Learning a Sparse Codebook of Facial and Body Microexpressions for Emotion Recognition

Yale Song
MIT CSAIL
Cambridge, MA 02139
yalesong@csail.mit.edu

Louis-Philippe Morency
USC ICT
Los Angeles, CA 90094
morency@ict.usc.edu

Randall Davis
MIT CSAIL
Cambridge, MA 02139
davis@csail.mit.edu

ABSTRACT

Obtaining a compact and discriminative representation of facial and body expressions is a difficult problem in emotion recognition. Part of the difficulty is capturing microexpressions, i.e., short, involuntary expressions that last for only a fraction of a second: at a micro-temporal scale, there are so many other subtle face and body movements that do not convey semantically meaningful information. We present a novel approach to this problem by exploiting the sparsity of the frequent micro-temporal motion patterns. Local space-time features are extracted over the face and body region for a very short time period, e.g., few milliseconds. A codebook of microexpressions is learned from the data and used to encode the features in a sparse manner. This allows us to obtain a representation that captures the most salient motion patterns of the face and body at a micro-temporal scale. Experiments performed on the AVEC 2012 dataset show our approach achieving the best published performance on the expectation dimension based solely on visual features. We also report experimental results on audio-visual emotion recognition, comparing early and late data fusion techniques.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Computer Vision*

Keywords

Audio-visual emotion recognition; microexpressions; spatio-temporal interest points; dictionary learning; sparse coding; data fusion.

1. INTRODUCTION

Humans express thoughts and emotions through multiple modalities, including speech, facial expression, and body gestures. Automatic human emotion analysis aims to infer the emotional state of a human from these modalities using techniques in various disciplines, including audio signal processing, speech understanding, computer vision, pattern recognition, and machine learning [50]. The multimodal nature of the problem poses a variety of interesting challenges to these disciplines both individually and collectively, including methods for signal representation and combination [43].

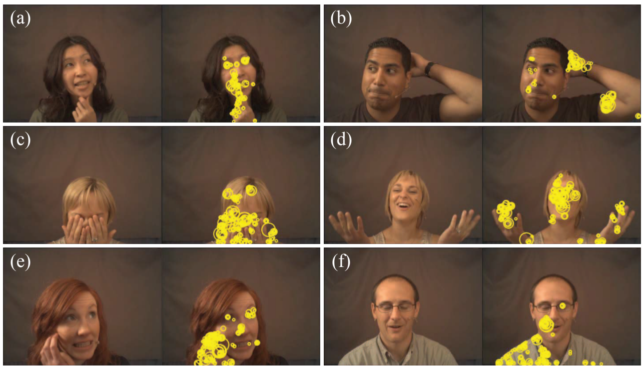


Figure 1: Body gestures provide complementary information about human emotion that may not be available from the face, e.g., (d) an 'open two palms up' gesture indicates high arousal, while the smiling face indicates positive valence. We detect spatio-temporal interest points over the face and body region (shown in yellow circles), and for each interest point extract local appearance-based features. We then use sparse coding to select the most salient patterns among the features. This allows us to obtain a compact, yet discriminative representation of the facial and body microexpressions.

Within the computer vision community, the key challenge is to obtain a compact and discriminative representation that captures how the face and body express emotion. Ekman's [9, 10] suggestion that the face reveals abundant information about human emotion has motivated systems focused on understanding facial expressions, which have been largely successful over the past decade [31, 50]. These systems do not, however, take advantage of information about body gestures, even though psychologists and behavioral scientists suggest that they are an integral part of expressing our thoughts and emotions [13, 17, 26, 8]. It is thus natural to expect that using body gestures, when combined with facial expressions, will improve the robustness of emotional recognition.

But finding an efficient way to represent both facial expressions and body gestures is a non-trivial problem. A common approach is to extract different feature descriptors from face (e.g., action units) and body (e.g., joint positions), and concatenate them into a single feature vector. Unfortunately, this can introduce errors when one of the feature extractors fails because of the face or body occlusion (e.g., see Figure 1 (c)), or when feature descriptors have different statistical properties (e.g., means and variances) and require appropriate feature normalization [38, 42].

ICMI'13, December 9–13, 2013, Sydney, Australia
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522851.

