Unsupervised Discovery of Facial Events

Feng Zhou† Fernando De la Torre† Jeffrey F. Cohn‡
†, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.
‡, University of Pittsburgh, Department of Psychology, Pittsburgh, Pennsylvania 15260.

Abstract

*Automatic facial image analysis has been a long standing research problem in computer vision. A key component in facial image analysis, largely conditioning the success of subsequent algorithms (*e.g.*, facial expression recognition), is to define a vocabulary of possible dynamic facial events. To date, that vocabulary has come from the anatomically-based Facial Action Coding System (FACS) or more subjective approaches (i.e. emotion-specified expressions). The aim of this paper is to discover facial events directly from video of naturally occurring facial behavior, without recourse to FACS or other labeling schemes. To discover facial events, we propose a novel temporal clustering algorithm, Aligned Cluster Analysis (ACA), and a multi-subject correspondence algorithm for matching expressions. We use a variety of video sources: posed facial behavior (Cohn-Kanade database), unscripted facial behavior (RU-FACS database) and some video in infants. ACA achieved moderate intersystem agreement with manual FACS coding and proved informative as a visualization/summarization tool.*

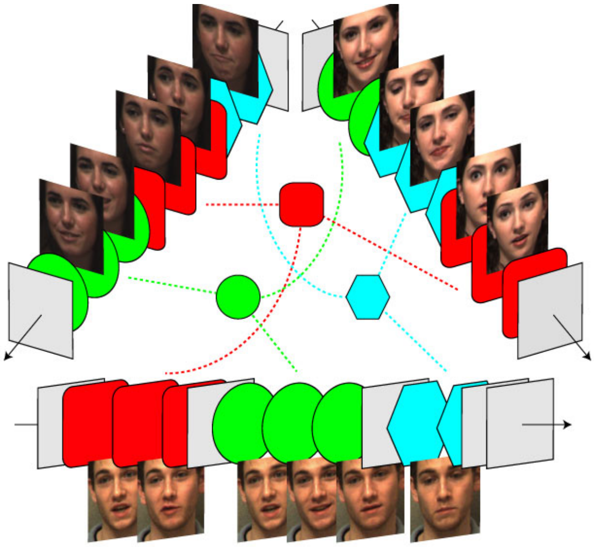


Figure 1. Selected video frames of unposed facial behavior from three participants. Different colors and shapes represent dynamic events discovered by unsupervised learning: smile (green circle) and lip compressor (blue hexagons). Dashed lines indicate correspondences between persons.

1. Introduction

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

Faces possess their own language. To represent the elemental units of this language, Ekman and Friesen [11] in the 70's proposed the Facial Action Coding System (FACS). FACS segments the visible effects of facial muscle activation into "action units". Each action unit is related to one or more facial muscles. The FACS taxonomy was develop by manually observing graylevel variation between expressions in images and to a lesser extent by recording the electrical activity of underlying facial muscles [3]. Because of its descriptive power, FACS has become the state of the art in manual measurement of facial expression and is widely used in studies of spontaneous facial behavior. In part for these reasons, much effort in automatic facial image analysis seeks to automatically recognize FACS action units [1, 27, 23, 26].

In this paper, we ask whether unsupervised learning can discover useful facial units in video sequences of one or more persons, and whether the discovered facial events correspond to manual coding of FACS action units. We propose extensions of an unsupervised temporal clustering algorithm, Aligned Cluster Analysis (ACA). ACA is an extension of kernel k-means to cluster multi-dimensional time series. Using this unsupervised learning approach it is possible to find meaningful dynamic clusters of similar facial expressions in one individual and correspondences between facial events across individuals in an unsupervised manner.

Fig. (1) illustrates the main idea of the paper. In addition, we show how our algorithms for temporal clustering of facial events can be used for summarization and visualization.

2. Temporal segmentation and clustering of human behavior

This section reviews previous work on temporal clustering and segmentation of facial and human behavior.

With few exceptions, previous work on facial expression or action unit recognition has been supervised in nature (i.e. event categories are defined in advance in labeled training data, see [1, 27, 23, 26] for a review of state-of-the-art algorithms). Little attention has been paid to the problem of unsupervised temporal segmentation or clustering prior to recognition. Essa and Pentland [12] proposed FACS+ a probabilistic flow-based method to describe facial expressions. Hoey [15] presented a multilevel Bayesian network to learn in a weakly supervised manner the dynamics of facial expression. Bettinger *et al.* [2] used AAM to learn the dynamics of person-specific facial expression models. Irani and Zelnik [33] proposed a modification of structure-from-motion factorization to temporally segment rigid and non-rigid facial motion. De la Torre *et al.* [8] proposed a geometric-invariant clustering algorithm to decompose a stream of one person's facial behavior into facial gestures. Their approach suggested that unusual facial expressions might be detected through temporal outlier patterns. In summary, previous work in facial expression addresses temporal segmentation of facial expression in a single person. The current work extends previous approaches to unsupervised temporal clustering across individuals.

