Unsupervised Extraction of Human-Interpretable Nonverbal Behavioral Cues in a Public Speaking Scenario

M. Iftekhar Tanveer
ROC HCI
Electrical and Computer Eng.
University of Rochester
itanveer@cs.rochester.edu

Ji Liu
ROC HCI
Computer Science
University of Rochester
jliu@cs.rochester.edu

M. Ehsan Hoque
ROC HCI
Computer Science
University of Rochester
mehoque@cs.rochester.edu

ABSTRACT

We present a framework for unsupervised detection of nonverbal behavioral cues—hand gestures, pose, body movements, etc.—from a collection of motion capture (MoCap) sequences in a public speaking setting. We extract the cues by solving a sparse and shift-invariant dictionary learning problem, known as *shift-invariant sparse coding*. We find that the extracted behavioral cues are human-interpretable in the context of public speaking. Our technique can be applied to automatically identify the common patterns of body movements and the time-instances of their occurrences, minimizing time and efforts needed for manual detection and coding of nonverbal human behaviors.

Categories and Subject Descriptors

G.1 [**NUMERICAL ANALYSIS**]: Optimization; J.5 [**ARTS AND HUMANITIES**]: Performing arts

Keywords

Public Speaking; Action Recognition; Unsupervised Analysis; Sparsity; Shift-Invariant Sparse Coding

1. INTRODUCTION

Public speaking is a widely-used method for articulating ideas. Understanding the influence of nonverbal behaviors in a public speaking setting is an interesting research topic [6]. It might be possible to improve our understanding of such influence by using data analytic approaches over a large collection of public speaking data. Human *Activity* or *Action* Recognition is a growing field of research that is already aiming towards such analysis. However, this domain is mostly focused on supervised classification of body language [1, 4]. It is often necessary to utilize unsupervised approaches in order to extract common body movement patterns without prior knowledge. For example, if we want to know all the common fidgeting patterns in a public speaking setting, we need an unsupervised algorithm to find these patterns. With supervised approach, we would have to provide samples of the fidgeting behaviors as an input to the algorithm. This is infeasible due to the lack of prior knowledge.

We propose a framework[1] to automatically identify and localize the common *behavioral cues* [6] in a public speaking video. In order to track the body movements of the speaker, we use a Kinect skeleton tracker [12]. The tracker provides time dependent, three dimensional signal of 20 joint locations of the speaker—collectively known as Motion Capture (MoCap) signal. We apply *shift-invariant sparse coding (SISC)* [9, 5] to extract the behavioral cues that are manifested as small temporal patterns in the MoCap signal. SISC is formulated as an optimization problem for learning a dictionary of temporal patterns, where the patterns appear within the signal in a sparse manner. We notice that the learned set of patterns are human-interpretable. People can relate the extracted behavioral cues to real-life public speaking scenarios.

Our work has several future applications. For instance, it can be used to mine different types of behavioral patterns from time sequence data (e.g. variations in facial expressions, vocal characteristics, body language, etc.). As the patterns are human-interpretable, this technique can be used to build an ontological coding schema for behavioral cues by simultaneously learning and manually annotating meaningful labels. Finally, the behavioral cues learned from this process could be used as "features" to train supervised algorithms in order to predict human performance for public speaking or any other forms of human communications.

2. RELATED LITERATURE

Computational analysis of nonverbal behaviors [14, 8, 1] has recently received a considerable attention from the researchers. Metaxas et al. [8] summarized works that focus on sensing body movements and representing them as a sequence of numbers (facial point trackers represent facial movements as a sequence of landmark points, Kinect skeleton trackers [12] represent full body movements as a sequence of joint locations, etc.)

Researchers also work on holistically predicting the outcome of specific tasks through automated analysis of nonverbal behaviors—outcomes of dating [11], job interviews [10], public speaking [2], etc. Most of these works use a fixed taxonomy and summary statistics like mean, variance, minima, maxima, count. These features may not be granular enough to segment or detect nonverbal behavioral cues.

[1]A demo is available in http://hci.cs.rochester.edu

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806350.