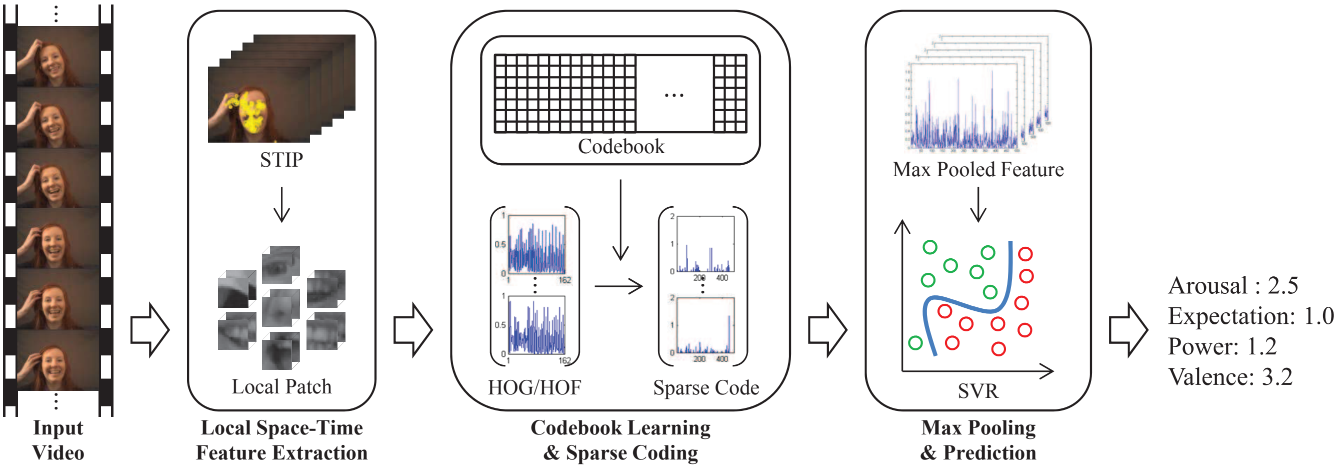


Figure 2: Our approach consists of three steps: extracting local space-time features; learning a codebook of microexpressions from the data and use it to obtain sparse representation of the local features; and performing spatio-temporal max pooling of features. This representation can then be used as an input to a prediction model (e.g., SVR [40]) to infer various emotional states.

Another challenge is capturing *microexpressions*.[1] A microexpression is defined as a short, involuntary facial muscle and body part movements caused by an internal emotional stimulus that lasts only 1/25 to 1/15 of a second [8]. These have been shown to be an important factor in understanding human emotion [11, 8]. Yet, detecting them is difficult because, apart from having to examine the video at a micro-temporal scale, there are many other subtle movements of the face and body that do not convey semantically meaningful information, which means that we need an efficient method to select a few most salient patterns.

We address the two challenges mentioned above – (1) representing both the face and body features and (2) capturing microexpressions – by proposing a compact yet discriminative representation of facial and body microexpressions. Our approach is motivated by the recent success of appearance-based local space-time feature descriptor [19] and a sparse representation of image features [29, 46]. The local space-time feature descriptor has been shown to perform particularly well in action recognition [33], because of its ability to represent the spatio-temporal patterns of short-term body part movements. We extract the local space-time features over the face and body region to describe micro-temporal movement patterns in the video. This makes it unnecessary to deal with the occlusion problem explicitly: instead of representing an occluded face as a “missing” observation, it describes *how* the face is being occluded, e.g., hands are covering the face (see Figure 1 (c)). Unfortunately, mainly because of their granularity, the extracted patterns of microexpressions are much more complex than human actions like walking, running, etc., and we need a method to select a few most discriminative patterns. To this end, we exploit the sparsity of frequent patterns in the extracted features: we learn a codebook of the local space-time features extracted over the face and body regions, and use sparse coding to select the most salient patterns of them. This allows us to obtain a representation that captures discriminative facial and body microexpressions, extracted over a short time period and encoded in a sparse manner.

We evaluated our approach in both unimodal (video only) and multimodal (audio-visual) settings, on a benchmark dataset provided by the second International Audio-Visual Emotion Challenge and Workshop (AVEC 2012) [37], where the goal is to predict continuous values of emotional states in four affective dimensions (arousal, expectation, power, and valence). Results suggest that our method performs particularly well in predicting the arousal dimension, and achieves the best published performance in the literature on the expectation dimension even without using audio and context-related features. On the audio-visual setting, we report that late fusion achieves better performance, and discuss possible reasons for why late fusion might work better than the early fusion in our task.

2. RELATED WORK

There is a large volume of literature on automatic human emotion recognition; here we discuss the most recent and relevant work. For comprehensive surveys, readers are referred to [31, 50, 43].

Perhaps the most popular and well-studied approach is based on facial expression analysis [31, 50]. These approaches detect a region of interest, i.e., the face, then extract features within the region using either appearance-based low-level descriptors [28, 35, 15] or facial landmark point detectors [4, 5]. Recently, Liu *et al.* [22] proposed an alternative approach by automatically learning action unit (AU)-like image features using deep learning (instead of trying to detect AUs directly). For these approaches to work properly, however, the face needs to be visible, without too much occlusion, and needs to be detected reliably so that features can be extracted from it. This may not be possible in natural scenarios in human communication, e.g., when a face is occluded by hands, or some parts of the body are not shown in the video (see Figure 1 for examples).

The idea of combining facial expressions and body gestures for emotion recognition has recently been explored in the literature [38, 2, 16]. Shan *et al.* [38] extract the local space-time features from two separate video feeds, one containing only facial expressions and another containing upper body movements, recorded simultaneously; the two streams of features are then fused using CCA [12]. Similarly, Joshi *et al.* [16] extract two types of features; facial landmark points from the face region and the local space-time features

[1] Haggard and Issacs [11] are formally acknowledged for the first discovery of the “micromomentary” expressions. Ekman [8] is acknowledged for coining the term “microexpression” and for using it in deception analysis.

from the upper body region, which are then used separately to construct bag-of-features (BoF) representations of face and body, respectively. They report that the BoF obtained from upper body outperforms the one obtained from face only, which let them to suggest that combining facial and body expressions improves performance.