Outside of the facial expression literature, unsupervised temporal segmentation and clustering of human and animal behavior has been addressed by several groups. Zelnik-Manor and Irani [32] extracted spatio-temporal features at multiple temporal scales to isolate and cluster events. Guerra-Filho and Aloimonos [13] presented a linguistic framework to learn human activity representations. The low level representation of their framework, motion primitives, referred to as kinetemes, were proposed as the foundation for a kinetic language. Yin *et al.* [30] proposed a discriminative feature selection method to discover a set of temporal segments, or units, in American Sign Language. These units could be distinguished with sufficient reliability to improve accuracy in ASL recognition. Wang *et al.* [29] used deformable template matching of shape and context in static images to discover action classes. Turaga *et al.* [28] presented a cascade of dynamical systems to cluster a video sequence into activities. Niebles *et al.* [21] proposed an unsupervised method to learn human action categories. They represented video as a bag-of-words model of space-time interest points. Latent topic models were used to learn their probability distribution, and intermediate topics corresponded to human action categories. Oh *et al.* [22] proposed parametric segmental switching dynamical models to segment honeybees behavior. Related work in temporal segmentation has been done, as well, in the area of data mining [17] and change point detection [14]. Unlike previous approaches, we propose the use of ACA. ACA generalizes kernel k-means to cluster time series, providing a simple yet effective and robust method to cluster multi-dimensional time series with few parameters to tune.

3. Facial feature tracking and image features

Over the last decade, appearance models [4, 19] have become increasingly prominent in computer vision. In the work below, we use AAMs [19] to detect and track facial features, and extract features. Fig. (2a) shows an example of AAM using image data from RU-FACS [1].

Sixty-six facial features and the related face texture are tracked throughout an image sequence. To register images to a canonical view and face, a normalization step registers each image with respect to an average face. After the normalization step, we build shape and appearance features for the upper and lower face regions. Shape features include, $\mathbf{x}_1^U$ the distance between inner brow and eye, $\mathbf{x}_2^U$ the distance between outer brow and eye, $\mathbf{x}_3^U$ the height of eye, $\mathbf{x}_1^L$ the height of lip, $\mathbf{x}_2^L$ the height of teeth, and $\mathbf{x}_3^L$ the angle of mouth corners. Appearance features are composed of SIFT descriptors computed at points around the outer outline of the mouth (at 11 locations) and on the eyebrows (5 points). The dimensionality of the resulting feature vector is reduced using PCA to retain 95% of the energy, yielding appearance features for the upper ($\mathbf{x}_4^U$) and lower ($\mathbf{x}_4^L$) face. For the task of clustering emotions, features from both face parts were used to obtain a holistic representation of the face. For more precise facial action segmentation, each face part was considered individually. See Fig. (2b) for an illustration of the feature extraction process.

4. Aligned Cluster Analysis (ACA)

This section describes Aligned Cluster Analysis (ACA), an extension of kernel k-means to cluster time series. ACA combines kernel k-means with Dynamic Time Alignment Kernel (DTAK). A preliminary version of ACA was presented at [36].

4.1. Dynamic time alignment kernel (DTAK)

To align time series, a frequent approach is Dynamic Time Warping (DTW). A known drawback of using DTW as a distance is that it fails to satisfy the triangle inequality. To address this issue, Shimodaira *et al.* [25] proposed Dynamic Time Alignment Kernel (DTAK). The DTAK between two sequences, $\mathbf{X} \doteq [\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}] \in \mathbb{R}^{d \times n_x}$ (see

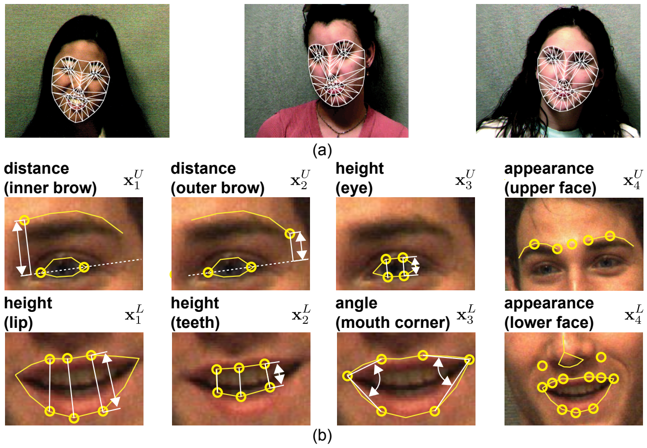

(a)
distance (inner brow) $\mathbf{x}_1^U$
distance (outer brow) $\mathbf{x}_2^U$
height (eye) $\mathbf{x}_3^U$
appearance (upper face) $\mathbf{x}_4^U$
height (lip) $\mathbf{x}_1^L$
height (teeth) $\mathbf{x}_2^L$
angle (mouth corner) $\mathbf{x}_3^L$
appearance (lower face) $\mathbf{x}_4^L$
(b)

Figure 2. Facial features used for temporal clustering. (a) AAM fitting across different subjects. (b) Eight different features extracted from distance between tracked points, height of facial parts, angles for mouth corners, and appearance patches.

notation [1]) and $\mathbf{Y} \doteq [\mathbf{y}_1, \cdots, \mathbf{y}_{n_y}] \in \mathbb{R}^{d \times n_y}$, is defined as:

$$\tau = \max_{\mathbf{Q}} \sum_{c=1}^{l} \frac{1}{n_x + n_y}(q_{1c} - q_{1c-1} + q_{2c} - q_{2c-1})\kappa_{q_{1c}q_{2c}},$$

where $\kappa_{ij} = \phi(\mathbf{x}_i)^T \phi(\mathbf{y}_j)$ represents the kernel similarity between frame $\mathbf{x}_i$ and $\mathbf{y}_j$. $\mathbf{Q} \in \mathbb{R}^{2 \times l}$ is an integer matrix that contains indexes to the alignment path between $\mathbf{X}$ and $\mathbf{Y}$. For instance, if the c^{th} column of $\mathbf{Q}$ is $[q_{1c}\ q_{2c}]^T$, the q_{1c} frame in $\mathbf{X}$ corresponds to the q_{2c} frame in $\mathbf{Y}$. l is the number of steps needed to align both signals.