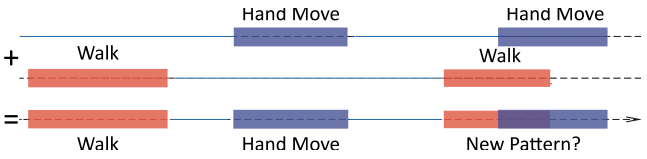

Figure 1: A simplified illustration of a problem that can be introduced from overlapping activities.

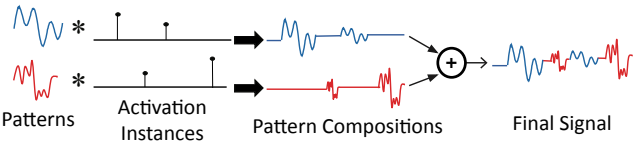

Figure 2: An assumption about the behavioral signal generation model.

Activity analysis [1, 4] is another area related to nonverbal behavioral analysis. As Aggarwal et al. described [1], the focus of this type of work is to identify the activities (walking, jumping, talking, etc.) within a video, as well as to identify the start and end times of the segments. This domain is mostly focused on supervised classification approaches.

There exist a few unsupervised methods for activity analysis as well. Zhou et al. [15] proposed a method named Aligned Cluster Analysis (ACA) for simultaneously segmenting and clustering temporal sequences using Dynamic Time Warping (DTW) and k-means clustering. In this approach, the temporal segments are assumed to be non-overlapping. However, in real life, the segments might overlap depending on the signal representations. For example, if we consider only the three dimensional coordinates of a person's hand, it may appear as different patterns depending on whether the person is showing a hand gesture or physically moving about (Figure 1). Now, if the person performs both actions together, the resulting pattern will be overlapped. Many different spurious patterns may emerge depending on how they overlap—these are typically difficult to detect. Our model can address this issue.

Our work is inspired by the SISC model presented by Mørup et al. [9]. An optimization problem similar to SISC is applied by Li et al. [7] in the action recognition problems. However, they considered the input as a set of K independent 1D signals. On the other hand, we consider it as a multivariate signal with K components. For MoCap data, our model allows the temporal patterns to capture the interdependence among the body-joints. In addition, we solve the optimization problem using a Gradient Descent (GD) approach, as opposed to the Orthogonal Matching Pursuit (OMP) used by Li et al [7]. GD relaxes the requirement of specifying a maximum number of repetitions for the patterns within the signal.

3. NONVERBAL BEHAVIORAL CUES

Behavioral scientists define nonverbal behavioral cues as patterns observed in gestures, posture, touching behavior, facial expressions, eye behavior, vocal behavior, etc. [6]. Vinciarelli et al. [14] notes that "The term behavioural cue is typically used to describe a set of temporal changes in neuromuscular and physiological activity that last for short intervals of time (milliseconds to minutes)."

To capture the essence of this definition, we assume a specific structure for the behavioral (MoCap) signals, as illustrated in Figure 2. We model the signals to be composed of several small patterns (behavioral cues). These patterns get activated sparsely at various time-instances. This sparse activation ensures that a pattern is not distorted by overlapping with itself. However, different patterns are allowed to overlap. A mathematical formulation of this model is discussed in the following section.

4. PROBLEM FORMULATION

Let $\mathbf{f}[n]$ be a finite length multivariate behavioral signal with K elements. More specifically, it is the output of the skeleton tracker at every frame, representing the x, y, z coordinates of the body joints. The length of the signal is N; i.e. $0 \leq n < N$. The mathematical model of the behavioral signal, $\mathbf{f}_{\text{model}}[n]$, is as follows:

$$\sum_{d=0}^{D-1} \alpha_d[n] * \psi_d[m] = \sum_{d=0}^{D-1} \sum_{u=0}^{N-1} \alpha_d[u]\psi_d[n-u] \qquad (1)$$

where, $\psi_d[m]$ denotes the d^{th} pattern ($0 \leq d < D$, $0 \leq m < M$). In general, the patterns are short; i.e. $M \ll N$. The activation sequence, $\alpha_d[n]$, consists of sparse impulses. These impulses represent the locations where the pattern d appears in the signal. Length of $\alpha_d[n]$ is the same as the signal length, N. The convolution ($*$) of $\psi_d[m]$ with a single impulse of $\alpha_d[n]$ shifts the pattern d to the location of the impulse. We use this idea to make the behavioral cues *shift-invariant*.