Our work is different in that, instead of extracting features from body and face separately [38], we extract features from the entire image volume, and use dictionary learning and sparse coding techniques to automatically learn the most salient patterns of the image. This lessens the burden of needing a special treatment to combine face and body features, e.g., feature normalization. We also evaluate our approach on a more challenging task: continuous emotion recognition, that is, a regression task, while previous work has concentrated on classification, e.g., [49, 51].

Capturing microexpressions in video has recently received much attention in computer vision and affective computing [47, 32, 39], and a few microexpression databases have been made available [32, 21, 48]. Wu *et al.* [47] applied the Gabor filter and GentleSVM to classify microexpressions, while Pfister *et al.* [32] proposed a method to temporally interpolate facial features using graph embedding. Shreve *et al.* [39] used facial strain patterns to detect both macro- and microexpressions. These approaches, however, have focused only on face regions and have not addressed the sparsity of microexpressions. Our work captures microexpressions over the face and body regions, addressing the sparsity issue directly.

3. OUR APPROACH

Our approach is composed of three steps (see Figure 2). First, we extract the local space-time features over the entire image volume (Section 3.1). We then obtain a sparse representation of the features by using a codebook learned directly from the data (Section 3.2). Finally, we perform max pooling of the sparse codes over the spatio-temporal domain, obtaining a compact representation of facial and body expressions (Section 3.3). This representation is then used as an input to a prediction model to recognize emotion.

3.1 Local Space-Time Features

One desirable property for an emotion recognition system is robustness to the challenges encountered in natural human communication scenarios, such as, the face occlusion. We therefore focus on using appearance-based low-level image descriptors, rather than attempting to extract mid/high-level descriptors of facial expressions (e.g., action units and eye gaze) or body gestures (e.g., joint positions and key poses). This has an additional benefit of not having to assign meanings to facial expressions and body gestures a priori, as many of them are context dependent and culturally specific.

The local space-time feature [18, 19, 7] has recently become a popular motion descriptor for action recognition [33]. It captures salient visual patterns in a space-time image volume by extending the local image descriptor (e.g., [23, 6]) to the space-time domain. One of our key ideas is that local space-time features extracted over the face and body region can be a good visual descriptor for facial and body microexpressions, for its ability to capture micro-temporal image patterns in a compact form.

Obtaining local space-time features is a two-step process – spatio-temporal interest point (STIP) detection followed by feature extraction – and several variants in each step have been developed in the past decade. Wang *et al.* [45] reports that using the Harris3D interest point detector [18] followed by a combination of the Histograms of Oriented Gradient (HOG) and the Histograms of Optical Flow (HOF) feature descriptors [19] provides overall good performance in various action recognition datasets. In this work, therefore, we use the Harris3D detector with HOG/HOF feature descriptors to extract local space-time features.

The Harris3D detector [18] constructs a linear scale-space representation of an image sequence $I(\cdot)$ as

$$L(\cdot;\sigma^2,\tau^2) = g(\cdot;\sigma^2,\tau^2) * I(\cdot) \tag{1}$$

where g is a Gaussian convolution kernel, and σ^2 and τ^2 are spatial and temporal variances, respectively. It then detects STIPs by computing the positive local maxima of a function $H = \det(\mu) - k\,\text{trace}^3(\mu)$, where μ is a spatio-temporal second-moment matrix, $\mu(\cdot;s\sigma^2,s\tau^2) = g(\cdot;s\sigma^2,s\tau^2) * (\nabla L(\nabla L)^T)$.

Figure 1 shows STIPs detected over the face and body region (yellow circles). Notice how well the interest point detector is able to capture subtle face and body movements performed at a micro-temporal scale. For example, Figure 1 (a) shows the STIPs capturing the 'talking while looking away' facial expression along with the 'resting chin with hand' body gesture, Figure 1 (e) capturing the 'looking away' facial expression along with the 'scratching face' gesture, and Figure 1 (d) capturing the 'smiling/laughing' facial expression along with the 'open two palms up' gesture. By extracting the HOG/HOF features around these STIPs, we obtain a feature descriptor that captures both the face and body microexpressions. Figure 1 (c) captures the 'covering face with hands' gesture, showing that our approach does not have to deal with the problems with occlusion (face or body) that happens often in a naturalistic setting.

3.2 Dictionary Learning and Sparse Coding

The local space-time features extracted over the face and body region are quite dense due to the nature of microexpressions, i.e., apart from meaningful microexpressions, there are many other subtle movements of the face and body that do not convey semantically meaningful information. We therefore need a method to represent them in a sparse manner so that only a few most salient patterns are recovered. This will allow our representation to focus on the patterns that appear most frequently in the given data (thus being more discriminative). To this end, we learn a codebook of microexpressions and use it to encode the local features in a sparse manner.