DTAK finds the path that maximizes the weighted sum of the similarity between sequences. A more revealing mathematical expression can be achieved by considering a new normalized correspondence matrix $\mathbf{W} \in \mathbb{R}^{n_x \times n_y}$, where $w_{ij} = \frac{1}{n_x+n_y}(q_{1c} - q_{1c-1} + q_{2c} - q_{2c-1})$ if there exist $q_{1c} = i$ and $q_{2c} = j$ for some c, otherwise $w_{ij} = 0$. Then DTAK can be rewritten:

$$\tau(\mathbf{X}, \mathbf{Y}) = tr(\mathbf{K}^T \mathbf{W}) = \psi(\mathbf{X})^T \psi(\mathbf{Y}), \qquad (1)$$

where $\psi(\cdot)$ denotes a mapping of the sequence into a feature space, and $\mathbf{K} \in \mathbb{R}^{n_x \times n_y}$. More details in [35].

4.2. k-means and kernel k-means

Clustering refers to the partition of n data points into k disjoint clusters. Among various approaches to unsupervised clustering, k-means [18, 34] and kernel k-means (KKM) [10, 31] are among the simplest and most popular. k-means and KKM clustering split a set of n objects into c groups by minimizing the within cluster variation. KKM finds the partition of the data that is a local optimum of the following energy function [34, 7]:

$$J_{kkm}(\mathbf{M}, \mathbf{G}) = ||\psi(\mathbf{X}) - \mathbf{MG}||_F^2, \qquad (2)$$

where $\mathbf{X} \in \mathbb{R}^{d \times n}$, $\mathbf{G} \in \mathbb{R}^{k \times n}$ and $\mathbf{M} \in \mathbb{R}^{d \times k}$. $\mathbf{G}$ is an indicator matrix, such that $\sum_c g_{ci} = 1$, $g_{ij} \in \{0, 1\}$ and g_{ij} is 1 if $\mathbf{d}_i$ belongs to class c, n denotes the number of samples. The columns of $\mathbf{X}$ contain the original data points, and the columns of $\mathbf{M}$ represent the cluster centroids; d is the dimension of the kernel mapping. In the case of KKM, d can be infinite dimensional and typically $\mathbf{M}$ cannot be computed explicitly. Substituting the optimal $\mathbf{M} = \psi(\mathbf{X})\mathbf{G}^T(\mathbf{GG}^T)^{-1}$ value, eq. (2) results in:

$$J_{kkm}(\mathbf{G}) = tr\Big(\mathbf{LK}\Big) \quad \mathbf{L} = \mathbf{I}_n - \mathbf{G}^T(\mathbf{GG}^T)^{-1}\mathbf{G}. \quad (3)$$

The KKM method typically uses a local search [10] to find a matrix $\mathbf{G}$ that makes $\mathbf{L}$ maximally correlated with the sample kernel matrix $\mathbf{K} = \psi(\mathbf{X})^T \psi(\mathbf{X})$.

4.3. ACA objective function

Given a sequence $\mathbf{X} \doteq [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ with n samples, ACA decomposes $\mathbf{X}$ into m disjointed segments, each of which corresponds to one of k classes. The i^{th} segment, $\mathbf{Z}_i \doteq [\mathbf{x}_{s_i}, \cdots, \mathbf{x}_{s_{i+1}-1}] \doteq \mathbf{X}_{[s_i, s_{i+1})} \in \mathbb{R}^{d \times w_i}$, is composed of samples that begin at position s_i and end at $s_{i+1} - 1$. The length of the segment is constrained as $s_{i+1} - s_i \leq n_{\max}$. $n_{\max}$ is the maximum length of the segment that controls the temporal granularity of the factorization. An indicator matrix $\mathbf{G} \in \{0, 1\}^{k \times m}$ assigns each segment to a class; $g_{ci} = 1$ if $\mathbf{Z}_i$ belongs to class c.

ACA combines kernel k-means with the DTAK to achieve temporal clustering by minimizing:

$$J_{aca}(\mathbf{G}, \mathbf{M}, \mathbf{s}) = ||[\psi(\mathbf{Z}_1) \ \cdots \ \psi(\mathbf{Z}_m)] - \mathbf{MG}||_F^2. \quad (4)$$

The difference between KKM and ACA is the introduction of the variable $\mathbf{s}$ that determines the start and end of each segment $\mathbf{Z}_i(\mathbf{s})$. $\psi(\cdot)$ is a mapping such that, $\tau_{ij} = \psi(\mathbf{Z}_i)^T \psi(\mathbf{Z}_j) = tr(\mathbf{K}_{ij}^T \mathbf{W}_{ij})$ is the DTAK. Observe that there are two kernel matrices, $\mathbf{T} \in \mathbb{R}^{m \times m}$ is the kernel segment matrix and $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the kernel sample matrix (kernel between samples). $\mathbf{T} \in \mathbb{R}^{m \times m}$ can be expressed re-arranging the $m \times m$ blocks of $\mathbf{W}_{ij} \in \mathbb{R}^{w_i \times w_j}$ into a global correspondence matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, that is:

$$\mathbf{T} = [\tau_{ij}]_{m \times m} = [tr(\mathbf{K}_{ij}^T \mathbf{W}_{ij})]_{m \times m} = \mathbf{H}(\mathbf{K} \circ \mathbf{W})\mathbf{H}^T,$$

where $\mathbf{H} \in \{0, 1\}^{m \times n}$ is the sample-segment indicator matrix; $h_{ij} = 1$ if j^{th} sample belong to i^{th} segment. Unfortunately, DTAK is not a strictly positive definite kernel

[1] Bold capital letters denote a matrix $\mathbf{X}$, bold lower-case letters a column vector $\mathbf{x}$, and all non-bold letters denote scalar variables. $\mathbf{x}_j$ represents the j^{th} column of the matrix $\mathbf{X}$. x_{ij} denotes the scalar in the row i and column j of the matrix $\mathbf{X}$. $\mathbf{I}_k \in \mathbb{R}^{k \times k}$ denotes the identity matrix. $||\mathbf{x}||_2^2$ denotes the norm of the vector $\mathbf{x}$. $tr(\mathbf{X}) = \sum_i x_{ii}$ is the trace of the matrix $\mathbf{X}$. $||\mathbf{X}||_F^2 = tr(\mathbf{X}^T\mathbf{X}) = tr(\mathbf{XX}^T)$ designates the Frobenius norm of a matrix. $\circ$ denotes the Hadamard or point-wise product.

