted using the leave-one-subject out protocol [10], where each fold consisted of data from just one subject. The prediction task was to predict the class of an unseen test sample and the performance metric being reported is average accuracy for the seven classes. The cross-validated values for number of HMM states was $N_s = 2$ (neutral and apex) and kernel parameters were $\rho = 0.8$ and $T = 35$.

Oulu-CASIA VIS Dataset: The Oulu-CASIA VIS dataset is a subset of the Oulu-CASIA NIR-VIS dataset [28], in which all videos were captured under the visible (VIS) light condition. It consisted of 480 samples from 80 subjects displaying one of the six basic emotion- anger, disgust, fear, happiness, sadness and surprise. The subjects were asked to make a facial expression matching an expression sequence shown on a monitor. The length of the sequences varies from 9 to 72 frames and each video begins at neutral expression and ends at apex expression (similar to CK+). d_{pca} in this case was 54. The experiments were conducted in a subject-independent format by using a 10 fold CV [9]. The evaluation task was to predict the class of an unseen test sample and the performance metric being reported is average accuracy for the six classes. The cross-validated values for number of HMM states was $N_s = 2$ (neutral and apex) and kernel parameters were $\rho = 0.9$ and $T = 30$.

FEEDTUM: The FEEDTUM facial expression dataset [23] consists of spontaneous facial expressions elicited by presenting the subjects with a set of carefully selected video stimuli. This is different from both CK+ and Oulu-CASIA VIS datasets, where the subjects were asked to perform specific facial movements. Moreover, in most cases the facial expressions sequences evolved from neutral to apex and then back to neutral, and the variability of the duration and timing of these phases was higher. d_{pca} in this case was 52. The dataset contains 19 subjects (320 videos) showing six basis emotions- anger, disgust, fear, happiness, sadness and surprise. As above, each video was described as a time series of facial landmark points computed using supervised gradient descent. The experiments were conducted using leave-one-subject out protocol, where each fold consisted of data from just one subject. The evaluation task was to predict the class of an unseen test sample and the performance metric being reported is average accuracy for the six classes. The cross-validated values for number of HMM states was $N_s = 3$ and kernel parameters were $\rho = 0.6$ and $T = 11$.

4.2. Experiments on Facial Action Time Series

To further evaluate the effectiveness of our approach, an additional set of experiments was conducted on another spontaneous facial expression dataset, the AM-FED dataset [12]. The AFER problem on this dataset was more challenging since the expression videos were not only unsegmented but also lacked prior information about the onset, duration and frequency of the target expression. Since the videos in this dataset involved large out-of-plane head movements, we were unable to use automatic facial landmark points as the frame-level feature representation as was done in the previous experiments. Instead we made use of the frame-level action unit (AU) annotations provided with this dataset for the time-series data. Action Units correspond roughly to movement of individual facial muscles. The focus of this paper is on modeling sequences, and not on feature extraction. The AU data provides a spontaneous biological time series, where the features are highly precise, thereby providing an alternate way to evaluate the sequence classification approach.

AM-FED: The AM-FED dataset consists of 242 sequences that were recorded on a web-cam while different

[2] d_{pca} for each dataset varied from $46 - 54$, which means that due to PCA feature dimensionality was reduced by almost 50%.