At a high level, the goal of sparse coding is to obtain a sparse representation of an input signal using an over-complete codebook (i.e., the number of codebook entries exceeds the dimension of input signal) such that only a small number of codebook entries are used to represent the input signal (hence the term "sparse"). Given an input signal $\mathbf{x} \in \mathbb{R}^N$ and an over-complete codebook $\mathbf{D} \in \mathbb{R}^{N \times K}, K \gg N$, this goal is formulated as finding a sparse signal $\alpha \in \mathbb{R}^K$ (i.e., most of its elements are zero) that minimizes the reconstruction error,

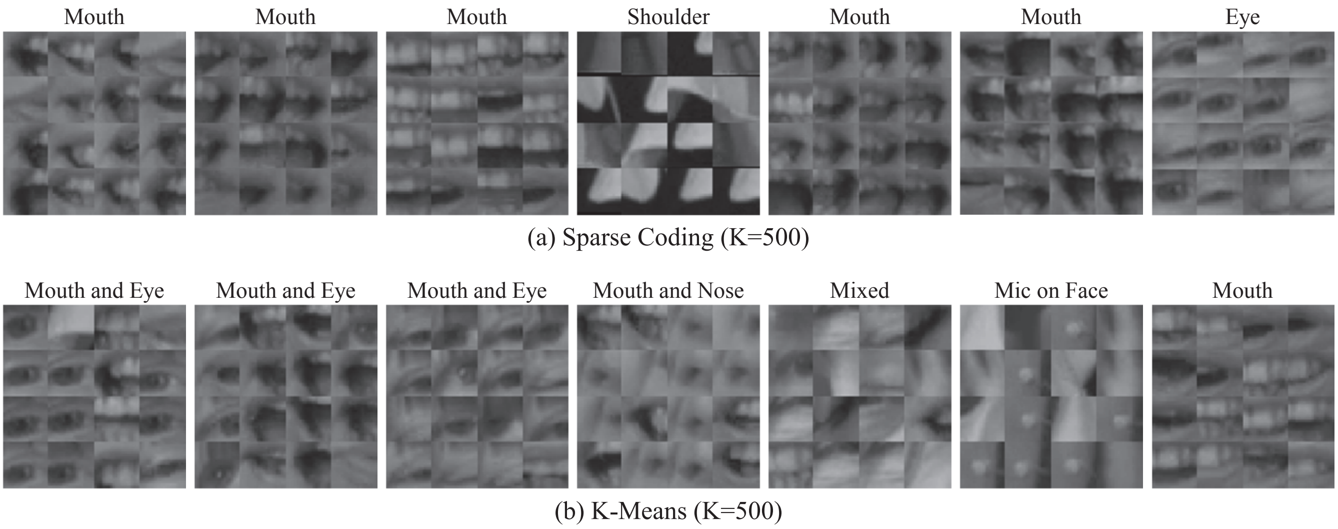
$$\min_{\alpha \in \mathbb{R}^K} \frac{1}{2} \left\|\mathbf{x} - \mathbf{D}\alpha\right\|_2^2 + \lambda \left\|\alpha\right\|_1 \tag{2}$$

where the first term measures reconstruction error, the second term is the L_1 regularization that encourages the sparsity of vector α, and λ controls the relative importance of the two terms. In words, α contains few non-zero linear coefficients to the codebook $\mathbf{D}$ that leads to the best approximation of $\mathbf{x}$.

The codebook $\mathbf{D}$ can be either manually defined or automatically learned from the data. In this work we automatically learn our codebook using a set of local space-time features $\{\mathbf{x}_1, \cdots, \mathbf{x}_M\}$:

$$\min_{\mathbf{D}} \frac{1}{M} \sum_{i=1}^{M} \min_{\alpha_i} \frac{1}{2} \left\|\mathbf{x}_i - \mathbf{D}\alpha_i\right\|_2^2 + \lambda \left\|\alpha_i\right\|_1 \tag{3}$$

This optimization problem is convex in $\mathbf{D}$ with $\mathbf{A} = [\alpha_1, \cdots, \alpha_M]$ fixed, and in $\mathbf{A}$ with $\mathbf{D}$ fixed, but not in both at the same time [24].



(a) Sparse Coding (K=500)

(b) K-Means (K=500)

Figure 3: Clustering of image patches based on their labels obtained using (a) sparse coding and (b) k-means algorithm. To obtain labels using sparse coding we find the dimension in a sparse code α that has the maximum value; for the k-means it is the index of the closest centroid. We show the seven largest clusters (i.e., most frequently occurring labels) found from the same video input. For the purpose of visualization, we cropped 20 × 20 pixel patches at the spatio-temporal center of each STIP. We can see that sparse coding provides more homogeneous (visually similar) clustering of image patches, implying that it encodes the most salient property of the local image features more precisely than the k-means algorithm.

It is thus solved by alternating the two convex optimization problems. Once the codebook $\mathbf{D}$ is learned off-line, we can use it to encode each local feature $\mathbf{x}$ into α by solving Equation 2.