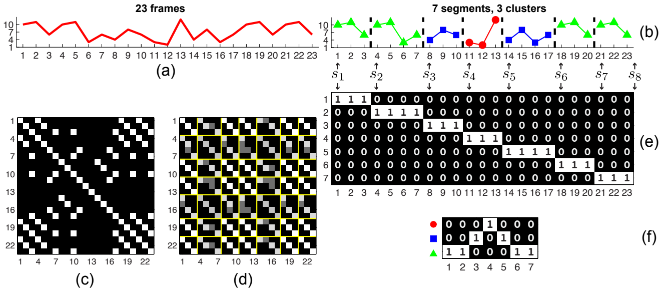


Figure 3. Example of temporal clustering. (a) 1-D sequence. (b) Results of temporal clustering. (c) Self-similarity matrix ($\mathbf{K}$). (d) Correspondence matrix ($\mathbf{W}$). (e) Frame-segment indicator matrix ($\mathbf{H}$). (f) Segment-class indicator matrix ($\mathbf{G}$).

[6]. Thus, we add a scaled identity matrix to $\mathbf{K}$; that is, $\mathbf{K} \leftarrow \mathbf{K} + \sigma \mathbf{I}_n$, were σ is chosen to be the absolute value of the smallest eigenvalue of $\mathbf{T}$ if it has negative eigenvalues.

After substituting the optimal value of $\mathbf{M}$ in eq. (4), a more enlightened form of J_{aca} can be derived by re-arranging the $m \times m$ blocks of $\mathbf{W}_{ij} \in \mathbb{R}^{w_i \times w_j}$ into a global correspondence matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$,

$$J_{aca}(\mathbf{G}, \mathbf{s}) = tr\Big((\mathbf{L} \circ \mathbf{W})\mathbf{K}\Big), \tag{5}$$

where $\mathbf{L} = \mathbf{I}_n - \mathbf{H}^T\mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{G}\mathbf{H}$. Recall $\mathbf{H}$ depends on $\mathbf{s}$. Fig. (3) illustrates the matrices $\mathbf{K}$, $\mathbf{H}$, $\mathbf{W}$ and $\mathbf{G}$ in a synthetic example of temporal clustering. Consider the special case when, $m = n$ and $\mathbf{H} = \mathbf{I}_n$; that is, each frame is treated as a segment. In this case, DTAK would be a kernel between two frames, *i.e.*, $\mathbf{W} = \mathbf{1}_n \mathbf{1}_n^T$ and ACA is equivalent to kernel k-means, eq. (3).

Optimizing ACA is a non-convex problem. We use a coordinate descent strategy that alternates between optimizing $\mathbf{G}$ and $\mathbf{s}$ while implicitly computing $\mathbf{M}$. Given a sequence $\mathbf{X}$ of length n, the number of possible segmentations is exponential, which typically renders a brute-force search infeasible. We adopt a dynamic programming (DP) based algorithm that has a complexity $O(n^2 n_{max})$ to exhaustively examine all the possible segmentations. See [35] for more details on the optimization and the code.

4.4. Learning good features for ACA

The success of kernel machines largely depends on the choice of the kernel parameters and the functional form of the kernel. As in previous work on multiple kernel learning [9, 5], we consider the frame kernel as a positive combination of multiple kernels, that is: $\mathbf{K}(\mathbf{a}) = \sum_{l=1}^{d} a_l \mathbf{K}_l$, s.t. $\mathbf{a} \geq \mathbf{0}_d$ where the set $\{\mathbf{K}_1, \cdots, \mathbf{K}_d\}$ is given and the a_l's are to be optimized.

In the ideal case [9, 5], if two samples belong to the same class, the kernel function outputs a similarity of 1 and 0 otherwise. In the case of temporal segmentation, the label of the i^{th} frame is given by $\mathbf{G}\mathbf{h}_i$. Assuming that all labels $(\mathbf{G}, \mathbf{H}, \mathbf{W})$ are known, we minimize the distance between the ideal kernel matrix and the parameterized one, that is:

$$J_{learn}(\mathbf{a}) = \|\mathbf{W} \circ \big(\mathbf{F} - \mathbf{K}(\mathbf{a})\big)\|_F^2, \tag{6}$$

where $\mathbf{F} = \mathbf{H}^T\mathbf{G}^T\mathbf{G}\mathbf{H}$, and the correspondence matrix $(\mathbf{W})$ weights individually the pair of frames that have been used in the calculation of DTAK.

To optimize J_{learn} with respect to $\mathbf{a}$, we rewrite eq. (6) as a quadratic programming problem: $J_{learn}(\mathbf{a}) = \mathbf{a}^T\mathbf{Z}\mathbf{a} - 2\mathbf{f}^T\mathbf{a} + c$, where $z_{ij} = tr((\mathbf{W} \circ \mathbf{K}_i)^T(\mathbf{W} \circ \mathbf{K}_j))$, $f_i = tr((\mathbf{W} \circ \mathbf{F})^T(\mathbf{W} \circ \mathbf{K}_i))$ and c is a constant.

