

# Exploiting Visual Quasi-Periodicity for Automated Chewing Event Detection using Active Appearance Models and Support Vector Machines\*

Steven Cadavid and Mohamed Abdel-Mottaleb

University of Miami, Department of Electrical and Computer Engineering  
s.cadavid1@umiami.edu, mottaleb@miami.edu

## Abstract

We present a method that automatically detects chewing events in surveillance video of a subject. Firstly, an Active Appearance Model (AAM) is used to track a subject's face across the video sequence. It is observed that the variations in the AAM parameters across chewing events demonstrate a distinct periodicity. We utilize this property to discriminate between chewing and non-chewing facial actions such as talking. A feature representation is constructed by applying spectral analysis to a temporal window of model parameter values. The estimated power spectra subsequently undergo non-linear dimensionality reduction via spectral regression. The low-dimensional representations of the power spectra are employed to train a Support Vector Machine (SVM) binary classifier to detect chewing events. Experimental results yielded a cross-validated percentage agreement of 93.4%, indicating that the proposed system provides an efficient approach to automated chewing detection.

## 1. Introduction

The detection of food consumption is key to the implementation of successful behavior modification in support of dietary monitoring and therapy [6]. Since the vast majority of humans consume food via mastication (chewing), we have designed an algorithm that automatically detects chewing behaviors in surveillance video of a person.

We are inspired by a recent study by Pogalin et al. [5] which demonstrates that quasi-periodic events, such as chewing, can be effectively modeled by the frequency content of PCA-based model parameter variations over time. In [5], the segmented image regions of a video

sequence are aligned and a robust variant of Principal Components Analysis (PCA), known as probabilistic PCA (pPCA), is subsequently applied. The segmented image regions can be reconstructed optimally by a weighted combination of the retained principal components. The frequency content of these weight variations over time is then estimated using the modified periodogram. A feature vector comprised of detected peak frequencies in the power spectra is then employed for event recognition.

Unlike [5] which only considers appearance variations, the proposed approach utilizes the Active Appearance Model (AAM) to jointly represent the shape and appearance components of the target object. The shape component can provide valuable information about an event that may not be adequately conveyed by the appearance. For instance, closed-mouth chewing may demonstrate larger variations in the motion of the jaw line than in the appearance variations of the lip region. Moreover, our method diverges from [5] in both the feature representation and classification scheme. The power spectra obtained by analyzing the frequency content of the AAM parameters over time reside in a redundant, high-dimensional space and are consequently subjected to non-linear dimensionality reduction via spectral regression [1]. Event classification is then performed using Support Vector Machines (SVM) [7].

The remainder of this paper is organized as follows. Sections 2 and 3 describe the AAM method and the spectral analysis of its parameters over time. In Section 4, the technique for non-linear dimensionality reduction of the power spectra is presented. Section 5 reviews the SVM classifier employed for detecting chewing events. Section 6 reports experimental results. Lastly, conclusions and future work are given in Section 7.

## 2. Active Appearance Models

AAM is a statistical representation of an object (e.g., face) introduced by Cootes et al. [2] and improved by

\*This research was supported in part by an NIH grant number 5R21DA024294.

others over recent years. AAM consists of a shape component,  $\mathbf{s}$ , and an appearance component,  $\mathbf{g}$ , that jointly represent the shape and texture variability seen in the object. The shape component represents a target structure by a parameterized statistical shape model obtained from training. The shape model is defined by a linear model:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^m p_i^{(s)} \mathbf{s}_i \quad (1)$$

where  $\mathbf{s}_0$  is the mean shape vector,  $\mathbf{s}_i$  is a set of orthogonal modes (i.e., eigenvectors with associated eigenvalues  $\mathbf{d}^{(s)} = \{d_i^{(s)}\}_{i=1}^m$ ) of shape variation calculated by applying PCA on the covariance matrix of the training shape data, and  $\mathbf{p}^{(s)} = \{p_i^{(s)}\}_{i=1}^m$  is a vector of non-rigid shape parameters. To account for global shape variations, the AAM concatenates the non-rigid shape parameters with four similarity transform parameters,  $a$ ,  $b$ ,  $t_x$ , and  $t_y$  (i.e., scaling, rotation, and translation).