Figure 3 compares the quality of image patch clusters obtained using sparse coding and the k-means algorithm (the caption explains how they were created). It shows that both the sparse coding and the k-means finds the most frequent micro-temporal patterns in a given video, e.g., talking and eye blinking. However, it shows that sparse coding tends to give more homogeneous (visually similar) clusters of image patches. For example, five out of the top seven largest clusters obtained using sparse coding represent talking (mouth); the fourth represents sharp edge movements over the shoulder and the seventh represents eye blinks. These clusters show higher homogeneity than those obtained using the k-means algorithm, whose top three clusters contain both mouth and eye movements, and whose fourth contains mouth and nose movements, etc. This visually demonstrates the ability of sparse coding to encode the most salient property of the local image features more precisely than the k-means algorithm, and is thus more discriminative.

3.3 Spatio-Temporal Feature Pooling

From each frame we obtain different numbers of local space-time features (and corresponding sparse codes). These features are typically pooled to obtain a vector of a fixed dimension, suitable as an input to many machine learning algorithms. The bag-of-features (BoF) approach [19] does this by labeling each feature using the k-means algorithm (as an index of the closest centroid) and generating a histogram of label occurrences within a space-time volume of either the entire sequence or subsequences of video input.

Using the sparse codes α obtained with Equation 2, we can compute a histogram representation using an average pooling operation,

$$\mathbf{z} = \frac{1}{M_v} \sum_{i=1}^{M_v} \alpha_i \tag{4}$$

where M_v is the number of sparse codes associated with a given space-time volume v, and $\mathbf{z} \in \mathbb{R}^K$ is the histogram representation of the sparse codes.

While the histogram representation by an average pooling has shown to perform well, recent progress has shown that an alternative approach, the max pooling operation, provides a better representation that is invariant to image transformations, noise, and clutter [49]. A theoretical analysis given by Boureau *et al.* [1] highlights the superiority of max pooling over average pooling. The max pooling operation is defined as

$$\mathbf{z} = \left[\max_{i=1\ldots M_v} |\alpha_{i,1}|, \cdots, \max_{i=1\ldots M_v} |\alpha_{i,K}| \right] \tag{5}$$

where $|\alpha_{i,k}|$ is an absolute value of the k-th dimension in the i-th sparse code.

In this work, we use max pooling to obtain a per-frame representation of the sparse codes. Specifically, we define a temporal window of predetermined size $\omega = 50$ (one second) centered at each frame, and pool over the sparse codes within the window using max pooling to obtain the per-frame representation $\mathbf{z}_t$. Given a video of length T, therefore, the output is a time-ordered sequence $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_T] \in \mathbb{R}^{K \times T}$. Note that different pooling methods can also be used, such as pooling from a pyramid of multiple spatio-temporal scales similar to [20], which has been shown to help capture surrounding context; for a deep discussion of various feature encoding and pooling methods, readers are referred to [3, 14].

3.4 Prediction

Once the features $\mathbf{Z}$ are max-pooled from each video, we can train a prediction model for either classification or regression. We form a training dataset $\mathcal{D} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{z}_i \in \mathbb{R}^K$ is the input feature vector and $y \in \mathbb{R}$ is a real-valued label, and use Support Vector Regression (SVR) [40] as our prediction model. We use the RBF kernel $K(\mathbf{z}_i, \mathbf{z}_j) = \exp(-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2 / 2\gamma^2)$ with the kernel width γ, which ensures that the data is separable from the origin in feature space.

4. EXPERIMENT

We evaluated our approach on audio-visual emotion recognition. We first evaluated our proposed approach that is purely vision-based, then combined our approach with audio features to evaluate it in a multimodal setting. In particular, we compare three popular modality fusion mechanisms: early fusion, early fusion using kernel CCA, and late fusion with voting.

4.1 Dataset

We used a benchmark dataset provided by the second International Audio-Visual Emotion Challenge (AVEC 2012) [37], which is a subset of the SEMAINE corpus [25]. The goal of this challenge was to estimate continuous values of emotional states in four affective dimensions: arousal (how dynamic or lethargic the person appears to feel), expectation (how surprised the person appears to be), power (how much control the person appears to have over him/herself and the surrounding), and valence (how positive or negative the person appears to feel). There were two sub-challenges in AVEC 2012; we selected the fully continuous sub-challenge (FCSC), where the goal was to predict for each frame in a video all four dimensions of emotional states.

The dataset contains 95 sessions of recordings (31 for training, 32 for development, 32 for test); each session lasts on average 4.75 minutes. Video was recorded at 49.979 FPS with a spatial resolution of 780 × 580 pixels; audio was recorded at 48 kHz. The dataset provides precomputed features for audio and visual modalities. For our experiments on audio-visual emotion recognition we use the provided audio features, which include various low-level descriptors commonly used in audio signal analysis. The audio features have 1,841 dimensions; we used PCA to reduce this to 500, which accounts for the 95% variance of the data. For the task of FCSC, audio features are computed only during speech, using a two-second sliding window at half-second intervals; when there was no vocalization the feature values were set to zero.

4.2 Methodology

We extract the local space-time features at the original frame rate of video (49.979 FPS), using the Harris3D detector and the HOG/HOF descriptor. Specifically, we use the software provided by Laptev *et al.* [19], which detects STIPs at multiple spatio-temporal scales, i.e., spatial scales $\sigma^2 = 2^k, k = [2, \cdots, 9]$ and temporal scales τ^2=[2, 4] (see Equation 1).