5. Experiments

This section reports experimental results for unsupervised temporal segmentation of facial behavior and compares them with emotion and FACS labels in two scenarios: first for individual subjects and then for sets of subjects.

5.1. Data sources

We use a variety of video sources: posed facial behavior from the Cohn-Kanade database [16], unscripted facial behavior from the RU-FACS database [1], and infants observed with their mothers [20]. The databases are:

- **Cohn-Kanade (CK) database**: The database contains a recording of posed facial behavior for 100 adults. With a few exceptions, all are between 18 and 30 years of age. There are small changes in pose and illumination, all expressions are brief (about 20 frames on average), begin at neutral, proceed to a target expression, and are well differentiated relative to unposed facial behavior in a naturalistic context (e.g., RU-FACS). Peak expressions for each sequences are AU- and emotion-labeled. The latter were used in the experiment reported below. The emotion labels were surprise, sadness, anger, fear and joy.

- **RU-FACS database**: The RU-FACS database [1] consists of digitized video and manual FACS of 34 young adults. They were recorded during an interview of approximately 2 minutes duration in which they lied or told the truth in response to an interviewer's questions. Pose orientation was mostly frontal with small to moderate out-of-plane head motion. Image data from five subjects could not be analyzed due to image artifacts. Thus, image data from 29 subjects was used.

- **Infant social behavior**: Image data were from a three-minute face-to-face interaction of a 6-month-old infant with her mother [20]. The infant was seated across from her mother. Mean head orientation was frontal but large changes in head orientation were common.

5.2. Facial event discovery for individual subjects

This section describes two experiments in facial event discovery of one individual. The first experiment compares the clustering results with the ones provided by FACS. The second experiments uses unsupervised temporal clustering to summarize facial behavior in an infant.

5.2.1 Individual subjects in RU-FACS

We randomly selected 10 sets of 19 subjects for learning ACA weights and 10 subjects for testing. We compared the performance of unsupervised ACA, ACA with learned weights (ACA+Learn), and KKM. We used 8 features (4 upper face and 4 lower face) as described in section 3.

For each realization (10 in total), we used 10 random subjects and ran ACA, ACA+Learn, and KKM 10 times for each subject starting from different initializations and selected the solution with least error. Because the number and frequency of action units varied among subjects, and to investigate the stability of the clustering w.r.t. the number of clusters, between 8 $\sim$ 11 clusters were selected for the lower face and 4 $\sim$ 7 for the upper face. The clustering results are the average over all clusters. The length constraint of the facial actions was set to be $n_{\max} = 80$. Accuracy is computed as the percentage of temporal clusters found by ACA that contain the same AU or AU combination using the confusion matrix:

$$\mathbf{C}(c_1, c_2) = \sum_{i=1}^{m_{alg}} \sum_{j=1}^{m_{truth}} g_{c_1 i}^{alg} g_{c_2 j}^{truth} |\mathbf{Z}_i^{alg} \cap \mathbf{Z}_j^{truth}| \quad (7)$$

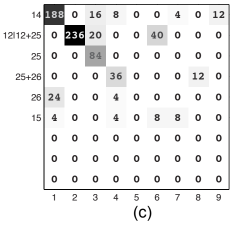
where $\mathbf{Z}_i^{alg}$ is the i^{th} segment returned by ACA (or KKM), and $\mathbf{Z}_j^{truth}$ is the j^{th} segment of the ground-truth data. $C(c_1, c_2)$ represents the total number of frames on the cluster segment c_1 that are shared by the cluster segment c_2 in ground truth. $g_{c_1 i}^{alg}$ is a binary value that indicates whether the i^{th} segment is classified as the c_1 temporal cluster of ACA. $|\mathbf{Z}_i^{alg} \cap \mathbf{Z}_j^{truth}|$ denotes the number of frames that the segment $\mathbf{Z}_i^{alg}$ and $\mathbf{Z}_j^{truth}$ share. The Hungarian algorithm is applied to find the optimum solution for the cluster correspondence problem. Empty rows or columns are inserted if the number of clusters is different from the ground truth, *i.e.*, $k^{alg} \neq k^{truth}$. Due to the possible occurrence of multiple AUs in the same frame, we consider AU combinations as distinct temporal clusters. We consider AUs with a minimum duration of 10 video frames. Any frames for which no AUs occurred were omitted.

(a)

Lower Face		Upper Face	
Feature	Weight	Feature	Weight
$\mathbf{x}_1^L$	.11(.02)	$\mathbf{x}_1^U$	.01(.01)
$\mathbf{x}_2^L$	.12(.07)	$\mathbf{x}_2^U$	.04(.02)
$\mathbf{x}_3^L$	.15(.06)	$\mathbf{x}_3^U$	.07(.04)
$\mathbf{x}_4^L$	.97(.03)	$\mathbf{x}_4^U$	1.47(.04)

(b)

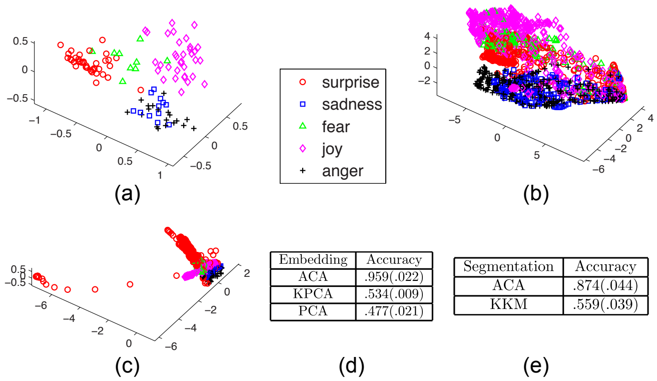
Segmentation	Lower Face	Upper Face
ACA + Learn	.704(.116)	.837(.123)
ACA	.687(.161)	.756(.154)
KKM	.434(.086)	.547(.138)



