Unsupervised Temporal Segmentation of Human Activities in Video

Feng Zhou
Robotics Institute
Carnegie Mellon University
www.f-zhou.com

Fernando de la Torre
Robotics Institute
Carnegie Mellon University
ftorre@cs.cmu.edu

Jessica K. Hodgins
Robotics Institute
Carnegie Mellon University
jkh@cs.cmu.edu

Abstract

Temporal segmentation of human actions in video is central to understand and build computational models of human motion. Several issues contribute to the challenge of temporal segmentation of human motion from video. These include the large variability in the temporal scale and periodicity of human actions, the complexity of representing articulated motion, and the exponential nature of all possible movement combinations. We formulate the temporal segmentation problem as an unsupervised learning problem, and propose Aligned Cluster Analysis (ACA), an extension of standard kernel k-means clustering to cluster time series. ACA extends standard kernel k-means clustering in two ways: (1) allows the cluster means contain a variable number of features and (2) introduces a generalized dynamic time warping (DTW) kernel as temporal metric between sequences. Experimental results reported on the Weizmann and KTH action datasets demonstrate the effectiveness of ACA for factorizing human actions in video.

1 Introduction

Visual analysis of human motion is currently one of the most active research topics in computer vision. Visual analysis of human motion attempts to detect, track and identify people, and more generally, to interpret human behaviors, from image sequences involving humans.

A key component to understanding humans in video is to be able to temporally segment activities in an unsupervised manner. Although previous research [1, 2, 3, 4, 5] has shown promising results, factorizing human motion into coherent temporal actions is still an unsolved problem in human motion analysis. The inherent difficulty of human motion segmentation stems from the large intra-person physical variability, irregularity in the periodicity of human actions and the exponential nature of possible movement combinations. Fig. 1 illustrates the problem of temporal segmentation: given a stream of video data, find a temporal decomposition of the video into coherent motion patterns.

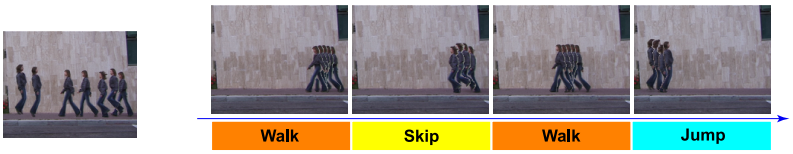


Figure 1: Unsupervised temporal segmentation of human activities in video.

This paper formulates the problem of temporal segmentation as one of time series clustering, and proposes Aligned Cluster Analysis (ACA). ACA presents a number of advantages: (a) The temporal segmentation problem is posed as a versatile energy minimization problem. The energy function is an extension of kernel k-means that allows to cluster temporal sequences, incorporate semi-supervised information and manually select the temporal granularity of the decomposition. (b) As any clustering algorithm, the exact solution of ACA is an NP complete problem. We propose an efficient coordinate descent minimization algorithm to solve ACA via dynamic programming (DP).

2 Previous Work

There exists an extensive literature in computer vision that addresses the problem of grouping human actions. Barbic *et al*. [6] proposed an algorithm to decompose human motion into distinct actions by detecting sudden changes in the intrinsic dimensionality of the Principal Component Analysis (PCA) model. Zelnik-Manor and Irani [7] extracted spatio-temporal features at multiple temporal scales to isolate and cluster events. De la Torre *et al*. [8] proposed a geometric-invariant clustering algorithm to decompose a stream of facial behavior into facial gestures. Guerra-Filho and Aloimonos [1] presented a linguistic framework for modeling and learning of human activity representations. Minnen *et al*. [9] discovered motifs in real-valued, multivariate time series data as locating regions of high density in the space of all time series subsequences. Preliminary version of this work have been presented at [10].

3 Temporal segmentation

This paper formulates the problem of temporal segmentation as an unsupervised clustering problem.