The appearance statistical model is built by warping each image instance so that its control points match the mean shape using a piece-wise affine warp or the thin-plate spline algorithm. The intensity variation is then sampled from the shape-normalized image over the region covered by the mean shape. Similarly, by applying PCA to the appearance data a linear model is defined:

$$\mathbf{g} = \mathbf{g}_0 + \sum_{i=1}^n p_i^{(g)} \mathbf{g}_i \quad (2)$$

where  $\mathbf{g}_0$  is the mean normalized grey-level vector,  $\mathbf{g}_i$  is a set of orthogonal modes (i.e., eigenvectors with associated eigenvalues  $\mathbf{d}^{(g)} = \{d_i^{(g)}\}_{i=1}^n$ ) of intensity variation and  $\mathbf{p}^{(g)} = \{p_i^{(g)}\}_{i=1}^n$  is a set of grey-level parameters. This generates shape data on the facial landmarks and appearance data on the grey-level intensity of each pixel in the face model. The set of eigenvalues are concatenated to form the vector  $\mathbf{v} = [\mathbf{d}^{(s)}, \mathbf{d}^{(g)}]$ , and its usage as weights will be described in the following section.

Given an image  $I$ , the objective of AAM fitting is to find the model parameters  $\mathbf{p} = [a, b, t_x, t_y, \mathbf{p}^{(s)}, \mathbf{p}^{(g)}]$  such that the error between the model-generated image and  $I$  is minimized. This is typically achieved by iteratively updating the model parameters  $\mathbf{p}$  from an initial state through an update function.

### 3. Visual Quasi-Periodicity Analysis

When applying AAM fitting to a video comprised of  $N$  images,  $\mathbf{I} = \{I^t\}_{t=1}^N$ , a set of model parameters,  $P = \{\mathbf{p}^t\}_{t=1}^N$ , is obtained. Similarity-normalized

model parameters can then be acquired by discarding the similarity parameters of  $\mathbf{p}^t$ , retaining only the non-rigid facial motion.

As described in [5], the quasi-periodic nature of chewing is conveyed within the model parameters. A non-parametric spectral analysis method, known as the modified periodogram, is employed to estimate the power spectrum of each model parameter over an  $M$ -frame sliding window. The power spectrum associated with the  $q^{\text{th}}$  model parameter of video frame  $t$  is expressed as:

$$F_q^t(f) = \frac{1}{M} \left| \sum_{k=0}^{M-1} w_k x_k^q \exp(-jk2\pi f) \right|^2 \quad (3)$$

$$\mathbf{x}^q = \{P_{qi}\}_{i=t-M}^{t-1}$$

where  $\mathbf{w}$  denotes the hamming window for periodogram smoothing. The weighted mean of the power spectra  $F_q^t(f)$  is then obtained as follows:

$$\bar{F}^t(f) = \sum_{q=1}^{m+n} d_q^* F_q^t(f), \quad d_q^* = \frac{d_q}{\sum_{i=1}^{m+n} d_i} \quad (4)$$

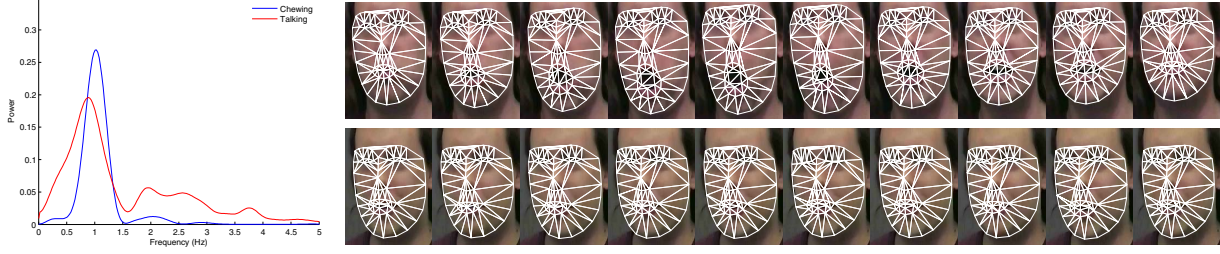
where  $d_q^*$  denotes the relative percentage of retained variance derived from eigenvalue  $d_q$ . We then normalize  $\bar{F}^t$  such that  $\|\bar{F}^t\| = 1$ .

In this application, to discriminate between chewing events and similar behaviors such as talking, a fairly high-resolution power spectra (e.g., 1024-point FFT resulting in 512 dimensions) must be constructed. However, due to the curse of dimensionality, the classification of high-dimensional feature vectors becomes difficult. Therefore, reducing the dimensionality of the power spectra becomes vital and is addressed in the following section.

### 4. Regularized Locality Preserving Indexing via Spectral Regression

Traditionally, linear techniques such as PCA and Linear Discriminant Analysis (LDA) are utilized to project a feature vector from a high dimensional space,  $\mathbb{R}^N$ , into a low-dimensional space,  $\mathbb{R}^n$  ( $n \ll N$ ). Linear techniques, however, have limited ability to represent complex non-linear data such as chewing events in a low-dimensional sub-space [3].