	1	2	3	4	5	6	7	8	9
14	188	0	16	8	0	0	4	0	12
12|12+25	0	236	20	0	0	40	0	0	0
25	0	0	84	0	0	0	0	0	0
25+26	0	0	0	36	0	0	0	12	0
26	24	0	0	4	0	0	0	0	0
15	4	0	0	4	0	8	8	0	0
	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0

(c)

Figure 4. Clustering performance on RU-FACS database. (a) Mean and standard deviation for the feature weights. (b) Temporal clustering accuracy. (b) Confusion matrix for subject S014.



Embedding	Accuracy
ACA	.959(.022)
KPCA	.534(.009)
PCA	.477(.021)

Segmentation	Accuracy
ACA	.874(.044)
KKM	.559(.039)

Figure 6. Clustering of 5 different facial expressions. (a) ACA embedding. (b) Kernel PCA embedding. (c) PCA embedding. (d) Clustering accuracy. (e) Temporal clustering accuracy.

Fig. (4b) shows the mean accuracy and variance of the temporal clustering for both the unsupervised and supervised (learned weights) versions of ACA and KMM. The clustering accuracy with unsupervised ACA was about 63% for lower face AUs and above 75% for upper face AUs. As expected, learning weights for ACA improved the temporal clustering in the upper and lower face although not substantially. The mean and variance for the weights for all the features in the lower and upper face are shown in Fig. (4a). As expected the weights give more importance to the appearance features. Fig. (4c) shows a representative lower-face confusion matrix for subject 14. The AUs were: AU 50 or AU 25, AU 50+12, AU 14, AU 12, AU 12+25+26, AU 12+15 and 17. More details are given in [35].

5.2.2 Infant subject

This experiment shows an application of the proposed techniques to summarize the facial expression of an infant. Infant facial behavior is known to be more temporally complex than that of adults. Fig. 5 shows the results of running unsupervised ACA with 10 clusters in 1000 frames. We used the appearance and shape features for the eyes and mouth. These 10 clusters provide a summarization of the infant's facial events.

5.3. Facial event discovery for sets of subjects

In this section we test the ability of ACA to cluster facial behavior corresponding to different subjects. We first report results for posed facial actions. We then report results for the more challenging case of unposed, naturally occurring facial behavior in an interview setting.

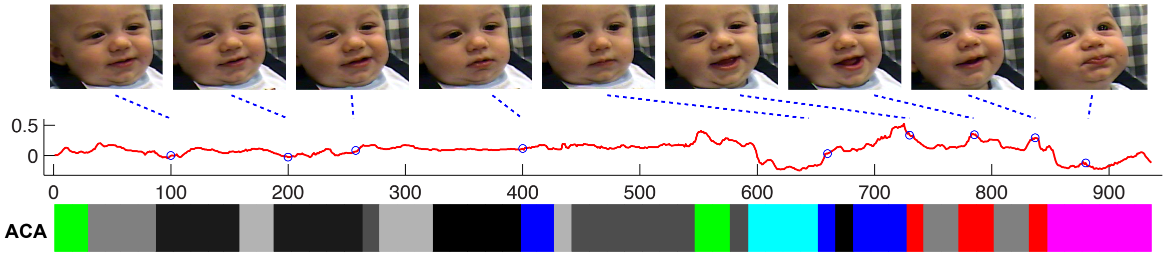


Figure 5. Temporal clustering of an infant facial behavior. Each facial gesture is coded with a different color. The feature represents the angle of the mouth. Observe how the frames of the same cluster correspond to a similar facial expressions.

5.3.1 Sets of subjects in CK

ACA was used to segment the facial expression data from CK database [16]. We used the six shape features and all features were normalized dividing them with respect to the first frame. The frame kernel was computed as a linear combination of 6 kernels with equal weighting.

In the first experiment on the CK database we tested the ability of ACA to temporally cluster several expressions of 30 randomly selected subjects (the number of facial expressions varies across subjects). Fig. (6e) shows the mean (and variance) results for 10 realizations. The number of clusters is five and $n_{max} = 25$. Both ACA and KMM are initialized 10 times and the solution with less error is selected. Again, ACA outperforms KKM. Fig. (7) shows one example of the temporal clustering achieved by unsupervised ACA.

The second experiment explores the use of ACA for visualization of facial events. Fig. (6a) shows the ACA embedding of 112 sequences from 30 randomly selected subjects (different expressions). The embedding is done by computing the first three eigenvectors of the kernel segment matrix ($\mathbf{T}$). In this experiment, the kernel segment matrix is computed using the ground-truth data (expression labels). Each point represents a video segment of facial expression. Fig. (6b) and Fig. (6c) represent the embedding found kernel PCA and PCA using independent frames (the frames are embedded using the first three eigenvectors of the kernel sample matrix $\mathbf{K}$). Because each frame represents a point it is harder to visualize the temporal structure of the facial events. To test the quality of the embedding for clustering, we randomly generated 10 sets of the facial expression for 30 subjects. For each set the ground-truth label is known and the ”optimal” three dimensional embedding is computed. Then we run KMM to cluster the data into five clusters. The results (mean and variance) of the clustering are shown in Fig. (6d). As expected, the segment embedding provided by ACA achieves higher clustering accuracy than kernel PCA or PCA on independent frames.