subjects were viewing three Superbowl commercials. After watching the videos, the subjects provided self-report ratings for two questions: "Did you like the video?" and "Would you like to watch this video again?". The self-report responses could be positive (1), neutral (0) or negative (-1). Similar to the protocol used in [13], only videos where the labels are either positive or negative were included and the target was to predict these binary self-report responses given a test video. The final dataset consisted of 103 and 170 sequences for "Watch/Not Watch again" and "Like/Does not like" prediction tasks respectively. In line with the evaluation protocol described in [12], the experiments were conducted in a (3 fold) leave-one-advertisement-out protocol, where the videos corresponding to one advertisement were used for testing and remaining for training. Average AUC score across all folds is reported as the evaluation metric. Of the frame-level annotations provided with this dataset, we discarded annotations for those facial actions that were either available for a few examples or had a low inter-coder reliability. The final frame-level representation comprised of annotations for AU 2, 4, 5, 14, 17, Unilateral left AU 12, Unilateral right AU 12, Negative AU 12 and Unilateral left AU 14 along with smile and expressability ratings (dimensionality was 11). In this dataset the value for each AU was the percent of annotators indicating presence of the AU in the frame. We thresholded all AU annotations $< 50\%$ to 0 and $\geq 50\%$ to 1. The cross-validated values for number of HMM states for task "Like/Does not like" was $N_s = 2$ and kernel parameters were $\rho = 1.2$ and $T = 35$, while for task "Watch/Not Watch again" was $N_s = 3$ and kernel parameters were $\rho = 0.6$ and $T = 5$.

4.3. Algorithms compared

We compared Exemplar-HMMs to three baseline algorithms. The first is a standard Disc-ST pipeline, where the frame-level features across a video are summarized using Mean-pooling and Max-pooling. The second baseline algorithm is LBPTOP, which is amongst the most common Disc-ST approaches for AFER. For implementing LBP-TOP, spatio-temporal features were extracted from non-overlapping blocks and concatenated into a single vector. The features from Disc-ST pipelines and LBPTOP were passed as input to a SVM classifier with rbf kernel. The third baseline algorithm is a generative HMM classifier (Section. 1). Same frame-level features were used in Disc-ST pipelines and HMM to facilitate a fair comparison. The parameters for HMM (number of states), LBPTOP (size of non-overlapping blocks) and rbf kernel (kernel width) were determined using the double CV procedure described in Section. 4. In addition, we have also provided the performance of a recent state-of-the-art algorithm [9] for making a relative comparison to our algorithm.

5. Results and Discussion

The results for CK+ and OULU-CASIA VIS datasets are shown in Table. 1a, AM-FED dataset in Table. 1b and FEEDTUM dataset in Table. 2. The results are divided into two blocks (using double lines), into Disc-ST methods and dynamic methods, where the term 'dynamic' refers to approaches explicitly incorporating temporal information inside the algorithm. Our contention that Exemplar-HMMs is able to overcome the limitations of HMMs by combining its modeling advantages with a discriminative classification paradigm is supported by the comparison to the standard HMM. The proposed approach outperforms the standard HMM on all the datasets by an appreciable margin. For example, the absolute performance improvement is approximately 10% accuracy for CK+ and OULU-CASIA, and 5% for FEEDTUM.

Our approach outperforms the baseline Disc-ST method with Max-pooling and Mean-pooling on all datasets. The performance advantage for Exemplar-HMMs is greatest for the AM-FED spontaneous expression dataset. For instance, on the Like/Don't Like task, the AUC score for Mean-pooling and Max-pooling are .66 and .61 respectively, compared to .84 for Exemplar-HMMs, and for the Watch-again/Don't Watch-again task, the AUC score for our approach is .92 compared to .87 and .89 for Mean-pooling and Max-pooling respectively. This increase in performance can be attributed to the ability of Exemplar-HMMs to address the key modeling issuess in Disc-ST methods (1) discriminative information, and (2) temporal alignment, as discussed in Section. 1. We further observed that our approach outperformed LBPTOP method on all three emotion datasets. This is particularly interesting since LBPTOP utilizes image textures that may be capable of capturing more information than the geometric features [27] employed in our method. Moreover the (absolute) performance increase is greater for the spontaneous expression dataset FEEDTUM (5.9% hike) than the posed datasets CK+ (3.3%) and OULU-CASIA VIS (3.5%). One possible explanation is that LBPTOP doesn't take into account the temporal structure inherent in a time-series, and this information might be important for expression recognition on spontaneous expressions such as those in FEEDTUM.