3.1 Dynamic time alignment kernel

Dynamic Time Warping (DTW) has typically been preferred over Euclidean metrics as a distance to match time series because of its robustness to temporal scaling. One of the drawbacks of using DTW as a metric is that it does not satisfy the triangle inequality. Shimodaira *et al*. [11] extended DTW to satisfy the Cauchy-Schwartz inequality and proposed Dynamic Time Alignment Kernel (DTAK). For two sequences, $\mathbf{X} \doteq [\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}] \in \mathbb{R}^{d \times n_x}$ and $\mathbf{Y} \doteq [\mathbf{y}_1, \cdots, \mathbf{y}_{n_y}] \in \mathbb{R}^{d \times n_y}$, DTAK is defined as,

$$\tau(\mathbf{X}, \mathbf{Y}) = \frac{p_{n_x, n_y}}{n_x + n_y} \text{ with } p_{i,j} = \max \begin{cases} p_{i-1,j} + \kappa_{ij} \\ p_{i-1,j-1} + 2\kappa_{ij} \\ p_{i,j-1} + \kappa_{ij} \end{cases} \text{ and } p_{0,0} = 0, \tag{1}$$

where $\kappa_{ij} = \phi(\mathbf{x}_i)^T \phi(\mathbf{y}_j)$ is the frame kernel that constitutes the Gram matrix $\mathbf{K} \in \mathbb{R}^{n_x \times n_y}$.

The recursive calculation of DTAK (1) maximizes the weighted sum of similarity over the path that the sequences ($\mathbf{X}$ and $\mathbf{Y}$) have been aligned, *i.e*.,

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

where l is the number of indexes (or steps) needed to align both signals and $\mathbf{Q} \in \mathbb{R}^{2 \times l}$ denotes the correspondence of frames, *i.e*., the q_{1c}^{th} frame in $\mathbf{X}$ and the q_{2c}^{th} frame in $\mathbf{Y}$ have been aligned. Let's denote the normalized correspondence matrix $\mathbf{W} \in \mathbb{R}^{n_x \times n_y}$ as, $w_{ij} = \frac{1}{n_x + n_y}(q_{1c} - q_{1c-1} + q_{2c} - q_{2c-1})$ if there exist $q_{1c} = i$ and $q_{2c} = j$ for some c, otherwise $w_{ij} = 0$. Then we obtain an alternative representation for DTAK,

$$\tau(\mathbf{X}, \mathbf{Y}) = \mathrm{Tr}(\mathbf{K}^T \mathbf{W}) = \psi(\mathbf{X})^T \psi(\mathbf{Y}) \tag{3}$$

where $\psi(\cdot)$ is to implicitly map the sequence into a feature space.

3.2 Objective function

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

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

$$J_{aca}(\mathbf{G}, \mathbf{s}) = \sum_{c=1}^{k} \sum_{i=1}^{m} g_{ci} \|\psi(\mathbf{Y}_i) - \psi(\mathbf{m}_c)\|^2, \tag{4}$$

where $\mathbf{Y}_i \doteq \mathbf{X}_{[s_i, s_{i+1})}$ is a segment and $\psi(\cdot)$ is a non-linear mapping[1] such that, $\tau_{ij} = \psi(\mathbf{Y}_i)^T \psi(\mathbf{Y}_j) = \mathrm{Tr}(\mathbf{K}_{ij}^T \mathbf{W}_{ij})$.

[1] Unfortunately DTAK is not strictly positive definite kernel [12]. We add the diagonal of Gram matrix $\mathbf{K} \leftarrow \mathbf{K} + \sigma \mathbf{I}_n$ with a constant value σ, which is chosen by the absolute value of the smallest eigenvalue of the DTAK matrix $[\tau_{ij}]_{n \times n}$ if the DTAK has negative eigenvalues.