To estimate the model parameters, $\psi_d[m]$ and $\alpha_d[n]$, we minimize the total squared difference between the actual signal and the model, $\mathbf{f}_{\text{model}}[n]$. As the behavioral cues activate sparsely over time, most of the elements of α should be zero. We enforce sparsity over the activation sequence by minimizing the ℓ_1 norm of α. In addition, we enforce non-negativity over α to avoid upside-down patterns. The overall optimization problem is shown in the Eq (2)

$$\hat{\psi}[m], \hat{\alpha}[n] = \underset{\psi,\alpha}{\text{argmin}} \underbrace{\frac{1}{2} \left\| \mathbf{f}[n] - \mathbf{f}_{\text{model}}[n] \right\|^2}_{P(\psi,\alpha)} + \underbrace{\lambda \left\| \alpha \right\|_1}_{Q(\alpha)} \qquad (2)$$

$$\text{s.t.} \quad \|\psi\|_F^2 \leq 1 \quad \text{and,} \quad \forall_n \alpha[n] \geq 0.$$

Here, $\|\alpha\|_1 := \sum_{d=0}^{D-1} \sum_{n=0}^{N-1} |\alpha_d[n]|$ represents the ℓ_1 norm of α. λ is the Lagrange multiplier controlling the weights imposed on the sparsity constraint of α. We also use a constraint $\|\psi\|_F^2 \leq 1$ to ensure that ψ is not affected by the values of λ.

5. OPTIMIZATION

The objective function shown in (2) is generally non-convex. However, when any one of the model parameters (α or ψ) is held fixed, it becomes convex over the other parameter. We use *alternating proximal gradient descent* approach to solve this optimization problem, as shown in Algorithm 1. In this approach, we alternatively update the parameters. For instance, we hold ψ fixed while updating α, and vice versa. It reduces the error after each iteration and guarantees to converge. However, there is no guarantee that it will converge to the global optimum. We rerun the algorithm multiple times with random initialization to make it likely to find the global optimum.

Algorithm 1: Learning the Behavioral Cues

Input: $\mathbf{f}[n]$, M, D and λ
Output: ψ, α
Initialize;
$i \leftarrow 0$;
$\alpha \leftarrow 0$, $\psi \leftarrow$ random;
while *not Converge* **do**
  Update ψ;
  reconstruct $\mathbf{f}_{model} \leftarrow \sum_{d=1}^{D-1} \alpha_d * \psi_d$;
  calculate $\nabla_\psi P$ using $\mathbf{f}, \mathbf{f}_{model}$ and α [Eq. (4)];
  $\psi^{(i+1)} \leftarrow$ **project**$(\psi^{(i)} - \gamma_\psi \nabla_\psi P)$;
  Update α;
  reconstruct $\mathbf{f}_{model} \leftarrow \sum_{d=1}^{D-1} \alpha_d * \psi_d$;
  calculate $\nabla_\alpha P$ using $\mathbf{f}, \mathbf{f}_{model}$ and ψ [Eq. (5)];
  $\alpha^{(i+1)} \leftarrow$ **shrink**$(\alpha^{(i)} - \gamma_\alpha \nabla_\alpha P)$ [Eq. (3)];
  $i \leftarrow i + 1$

The objective function in (2) is a composite function with a smooth part, $P(\psi, \alpha)$, and a non-smooth part, $Q(\alpha)$. We use an *iterative soft-thresholding* approach for this minimization, as reviewed by Beck et al. [3]. We update α by the gradients of $P(\psi, \alpha)$, followed by a **shrink** operation, as shown in (3). This operation reduces each component of α towards zero and thus resulting in a sparse solution. We simultaneously project α to the set of positive numbers to enforce non-negativity.

$$\begin{aligned} &\alpha[n] \leftarrow \mathrm{sgn}(\alpha[n]) \max(0, |\alpha[n]| - \gamma\lambda) \quad \forall_{0 \le n < N} \\ &\alpha[n] \leftarrow max(0, \alpha[n]) \quad \forall_{0 \le n < N} \end{aligned} \tag{3}$$

The gradients of P with respect to ψ and α ($\nabla_\psi P$ and $\nabla_\alpha P$) are given by Eq (4) and (5), respectively.