We also compared the performance of Exemplar-HMMs with a recent (2014) state-of-the-art algorithm (STM-ExpLet [9]) that explicitly adds temporal information inside the algorithm. Exemplar-HMMs achieved 75.62% accuracy on OULU-CASIA VIS dataset, which was similar to the accuracy of STM-ExpLet (74.59%). For CK+ our results were calculated using the standard leave-one-subject-out protocol as mentioned in [10], while 10 fold CV was used to calculate results in STM-ExpLet paper. Thus for making a fair comparison we ran a 10 fold CV (randomized over 10 runs) for CK+ and obtained an accuracy of 93.89% com-

Method	**CK+**	**OULU-CASIA VIS**
Geom. + Mean-pooling	92.42 ($\pm$1.58)	70.83 ($\pm$2.84)
Geom. + Max-pooling	92.74 ($\pm$1.67)	69.76 ($\pm$1.73)
LBPTOP [29]	91.30($\pm$1.79)	72.08($\pm$2.22)
HMM	85.35($\pm$2.16)	63.54($\pm$3.10)
STM-ExpLet [9]	94.19 (N/A)	74.59 (N/A)
Exemplar-HMMs	**94.60**($\pm$**1.37**)	**75.62**($\pm$**2.10**)

(a) Results (mean % accuracy) for CK+ and OULU-CASIA VIS

Method	**Like/Don't Like**	**Watch-again/ Don't Watch-again**
AU + Mean-pooling	.66	.87
AU + Max-pooling	.61	.89
HMM	.58	.84
Exemplar-HMMs	**.84**	**.92**

(b) Results (mean area under the ROC curve) for AM-FED.

Table 1: Results from experimental evaluation.

pared to 94.16% in STM-ExpLet. Although the results from our algorithm seems comparable to the state-of-the-art, it is important to note that our algorithm was based on geometric features while STM-ExpLet used 3D texture based appearance features, which are capable of capturing more information and have also previously been shown to outperform geometric features [27]. Thus the observation that the present classification paradigm is able to match the state-of-the-art results on current AFER problems despite using simpler features highlights both its performance advantages and the promising nature of the present research direction for tackling AFER. In order to extend current approach to texture based features, we have discussed our ongoing work to exploit high-dimensional features in the current pipeline in the next section.

6. Conclusion and Future Work

This paper builds upon the idea that facial expressions have a specific temporal structure that can be explicitly modeled using latent variable sequential models. Focusing on Hidden Markov Models, we argue that they provide certain modeling benefits over temporally holistic feature based approaches for facial expression recognition. However owing to their generative nature, HMMs typically have a lower classification performance than discriminative classifiers such as Support Vector Machines (SVMs). This paper explored an approach for combining the modeling strength of HMMs with the discriminative power of SVMs via probabilistic kernels for the task of facial expression recognition. This combination was achieved by modeling each example with an HMM, followed by computing a kernel matrix, via Probabilistic Product Kernels, that comprised the input to an SVM. By achieving state-of-the-art results on both posed and spontaneous datasets, this approach highlighted its performance advantage for video-based facial expression recognition compared to traditional HMMs, and compared to discriminative approaches based on temporally holistic features.

This preliminary work showed both the modeling and performance advantages of an approach relying on model-based similarity for AFER. Building on this line of research, our current focus is on extending both HMMs and probabilistic kernels to handle high-dimensional spatio-temporal features that can further enhance its modeling advantages. This is being accomplished by the application of the kernel trick for embedding data-points in an implicit projection space, and then extending HMMs and probabilistic measures to work with kernels. We shall also explore the possibility of using other probabilistic kernels (Fisher, KL-divergence etc.) as part of our future work.