The matrix 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 $\mathbf{W} \in \mathbb{R}^{n \times n}$,

$$J_{aca} = \mathrm{Tr}\left((\mathbf{L} \circ \mathbf{W})\mathbf{K}\right) \text{ with } \mathbf{L} = \mathbf{I}_n - \mathbf{H}^T\mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{G}\mathbf{H}, \tag{5}$$

where $\mathbf{H} \in \{0,1\}^{m \times n}$ is the frame-segment indicator matrix; $h_{ij} = 1$ if j^{th} frame belong to i^{th} segment. Consider the special case when, $m = n$ and $\mathbf{H} = \mathbf{I}_n$, which implies each frame is treated as a segment. In this case, the dynamic kernel is simply the frame kernel, *i.e.*, $\mathbf{W} = \mathbf{1}_n\mathbf{1}_n^T$ and ACA is equivalent to the kernel k-means,

$$J_{kkm}(\mathbf{G}) = \mathrm{Tr}\left(\mathbf{L}\mathbf{K}\right) \text{ with } \mathbf{L} = \mathbf{I}_n - \mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{G}. \tag{6}$$

3.3 Coordinate descent optimization

Similar to kernel k-means, we take a coordinate-descent algorithm to optimize (4):

$$\arg\min_{\mathbf{G},\mathbf{s}} \sum_{c=1}^{k}\sum_{i=1}^{m} g_{ci}\|\psi(\mathbf{Y}_i) - \psi(\mathbf{m}_c^0)\|^2, \tag{7}$$

where $\mathbf{m}_c^0$ is the cluster mean implicitly computed from the segmentation $(\mathbf{G}^0, \mathbf{s}^0)$ derived in the previous step.

Given a sequence $\mathbf{X}$ of length n, the number of possible segmentations is exponential (*i.e.*, $O(2^n)$), which makes a brute-force search infeasible to be carried out. We adopt a DP-based algorithm with complexity $O(n^2 w_{max})$ to exhaustively examine all the possible segmentations,

$$J(1,v) = \min_{v - w_{\max} < i \leq v}\left(J(1,i-1) + \min_{\mathbf{g}} \sum_{c=1}^{k} g_c\|\psi(\mathbf{X}_{[i,v]}) - \mathbf{m}_c^0\|^2\right), \tag{8}$$

where $J(u,v) \doteq \min_{\mathbf{G},\mathbf{s}} J_{aca}|_{\mathbf{X}_{[u,v]}}$ is an auxiliary function to relate the minimum energy directly with anchors (the position of head u and tail v) of the subsequence $[\mathbf{x}_u, \mathbf{x}_{u+1}, \cdots, \mathbf{x}_v]$. When $v = n$, the $J(1,n)$ is the optimal cost of the segmentation that we seek.

4 Experiments

4.1 Synthetic Data

We synthetically generated time series, $\mathbf{X} \in \mathbb{R}^{2 \times n}$, by randomly sampling twelve 2-D Gaussian distributions that have been grouped into three temporal clusters. Several artificial samples are randomly inserted into the sequence, acting as the temporal noise. The parameter, p_{noise}, controls the amount of noise. We compare the clustering performance of ACA with the GMM method proposed by Barbic *et al*. [6] based on the confusion matrix,

$$\mathbf{C}(c_1, c_2) = \sum_{i=1}^{m_{alg}}\sum_{j=1}^{m_{truth}} g_{c_1 i}^{alg} g_{c_2 j}^{truth} |\mathbf{Y}_i^{alg} \cap \mathbf{Y}_j^{truth}|, \tag{9}$$

where $C(c_1, c_2)$ represents the total number of frames on the cluster segment c_1 that are shared by the cluster segment c_2 in ground truth. Fig. 2 shows the accuracy of ACA and GMM.

4.2 Video

In the second experiment we demonstrate the ability of ACA to robustly segment a video sequence of several humans performing different actions on two databases, Weizmann [13] and KTH [14].

The set of testing videos is synthesized by concatenating the clips that are randomly selected from these two databases. We generate 50 testing videos for both Weizmann and KTH dataset. Each of the videos contains 10-20 clips of different actions. The similarity matrix is constructed by using the silhouette-based (for Weizmann) and flow-based features (for KTH) respectively. After that, we run ACA and GMM 10 times, one for each concatenated video. We set the length constraint of the actions $w_{\max} = 24$ for Weizmann and $w_{\max} = 16$ for KTH. These parameters are selected based on the video. The obtained by ACA and GMM are shown in Fig. 3.