$$\frac{\partial P}{\partial \psi_{d',k'}[m']} = \sum_{n=0}^{N-1} \{f_{\mathrm{model},k'}[n] - f'_k[n]\} \alpha_{d'}[n - m'] \tag{4}$$

$$\frac{\partial P}{\partial \alpha_{d'}[n']} = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \{f_{\mathrm{model},k}[n] - f_k[n]\} \psi_{d',k}[n - m'] \tag{5}$$

In order to determine a correct learning rate γ_ψ or γ_α, we use a *backtracking line search* approach. We gradually decrease the learning rate until the objective function, $\mathcal{F}(x_i)$, satisfies $\mathcal{F}(x_{i+1}) \le \mathcal{M}_{x_i,\gamma_i}(x_{i+1})$, where, $\mathcal{M}_{x_i,\gamma_i}$ represents a function as defined in (6).

$$\begin{aligned} &\mathcal{M}_{x_i,\gamma_i}(x_{i+1}) := \mathcal{F}(x_i) + \langle \nabla \mathcal{F}(x_i), x_{i+1} - x_i \rangle \\ &+ \frac{1}{2\gamma_i} \|x_{i+1} - x_i\|^2 + Q(\alpha). \end{aligned} \tag{6}$$

Here, x_i refers to the model parameters (ψ or α) in the i^{th} iteration.

In the **project** procedure, we use (7) to project $\psi_d[m]$ on the set $\{\psi_d[m] \mid \|\psi_d[m]\|_F^2 \le 1\}$ at every iteration. This enforces the constraint $\|\psi_d[m]\|_F^2 \le 1$.

$$\psi_d[m] \leftarrow \min\left(\|\psi_d[m]\|_F, 1\right) \frac{\psi_d[m]}{\|\psi_d[m]\|_F} \tag{7}$$

We set the value of λ using the L-curve method, as described by Mørup et al. [9].

Figure 3: A sample frame from one of the videos in the public speaking dataset.

6. DATA

We applied this algorithm on two different types of datasets. The first type was a synthetically-formulated dataset with overlapping patterns. We used this data to empirically check if the algorithm could solve the problem described in Figure 1. The second type was a real dataset, collected in an actual public speaking scenario [13].

The public speaking dataset contained videos (Figure 3) and MoCap sequences of 55 public speeches given by 20 students—12 males and eight females. Each speech was approximately three minutes long. The MoCap sequences were captured by a Kinect skeleton tracker. All the sequences contained x, y, and z coordinates of 20 joints of the speaker's body. Therefore, each MoCap signal was a time-varying signal with 60 components.

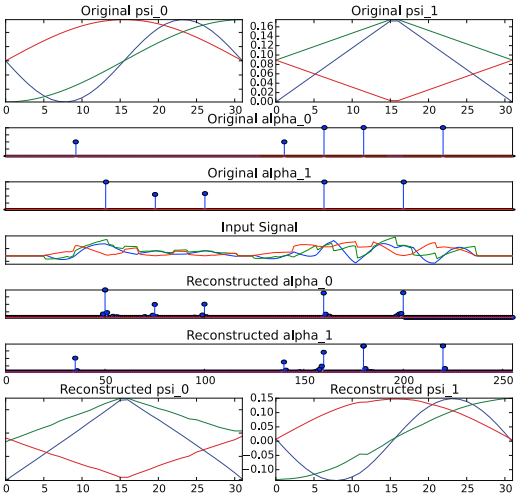

Figure 4: The test of SISC algorithm using synthetically-generated data.

7. RESULTS

The results of running the algorithm on both synthetic and real data are discussed below.

7.1 Synthetic Data

We illustrate a synthetically-generated data in Figure 4. The top row shows two patterns (ψ_0 and ψ_1) in a three dimensional signal (i.e. K = 3) that were used to construct the input data (row four). The second and the third row illustrate the activation sequences α_0 and α_1, respectively. We notice that the reconstructed patterns (row seven) are