For sparse coding we learn a codebook $\mathbf{D}$ of size K=[100, 200, 500, 1000] using 360,000 randomly selected HOG/HOF features extracted from the training video data (about one-tenth of the original size). We fix the L_1 regularization weight $\lambda = 0.1$ (see Equation 2 and Equation 3) throughout our experiment, determined based on our preliminary experiment with the dataset. We use the software provided by Mairal *et al.* [24] to learn our codebook and obtain sparse codes.

As noted in [37], running an experiment with the original frame rate may cause difficulty due to the heavy memory consumption (there are 3.6 million sparse codes in the training split alone). For the max pooling, therefore, we pool the features at half-second intervals using a temporal window of one second. We chose the window size and sampling rate both to reduce the memory requirement and to match the sampling rate of the audio features.

In addition to evaluating our method on visual data alone, we also used the audio features provided from AVEC 2012 to see how our method performed in a multimodal setting. Specifically, we compared three popular fusion methods: early fusion, early fusion using kernel CCA [12], and late fusion with voting.

For the early fusion method, we concatenated the audio-visual features for each frame and used it as an input feature. For early fusion with kernel CCA, we followed the approach explained in Song *et al.* [42], but used the SVR as a prediction model, to keep our experimental protocol consistent. For late fusion with voting, we trained the SVR on audio and visual features separately, and took a weighted average of the prediction results for each frame t:

$$y_t = (1 - \psi) y_t^A + \psi y_t^V \tag{6}$$

where y_t^A and y_t^V are the prediction results from audio and visual modalities, respectively, and $0 \leq \psi \leq 1$ is a weight parameter.

Note that the provided audio features have missing values when there is no vocalization from the subject – in this case we set $\psi = 1$. We varied the weight parameter ψ from 0 to 1, increasing by 0.1.

We train the SVR using an RBF kernel with the kernel width of $\gamma = 10^k, k = [-2, \cdots, -6]$; we set the cost term $C = 1$ throughout the experiment. Once we get the per-frame prediction result $y_t, t = 1, \cdots, T$, we upsample it to match the sampling rate of the ground truth labels. We then perform exponential smoothing with a smoothing factor $\eta = 10^k, k = [-1, -2, -3]$ to ensure the smoothness of the prediction results over time,

$$y'_t = \begin{cases} \eta y_t + (1 - \eta) y'_{t-1} & \text{if } t > 1 \\ y_1 & \text{otherwise.} \end{cases} \tag{7}$$

We find the optimal values of the hyper-parameters – codebook size K, RBF kernel width γ, the exponential smoothing factor η, and the late fusion weight factor ψ – based on the best performance on the development split.

4.3 Results on Visual Input Alone

Table 1 shows the cross-correlation coefficients between predicted and ground-truth labels, averaged over all sequences. We include the baseline result from [37] and the results of the top four contenders from AVEC 2012 [27, 41, 36, 30]. We also include our results on audio-visual input (bottom row), discussed in Section 4.4.

Our approach (video only) gives a dramatic improvement on all four dimensions over the baseline (e.g., on the test split, on average from 0.093 to 0.413) and performs comparably to the previous state-of-the-art results. Note that the results from the top four contenders are based on audio-visual input features, with additional information such as context [41] (i.e., knowing the personality of a virtual human that the user is interacting with) or the duration of conversation [30] (i.e., knowing how long the user has been interacting with a virtual human). Considering our approach is purely vision-based, this is a significant improvement over other approaches. An interesting future work could combine our vision-based approach with those additional features.

Notably, our approach achieved the best published result on the expectation dimension: based solely on visual features we achieve 0.419 (on the test split), while the previous state-of-the-art is 0.330. Among the four dimensions, our approach performed particularly well on the arousal dimension. This may be attributed to how humans might perceive arousal, i.e., how dynamic or lethargic the person appears to be. Intuitively, a person may look dynamic if he/she makes frequent facial and body microexpressions. Our approach, which captures the most salient spatio-temporal patterns of microexpressions, is thus likely to be particularly well suited to predicting the arousal dimension.

Table 1: Experimental results on the development and test splits. We show the cross-correlation coefficients between predicted and ground-truth labels, averaged over all sequences. The result on audio-visual input is obtained using late-fusion approach.

Method	Development Split					Test Split				
	Arousal	Expectation	Power	Valence	Mean	Arousal	Expectation	Power	Valence	Mean
Baseline [37] (Video Only)	0.151	0.122	0.031	0.207	0.128	0.077	0.128	0.030	0.134	0.093
Baseline [37] (Audio Visual)	0.181	0.148	0.084	0.215	0.157	0.141	0.101	0.072	0.136	0.112
Nicolle *et al.* [27] (Audio Visual)	**0.644**	0.341	0.511	0.350	0.461	**0.612**	0.314	0.556	0.341	**0.456**
Soladie *et al.* [41] (Audio Visual)	0.520	0.300	**0.590**	0.470	**0.470**	0.420	0.330	**0.570**	**0.420**	0.430
Savran *et al.* [36] (Audio Visual)	0.383	0.266	0.556	**0.473**	0.420	0.359	0.215	0.477	0.325	0.344
Ozkan *et al.* [30] (Audio Visual)	0.396	0.246	0.476	0.235	0.338	0.325	0.311	0.451	0.182	0.317
Our Approach (Video Only)	0.583	0.368	0.491	0.345	0.447	0.575	0.419	0.427	0.230	0.413
Our Approach (Audio Visual)	0.581	**0.380**	0.510	0.334	0.451	0.576	**0.429**	0.427	0.235	0.417