Method	Accuracy %
Geom. + Mean-pooling	48.91($\pm$3.70)
Geom. + Max-pooling	53.87($\pm$2.59)
LBPTOP [29]	48.17($\pm$3.31)
HMM	48.23($\pm$3.88)
Exemplar-HMMs	**54.14**($\pm$**3.72**)

Table 2: Results (mean % accuracy) from FEEDTUM.

7. Acknowledgment

Support for this work was provided by NIH grant NIH R01NR013500. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agency. We would also like to thank Mohsen Malmir and Ritwik Giri for helpful discussions.

References

[1] Y. Altun, I. Tsochantaridis, T. Hofmann, et al. Hidden markov support vector machines. In *International Conference on Machine Learning*, volume 3, pages 3–10, 2003. 2

[2] M. Bicego, V. Murino, and M. A. Figueiredo. Similarity-based clustering of sequences using hidden markov models. In *Machine learning and data mining in pattern recognition*, pages 86–95. Springer, 2003. 2, 3

[3] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1):160–187, 2003. 2, 3

[4] F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual Analysis of Humans*, pages 377–409. Springer, 2011. 1

[5] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and audio processing*, 2(2):291–298, 1994. 3

[6] T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999. 2

[7] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *The Journal of Machine Learning Research*, 5:819–844, 2004. 2, 3, 4

[8] L. A. Jeni, D. Takacs, and A. Lorincz. High quality facial expression recognition in video streams using shape related information only. In *IEEE International Conference on Computer Vision Workshops*, pages 2168–2174. IEEE, 2011. 1, 5

[9] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *IEEE Computer Vision and Pattern Recognition*, pages 1749–1756. IEEE, 2014. 1, 2, 3, 5, 6, 7

[10] P. Lucey, J. F. 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 *IEEE Computer Vision and Pattern Recognition Workshops*, pages 94–101. IEEE, 2010. 5, 6

[11] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *IEEE Automatic Face and Gesture Recognition Workshops*, pages 57–64. IEEE, 2011. 1

[12] 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 *IEEE Computer Vision and Pattern Recognition Workshops*, pages 881–888. IEEE, 2013. 1, 5, 6

[13] 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. 6

[14] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 2, 3, 5

[15] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 1

[16] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1848, 2007. 2

[17] J. A. Rodríguez-Serrano, F. Perronnin, J. Lladós, and G. Sánchez. A similarity measure between vector sequences with application to handwritten word image retrieval. In *IEEE Computer Vision and Pattern Recognition*, pages 1722–1729. IEEE, 2009. 2, 3, 4

[18] A. Saeed, A. Al-Hamadi, and R. Niese. The effectiveness of using geometrical features for facial expression recognition. In *IEEE International Conference on Cybernetics*, pages 122–127. IEEE, 2013. 1, 5

[19] L. Shang and K.-P. Chan. Nonparametric discriminant hmm and application to facial expression recognition. In *IEEE Computer Vision and Pattern Recognition*, pages 2090–2096. IEEE, 2009. 2

[20] K. Sikka, A. Dhall, and M. S. Bartlett. Classification and weakly supervised pain localization using multiple segment representation. *Image and Vision Computing*, 32(10):659 – 670, 2014. Best of Automatic Face and Gesture Recognition 2013. 1, 2

[21] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision, Workshops and Demonstrations*, volume 7584, pages 250–259. Springer Berlin Heidelberg, 2012. 1

[22] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(1):28–43, 2012. 2

[23] F. Wallhoff. Feedtum facial expression and emotion dataset, 2004. 5

[24] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *IEEE Computer Vision and Pattern Recognition*, pages 3422–3429. IEEE, 2013. 1, 2

[25] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Computer Vision and Pattern Recognition*, pages 532–539. IEEE, 2013. 5

[26] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas. Facial expression recognition using encoded dynamic features. In *IEEE Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1

[27] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *IEEE Automatic Face and Gesture Recognition*, pages 454–459. IEEE, 1998. 6, 7

[28] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 5

[29] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. 1, 7