5.3.2 Sets of subjects in RU-FACS

This section tested the ability of ACA to discover dynamic facial events in a more challenging database of naturally occurring facial behavior of multiple people. Several issues contribute to the challenge of this task on the RU-FACS database. These include non-frontal pose, moderate out-of-plane head motion, subject variability and the exponential nature of possible facial action combinations.

Segmentation	Lower Face	Upper face
ACA + MDA	.522(.045)	.688(.087)
ACA + CAT	.493(.064)	.545(.105)

Figure 8. Temporal clustering across individuals (RU-FACS).

To solve this challenging scenario two strategies are considered: (1) ACA+CAT: concatenate all videos and run ACA in the concatenated video sequence, (2) ACA+MDA: Run ACA independently for each individual and solve for the correspondence of clusters across people using the Multidimensional Assignment Algorithm (MDA) [24]. We propose a heuristic approach to solve the multidimensional assignment problem called Pairwise Approximation Multidimensional Assignment(PA-MDA). Details of the algorithm are given in [35].

Using the same features as section 5.2.1, we randomly selected 10 sets of 5 people and report the mean clustering results and variance. For ACA+MDA, we kept the same parameter setting as in the previous segmentation of one subject. The number of clusters in ACA+CAT was set to $14 \sim 17$ and the length constraint is the same as before (80). As shown in Fig. (8), ACA+MDA achieved more accurate segmentation than ACA+CAT. Moreover, ACA+MDA scales better for clustering many videos. Recall that ACA+CAT scales quadratically in space and time, and this can be a limitation when processing many subjects. As expected the clustering performance is lower than in the case of using only clustering one individual.

Fig. (9a) shows the results for temporal segmentation achieved by ACA+MDA on subjects S012, S028 and S049. Each color denotes a temporal cluster discovered by ACA. Fig. (9) shows some of the dynamic vocabularies for facial expression analysis discovered by ACA+MDA. The algorithm correctly discovers smiling, silent, and talking as different facial events. Visual inspection of all subjects' data suggests that the vocabulary of facial events is moderately

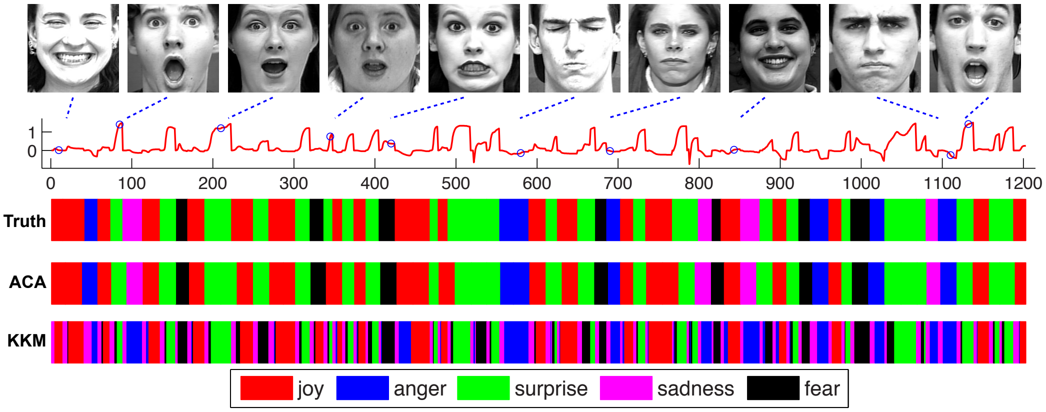


Figure 7. (a) Mouth angle. Blue dots corresponds to frames. (b) Manual labels, Unsupervised ACA and KKM.

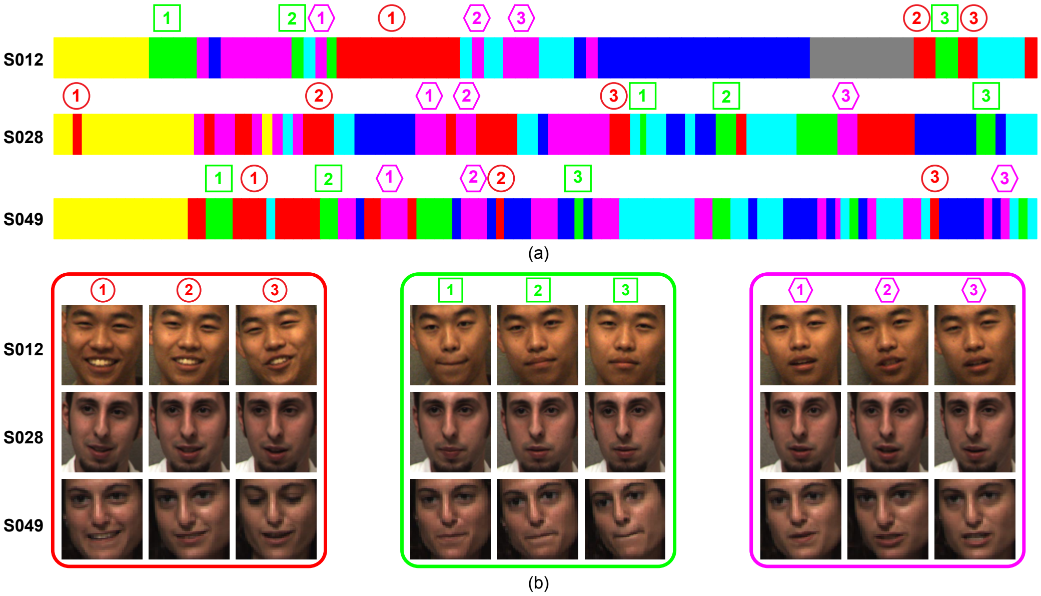


Figure 9. a) Results obtained by ACA on subjects S012, S028 and S049. b) Corresponding frames found by ACA+MDA.

consistent with human evaluation. More details and updated results are given in [35].