On the other hand, our approach performed only marginally better than the baseline on valence: on the test split we achieve 0.230 while the baseline is 0.134. Soladie *et al.* [41] report that high-level descriptors such as the duration of a smile can be a discriminative feature for valence prediction. Our approach captures an implicit notion of a smile, e.g., when the corner of the mouth moves upward, etc. However, it does not extract how long a person is smiling. This is mainly due to the Harris3D interest point detector we used: we detected STIPs using the temporal scale of τ^2 =[2 4] that corresponds to 11 and 16 frames, which, in the dataset we used, corresponds to 0.2 and 0.3 seconds. Although this short temporal scale allows us to capture microexpressions, it limits the duration each descriptor can handle. Also, because the Harris3D detector finds a moving corner in a space-time image volume, if there is not enough visual difference from frame to frame (e.g., keep smiling without moving), there will be no features to track. We believe that dense sampling methods such as dense trajectory tracker [44] followed by our sparse encoding step may achieve better performance; we leave this as interesting future work.

4.4 Results on Audio-Visual Input

Table 2 compares the cross-correlation coefficients on audio-visual input data obtained with three different fusion methods: early fusion, early fusion with KCCA, and late fusion with voting. The late fusion approach consistently outperformed the two early fusion approaches in terms of the mean score, both on the development (0.451) and the test splits (0.417). The two early fusion methods performed similarly on the test split (0.351 and 0.341, respectively), although on the development split the early fusion (0.425) outperformed the early fusion with KCCA (0.402).

The two early fusion approaches performed no better than the late fusion approach. In fact, the performance was even worse than using our visual feature alone (see Table 1). We believe this is due, in part, to the missing audio features in some frames. Note that the audio features are set to zero when there is no vocalization. Those frames with zero-valued features are likely to have various levels of annotated emotional states, depending on the context and shown facial and body expressions (e.g., the user can look happy or sad without vocalization). Likewise, some frames will have the same level of annotated emotional states with or without vocalization (e.g., the user can laugh quietly or aloud). We can therefore regard the zero-valued frames as “missing” observations. The early fusion approach combines audio and visual modalities in the feature space, thereby creating a combinatorial feature space. We believe that the early fusion, by forcing the features to be combined regardless of the missing observations, may have confused the prediction model learning meaningful mapping between the combinatorial input feature space (audio-visual features) and the output label space (emotional states).

The late fusion approach, on the other hand, combines the prediction results only when the features from both modalities are available – when the audio modality is not available the prediction is made entirely based on the visual modality alone (by setting $\psi = 1$ for that frame). Figure 4 shows the mean correlation coefficient score as a function of the fusion weight. We can see that, except for the expectation dimension, prediction results from the visual modality were preferred. This implies that visual modality (represented by our proposed method) is more indicative of the three affective dimensions (arousal, power, and valence) than the audio modality (represented by the baseline features provided in AVEC 2012). As seen in Table 1, this late fusion approach improved performance in all four dimensions (see the results on the test split). We believe that more careful design of the audio features will improve the performance further; we leave this as interesting future work.

5. CONCLUSIONS

We presented a compact, yet discriminative representation of facial and body microexpressions. Our approach extracts local space-time features from the face and body region and learns a codebook from them. We then encode the local space-time features using sparse coding, selecting only a few most salient spatio-temporal patterns of microexpressions. This allows our representation to focus on the microexpressions that appear most often in the given data, which will likely contain semantically meaningful informa-

Table 2: Experimental results on different fusion methods. We show the cross-correlation coefficients between predicted and ground-truth labels, averaged over all sequences.

Fusion Method	Development Split					Test Split				
	Arousal	Expectation	Power	Valence	Mean	Arousal	Expectation	Power	Valence	Mean
Early Fusion	0.494	0.366	0.487	**0.351**	0.425	0.409	0.401	0.406	0.186	0.351
Early Fusion with KCCA	0.510	0.372	0.466	0.259	0.402	0.394	0.404	0.389	0.175	0.341
Late Fusion with Voting	**0.581**	**0.380**	**0.510**	0.334	**0.451**	**0.576**	**0.429**	**0.427**	**0.235**	**0.417**