Recently, Cai et al. [1] presented a computationally efficient, non-linear method which has shown success in representing large dimensional data in a low-dimensional sub-space [3]. This algorithm, termed regularized locality preserving indexing, aims to find the mapping function  $\mathbf{a}$  that projects a set of  $K$  samples



**Figure 1. (right) Sample chewing (top) and talking (bottom) sequences. (left) The power spectra associated with the chewing (blue) and talking (red) sequences.**

$X = \{\mathbf{x}_i\}_{i=1}^K \in \mathbb{R}^N$ , to samples  $Y = \{y_i\}_{i=1}^K \in \mathbb{R}^n$  (i.e.,  $\mathbf{y}_i = \mathbf{a}^T \mathbf{x}_i$ ) with corresponding class labels  $\mathbf{c} = \{c_i\}_{i=1}^K$ . This approach is described as follows:

1. Find  $Y$  by solving the following optimization problem:

$$\begin{aligned} Y &= \arg \min_{Y D Y^T = 1} \sum_{i=1}^K \sum_{j=1}^K (\mathbf{a}^T \mathbf{x}_i - y_j)^2 W_{ij} \\ &= \arg \min_{Y D Y^T = 1} Y L Y^T \end{aligned} \quad (5)$$

where  $D$  is a diagonal matrix whose elements are column sums of  $W$  ( $D_{ii} = \sum_j W_{ij}$ ) and  $L = D - W$  is the Laplacian graph. The constraint  $Y D Y^T = 1$  effectively fixes a scaling factor of the solution. The similarity matrix  $W$  is constructed as follows:  $W_{ij} = 1$  if the class labels,  $c_i$  and  $c_j$ , are equivalent and  $W_{ij} = 0$ , otherwise.

This optimization problem, which is also known as Laplacian Eigenmap [4], can be solved efficiently by calculating the eigenvectors of the generalized eigen-problem  $LY = \lambda DY$ .

2. Find  $\mathbf{a}$  such that  $Y = \mathbf{a}^T X$  by solving a regularized least squares problem:

$$\mathbf{a} = \arg \min_{\mathbf{a}} \left\{ \sum_{i=1}^K (\mathbf{a}^T \mathbf{x}_i - y_i)^2 + \alpha \|\mathbf{a}\|^2 \right\} \quad (6)$$

The regularization term guarantees that the least squares problem is well-posed and has a unique solution.

In the following section, we describe the use of SVM to classify each low-dimensional power spectrum as representing either a chewing event or a non-chewing event.

## 5. Chewing Event Detection

SVMs have been used in the fields of machine learning and pattern recognition, and have shown success in

the recognition of facial actions [3]. They derive a decision boundary based on a typically small subset of training samples known as Support Vectors. In order for this property to carry over to SV Regression, Vapnick [7] devised an  $\epsilon$ -intensive loss function which does not penalize errors below some  $\epsilon > 0$ , chosen a priori. This method, known as  $\epsilon$ -SVR, seeks to estimate the functional:

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b \quad (7)$$

based on samples  $\{(\mathbf{x}_i, c_i)\}_{i=1}^K \in \mathbb{R}^n \times \mathbb{R}$  by minimizing the regularized risk functional  $\frac{1}{2} \|\mathbf{w}\|^2 + C \cdot R_{emp}^\epsilon$  where  $c_i$  is the class label associated with the low-dimensional power spectrum  $\mathbf{x}_i$ ,  $C$  is a constant determining the trade-off between minimizing training errors and minimizing the model complexity term  $\|\mathbf{w}\|^2$ , and  $R_{emp}^\epsilon := \frac{1}{K} \sum_{i=1}^K |c_i - f(\mathbf{x}_i)|_\epsilon$ .

Additionally, kernel functions, such as the Radial Basis Function (RBF) kernel, are employed to map samples which may not be linearly separable to a feature space where linear methods can then be applied.

Chewing events typically take place over several seconds, spanning several hundreds of consecutive frames. The proposed method predicts a class label for a given frame independent of the predicted class labels of previous frames. This may give rise to abrupt changes in the class label between subsequent frames. Therefore, we introduce a regularization term to the following classification scheme to impose smoothness over the predicted class labels:

$$\arg \min_{\varsigma} \left\{ |\varsigma - f(\mathbf{x}_t)| + \lambda \left( 1 - \frac{1}{N} \sum_{i=t-N}^{t-1} \varphi(\varsigma, c_i) \right) \right\} \quad (8)$$

The first term in (8) measures the  $L^1$ -norm distance between the class label,  $\varsigma \in \{0, 1\}$ , and the real-valued functional  $f(\mathbf{x}_t)$ . The second term, weighted by  $\lambda$ , imposes smoothness by promoting the class label with the greatest number of occurrences over the  $N$  previous frames ( $\varphi(a, b) = 1$  if  $a = b$  and 0, otherwise). The class label that minimizes (8) is designated the class label of frame  $t$ ,  $c_t$ .