6. Conclusions and future work

At present, taxonomies of facial expression are based on FACS or other observer-based schemes. Consequently, approaches to automatic facial expression recognition are dependent on access to corpuses of FACS or similarly labeled video. This is a significant concern, in that recent work suggests that extremely large corpuses of labeled data may be needed to train robust classifiers. This paper raises the question of whether facial actions can be learned directly from video in an unsupervised manner.

We developed a method for temporal clustering of facial behavior that solves for correspondences between dynamic events and has shown promising concurrent validity with manual FACS. In experimental tests using the RU-FACS database, agreement between facial actions identified by unsupervised analysis of face dynamics and FACS approached the level of agreement that has been found between independent FACS coders. These findings suggest that unsupervised learning of facial expression is a promising alternative to supervised learning of FACS-based actions. At least three benefits follow. One is the prospect that automatic facial expression analysis may be freed from its dependence on observer-based labeling. Second, because the current approach is fully empirical, it potentially can identify regularities in video that have not been anticipated by the top-down approaches such as FACS. New discoveries become possible. This becomes especially important as automatic facial expression analysis increasingly develops new metrics, such as system dynamics, not easily captured by observer-based labeling. Three, similar benefits may accrue in other areas of image understanding of human behavior. Recent efforts to develop vocabularies and grammars of human actions [13] depend on advances in unsupervised learning. The current work may contribute to this effort. Current challenges include how best to scale ACA for very large databases and increase accuracy for subtle facial actions. We are especially interested in applications of ACA to detection of anomalous actions and efficient image indexing and retrieval. See [35] for more results.

Acknowledgements This work was partially supported by National Institute of Health Grant R01 MH 051435. Thanks to Tomas Simon for helpful discussions and Iain Matthews for assistance with the AAM code.

References

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *J. Multimedia*, 2006.

[2] F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *BMVC*, 2002.

[3] 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*, 2007.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, 1998.

[5] N. Cristianini, J. Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In *NIPS*, 2002.

[6] M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui. A kernel for time series based on global alignments. In *ICASSP*, 2007.

[7] F. de la Torre. A unification of component analysis methods. In *Handbook of Pattern Recognition and Computer Vision (4th edition)*, October 2009.

[8] F. de la Torre, J. Campoy, Z. Ambadar, and J. Cohn. Temporal segmentation of facial behavior. In *ICCV*, 2007.

[9] F. de la Torre and O. Vinyals. Learning kernel expansions for image classification. In *CVPR*, 2007.

[10] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *PAMI*, 29(11):1944–1957, 2007.

[11] P. Ekman and W. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.

[12] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *PAMI*, pages 757–763, 1997.

[13] G. Guerra-Filho and Y. Aloimonos. A language for human action. *IEEE Computer*, 40(5):42–51, 2007.

[14] Z. Harchaoui, F. Bach, and E. Moulines. Kernel change-point analysis. In *NIPS*, 2009.

[15] J. Hoey. Hierarchical unsupervised learning of facial expression categories. In *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[16] T. Kanade, Y. li Tian, and J. F. Cohn. Comprehensive database for facial expression analysis. In *FG*, pages 46–53.

[17] E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani. An online algorithm for segmenting time series. In *ICDM*, 2001.

[18] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[19] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, pages 135–164, 2004.

[20] D. S. Messinger, M. H. Mahoor, S. M. Chow, and J. F. Cohn. Automated measurement of facial expression in infant-mother interaction: A pilot study. *Infancy*, 14(3):285–305, 2009.

[21] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, (79):299–318, 2008.

[22] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *IJCV*, 77(1-3):103–124, 2008.

[23] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Syst. Man. Cybern. B Cybern.*, 36(2):433–449, 2006.

[24] W. P. Pierskalla. The multidimensional assignment problem. *Oper. Res.*, 16(2):422–431, 1968.

[25] H. Shimodaira, K.-I. Noma, M. Nakai, and S. Sagayama. Dynamic time-alignment kernel in support vector machine. In *NIPS*, 2001.

[26] T. Simon, M. H. Nguyen, F. de la Torre, and J. F. Cohn. Action unit detection with segment-based SVMs. In *CVPR*, 2010.

[27] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *PAMI*, 29(10):1683–1699, 2007.

[28] P. Turaga, A. Veeraraghavan, and R. Chellappa. Unsupervised view and rate invariant clustering of video sequences. *CVIU*, (113):353–371, 2009.

[29] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *CVPR*, 2006.

[30] P. Yin, T. Starner, H. Hamilton, I. Essa, and J. M. Rehg. Learning the basic units in American sign language using discriminative segmental feature selection. In *ICASSP*, 2009.

[31] R. Zass and A. Shashua. A unifying approach to hard and probabilistic clustering. In *ICCV*, 2005.

[32] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, 2001.

[33] L. Zelnik-Manor and M. Irani. Temporal factorization vs. spatial factorization. In *ECCV*, 2004.

[34] H. Zha, X. He, C. H. Q. Ding, M. Gu, and H. D. Simon. Spectral relaxation for k-means clustering. In *NIPS*, pages 1057–1064, 2001.

[35] F. Zhou, F. de la Torre, J. F. Cohn, and T. Simon. Unsupervised discovery of facial events. In *Technical Report TR-10-10. CMU May*, 2010.

[36] F. Zhou, F. de la Torre, and J. K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *FG*, 2008.