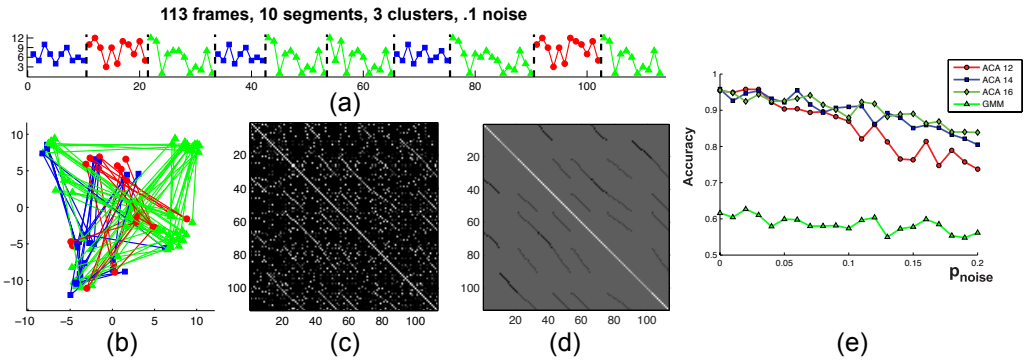
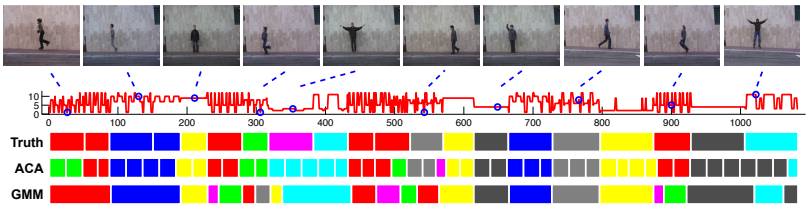


Figure 2: Segmentation of synthetic data. (a)-(d) An example. (a) Time series as a string. (b) Time series in 2-D space. (c) The Gram matrix $\mathbf{K}$. (d) $\mathbf{L} \circ \mathbf{W}$. (e) Accuracy.



	Weizmann	KTH
ACA	**0.76**	**0.80**
GMM	0.63	0.72

(a) (b)

Figure 3: Segmentation of human activities. (a) An example. (b) Accuracy.

References

[1] G. Guerra-Filho and Y. Aloimonos. Understanding visuo-motor primitives for motion synthesis and analysis. *Comp. Anim. Virtual Worlds*, 17:207–217, 2006.

[2] C. Lu and N. J. Ferrier. Repetitive motion analysis: Segmentation and event classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):258–263, 2004.

[3] 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.

[4] X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *ICML*, pages 1055–1062, 2007.

[5] Z. Harchaoui and O. Cappé. Retrospective mutiple change-point estimation with kernels. In *IEEE Workshop on SSP*, pages 768–772, 2007.

[6] J. Barbic, A. Safonova, J. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard. Segmenting motion capture data into distinct behaviors. In *Graphics Interface*, pages 185–194, 2004.

[7] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, pages 123–130, 2001.

[8] F. de la Torre, J. Campoy, Z. Ambadar, and J. F. Cohn. Temporal segmentation of facial behavior. In *ICCV*, pages 1–8, 2007.

[9] D. Minnen, C. Lee Isbell, I. A. Essa, and T. Starner. Discovering multivariate motifs using subsequence density estimation and greedy mixture learning. In *AAAI*, pages 615–620, 2007.

[10] F. Zhou, F. de la Torre, and J. K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *FGR*, pages 1–7, 2008.

[11] H. Shimodaira, K.-I. Noma, M. Nakai, and S. Sagayama. Dynamic time-alignment kernel in support vector machine. In *NIPS*, pages 921–928, 2001.

[12] M. Cuturi, J.-P.Vert, O. Birkenes, and T. Matsui. A kernel for time series based on global alignments. In *ICASSP*, pages 15–20, 2007.

[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2247–2253, 2007.

[14] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, pages 32–36, 2004.