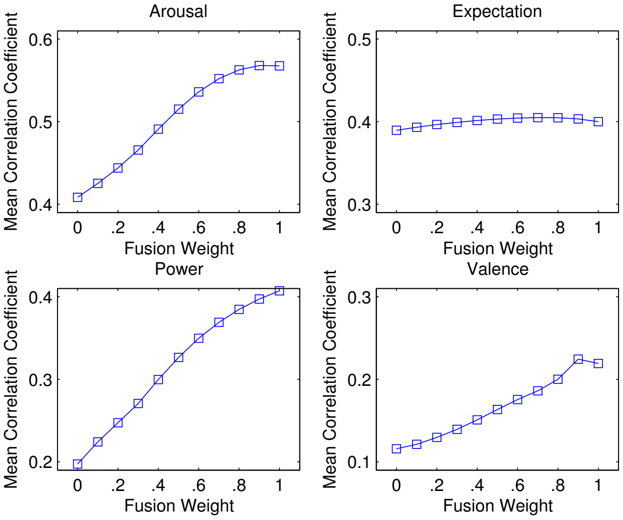


Figure 4: Mean correlation coefficient scores as a function of the fusion weight ψ. The higher the fusion weight, the more the visual modality is weighted in the final prediction. Results are based on the development split.

tion. Max pooling is performed to obtain a compact representation of the features, and an SVR is used to predict continuous values of four affective dimensions. Experimental results on the AVEC 2012 dataset show that our approach achieves the best published performance on the expectation dimension even without using audio and context-related features.

This work has focused on obtaining a good representation of visual modality. In the future, we plan to work on an audio-equivalent of what we have proposed here; that is, obtaining a compact yet discriminative representation of speech and prosodic characteristics of humans. This will potentially enable us to perform a fusion of audio-visual modalities in a more meaningful way. We also plan to work on improving the prediction part of our approach by considering temporal dynamics in streaming data [34, 42].

6. ACKNOWLEDGMENTS

This work was supported in part by ONR #N000140910625, by NSF IIS-1018055, by NSF IIS-1118018, and by the U.S. Army Research, Development, and Engineering Command (RDECOM). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of Naval Research, or the U.S. government.

7. REFERENCES

[1] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, 2010.

[2] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image Vision Comput.*, 31(2), 2013.

[3] A. Coates and A. Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *PAMI*, 23(6), 2001.

[5] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[8] P. Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. WW Norton & Company, 2009.

[9] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[10] P. Ekman and W. V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. I S H K, 2003.

[11] E. A. Haggard and K. S. Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*. Springer, 1966.

[12] D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *NECO*, 16(12), 2004.

[13] J. M. Iverson and S. Goldin-Meadow. Why people gesture when they speak. *Nature*, 396(6708), 1998.

[14] Y. Jia, O. Vinyals, and T. Darrell. On compact codes for spatially pooled features. In *ICML*, 2013.

[15] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *FG*, 2011.

[16] J. Joshi, R. Goecke, M. Breakspear, and G. Parker. Can body expressions contribute to automatic depression analysis? In *FG*, 2013.

[17] A. Kendon. *Gesture – Visible Action as Utterance*. Cambridge University Press, 2004.

[18] I. Laptev. On space-time interest points. *IJCV*, 64(2-3), 2005.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[21] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *FG*, 2013.

[22] M. Liu, S. Li, S. Shan, and X. Chen. AU-aware Deep Networks for Facial Expression Recognition. In *FG*, 2013.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004.

[24] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *JMLR*, 11, 2010.

[25] G. McKeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schröder. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *T. Affective Computing*, 3(1), 2012.

[26] D. McNeill. *Gesture and thought*. University of Chicago Press, 2008.

[27] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *ICMI*, 2012.

[28] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7), 2002.

[29] B. A. Olshausen and D. J. Fieldt. Sparse coding with an overcomplete basis set: a strategy employed by V1. *Vision Research*, 37, 1997.

[30] D. Ozkan, S. Scherer, and L.-P. Morency. Step-wise emotion recognition using concatenated-hmm. In *ICMI*, 2012.

[31] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *PAMI*, 22(12), 2000.

[32] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen. Recognising spontaneous facial micro-expressions. In *ICCV*, 2011.

[33] R. Poppe. A survey on vision-based human action recognition. *IVC*, 28(6), 2010.

[34] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li. Global ranking using continuous conditional random fields. In *NIPS*, 2008.

[35] E. Rahtu, J. Heikkilä, V. Ojansivu, and T. Ahonen. Local phase quantization for blur-insensitive image analysis. *Image Vision Comput.*, 30(8), 2012.

[36] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *ICMI*, 2012.

[37] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *ICMI*, 2012.

[38] C. Shan, S. Gong, and P. W. McOwan. Beyond facial expressions: Learning human emotion from body gestures. In *BMVC*, 2007.

[39] M. Shreve, S. Godavarthy, D. B. Goldgof, and S. Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *FG*, 2011.

[40] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. Technical report, Statistics and computing, 2003.

[41] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In *ICMI*, 2012.

[42] Y. Song, L.-P. Morency, and R. Davis. Multimodal human behavior analysis: learning correlation and interaction across modalities. In *ICMI*, 2012.

[43] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *T. Affective Computing*, 3(1), 2012.

[44] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[45] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[46] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6), 2010.

[47] Q. Wu, X. Shen, and X. Fu. The machine knows what you are hiding: An automatic micro-expression recognition system. In *ACII*, 2011.

[48] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu. Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *FG*, 2013.

[49] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[50] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *PAMI*, 31(1), 2009.

[51] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu. Sparse coding on local spatial-temporal volumes for human action recognition. In *ACCV*, 2010.