## 6. Experimental Results

We applied the proposed approach to a publicly-available dataset consisting of seven three-minute videos (captured at 24 fps) of subjects performing both chewing and non-chewing actions<sup>1</sup>. Each video is comprised of five action segments: 1) closed-mouth chewing, 2) open-mouth chewing, 3) an assortment of facial expressions, 4) talking, and 5) still face. There was no condition placed on the subjects' head movements, although the majority of the subjects maintain a frontal head pose. All of the captured frames were used in our experiments except for the first 72 frames (three seconds) of every video (a 72-frame window is needed to build a frequency spectrum). This resulted in 29,155 frames being used in our experiments.

We conducted a series of experiments using the Leave-One-Subject-Out (LOSO) validation method. That is, six of the videos are used for training and the remaining video is used for testing. This process is repeated seven times to test every video in the dataset. A human coder manually coded the presence of chewing in each video frame (i.e., class label of 1 for frames containing chewing events and 0, otherwise). The manual coding is then employed for both the training and testing of our system. Subject-dependent AAMs were trained based on the manual annotations of a 3% frame sampling of each video. The frequency spectra of every 4<sup>th</sup> sample of the training set were used to learn the projection matrix,  $\mathbf{a}$ , based on spectral regression. The low-dimensional samples were subsequently used to train the binary SVM classifier.

We compare the predicted and manually coded class labels in Table 1. The second column reports the percentage agreement (PA) between the predicted and actual class labels. Columns 3 and 4 illustrate the percentage distribution between chewing (C) and non-chewing (NC) frames. Lastly, columns 5 and 6 present the percentage of NC frames which are incorrectly labeled as C frames (FAR), and the percentage of C frames which are incorrectly labeled as NC frames (FRR), respectively. As the table demonstrates, our approach performs well in detecting chewing events; The average percentage agreement, FAR, and FRR are 94.3%, 5.5%, and 5.4%, respectively. False accepts were primarily contained within the talking segments. This is due to the similar appearance, motion and quasi-periodicity of the two actions. False rejects, conversely, were primarily caused by a failure to detect subtle, closed-mouth chewing.

<sup>1</sup>Data Set 2 can be downloaded from [http://www.icta.ufl.edu/projects\\_nih/data/chewingV1.htm](http://www.icta.ufl.edu/projects_nih/data/chewingV1.htm)

Sub.	PA	C	NC	FAR	FRR
1	96.1	47.6	52.4	3.1	4.7
2	94.1	44.2	55.8	4.5	7.6
3	93.5	38.4	61.6	3.4	11.4
4	93.6	35.1	64.9	3.7	11.4
5	89.5	35.3	64.7	16.2	0.0
6	97.3	38.5	61.5	2.7	2.7
7	95.8	13.3	86.7	4.9	0.0
<b>Avg.</b>	<b>94.3</b>	<b>36.1</b>	<b>63.9</b>	<b>5.5</b>	<b>5.4</b>

**Table 1. Percentage agreement, FAR, and FRR between the predicted and actual class labels.**

## 7. Conclusion

In this paper, we presented a framework that utilizes the concept of visual quasi-periodicity to detect chewing events from surveillance video. Applications of this algorithm include obesity, diabetes, and cardiovascular disease control. Future work will include testing on additional subjects as well as the development of a face model that demonstrates greater subject independence.

## References

- [1] D. Cai, X. He, W. V. Zhang, and J. Han. Regularized locality preserving indexing via spectral regression. In *Proc. of the ACM conference on Conf. on Information and Knowledge Management*, pages 741–750, 2007.
- [2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Proc. of the European Conf. on Computer Vision*, 2:484–498, 1998.
- [3] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *CVPR Workshop on Human communicative Behavior analysis (CVPR4HB)*, June 2009.
- [4] P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.
- [5] E. Pogalin, A. Smeulders, and A. Thean. Visual quasi-periodicity. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.
- [6] M. Schmalz, A. Mendez-Vazquez, and A. Helal. Algorithms for the detection of chewing behavior in dietary monitoring applications. In *Proc. of SPIE Technical Symposium: Mathematics of Data/Image Coding, Compression, and Encryption with Applications XII*, volume 7444A, August 2009.
- [7] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.