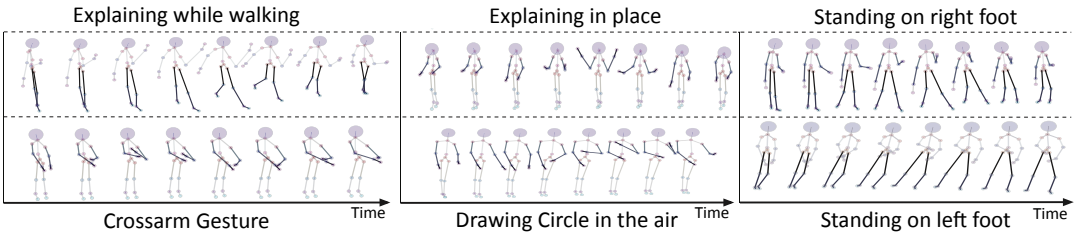

Figure 5: A few behavioral cues automatically extracted from the public speaking data. The length of each behavioral cue (M) was set to 2 seconds. We highlighted the bones showing large movements in time.

almost identical to the original patterns (row one), although their order is not same. Notice that the original activation sequences (second and third rows) contain an impulse at index 160, indicating an overlap of the patterns. The reconstructed plots of the activation sequences (fifth and sixth rows) show that this overlap was perfectly identified and decoupled. This shows that SISC can handle overlapping patterns. However, this is not true for all possible cases. For example, if the result of overlapping signals is orthogonal to any of its constituents, SISC will not be able to decouple them. Nevertheless, such orthogonality is not likely in practice.

7.2 Real Data

The algorithm was run individually on each public speaking MoCap sequence. A few behavioral cues captured by the algorithm are shown in Figure 5. Each sequence of the skeletons represents a single behavioral cue that the algorithm extracted automatically. The position of the skeleton represents different instances in time. We observed that, the patterns represent the most common body movements in the sequence. We say the behavioral cues are "human-interpretable" if we can associate the retrieved movement patterns of the skeleton with the actual body movements shown in the videos purely through visual observation. However, both the notion of interpretibility and the actual interpretation of the patterns are subjective and manually assigned by the authors.

The time length of the behavioral cues, M, was heuristically set to two seconds. If M is too small (< 0.5 sec), only short parts of the original sequence are captured. In that case, the captured patterns become difficult to interpret. On the other hand, increasing M ($2 \text{ sec} < M < 5$ sec) does not seem to have much influence on the interpretibility of the patterns. However, we've noticed multiple patterns of body movements to be merged together when M is too large (e.g. $M > 8$ seconds).

The parameter D controls the maximum number of patterns to be extracted from the signal. The extracted patterns lose interpretibility if D is set too low (e.g. $D \leq 2$). On the other hand, setting D to a large value does not have any negative effect on the interpretibility. In that case, the additional patterns show up as a signal of all zeros. However, increasing D slows down the program linearly.

8. CONCLUSION

In this paper, we presented a mathematical framework for unsupervised extraction of body movement patterns. We used shift-invariant sparse coding to extract the patterns. We noticed that the extracted patterns are human-interpretable, that is, we could manually associate the patterns with the movements shown in the video. In the future, this framework could be used to mine common body movement patterns (e.g., fidgeting) in public speaking scenarios. We shall try to apply this framework to extract patterns from other behavioral signals (e.g., facial expressions, prosody, etc.). A demo of our framework is available in hci.cs.rochester.edu.

9. REFERENCES

[1] J. K. Aggarwal et al. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 2011.

[2] L. Batrinca et al. Cicero-towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents*, 2013.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[4] G. Cheng et al. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*, 2015.

[5] R. Grosse et al. Shift-invariant sparse coding for audio classification. In *UAI*, 2007.

[6] M. Knapp, J. Hall, and T. Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[7] Y. Li et al. Learning shift-invariant sparse representation of actions. In *CVPR*, 2010.

[8] D. Metaxas and S. Zhang. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 2013.

[9] M. Mørup et al. Shift invariant sparse coding of image and music data. Technical Report IMM2008-04659, Technical University of Denmark, 2008.

[10] I. Naim et al. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *FG'15*, 2015.

[11] R. Ranganath et al. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *EMNLP*, 2009.

[12] J. Shotton et al. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013.

[13] M. I. Tanveer et al. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *IUI'15*, pages 286–295. ACM, 2015.

[14] A. Vinciarelli et al. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*.

[15] F. Zhou et al. Aligned cluster analysis for temporal segmentation of human motion. In *FG'08*, 2008.