Automatic Human Motion Segmentation

Vikramjit Singh Sidhu

Motion Synthesis for Virtual Characters
Advised by: Somayeh Hosseini
Supervised by: Klaus Fischer

Previously on MSVC

- So far we have examined motion synthesis approaches directly applied on data
  - Motion Graphs
  - Controllers
  - …
- Today we tackle motion synthesis from another angle

Problem

- How can we manipulate the input data to make motion synthesis easier?

- We know that motion primitives make synthesis easier

- Can we divide the data into motion primitives in an unsupervised manner?

- Can we operate on easily available data (e.g. video) to make synthesis easier?

Approaches

- Clustering Based Method
  - Hierarchical Aligned Cluster Analysis - Zhou et. al. 2013
  - Aligned Cluster Analysis - Zhou et. al. 2008
- Graph Based Method
  - Efficient Unsupervised Temporal Segmentation of Human Motion - Vögele et. al. 2014
- PCA Based Method
  - Complex Non-Rigid Motion 3D Reconstruction by Union of Subspaces - Zhu et. al. 2014

HACA

- Zhou, Feng, Fernando De la Torre, and Jessica K. Hodgins. "Hierarchical aligned cluster analysis for temporal clustering of human motion." *IEEE Transactions on Pattern Analysis and Machine Intelligence*

HACA - Overview

- Formulates the problem as one of temporal clustering via a variation of Kernel K-means
- Various tools are introduced to get an energy function
- Optimization done via a mixture of Dynamic Programming and Co-ordinate descent
- Can also temporally cluster video data

HACA - Agenda

1. Kernel K-means formulation
2. Aligned Cluster Analysis
    1. Frame Kernel Matrix
    2. Dynamic Time Alignment Kernel
    3. Energy function
3. Sketch of Optimization
4. Hierarchical Aligned Cluster Analysis
5. Experiments and Results

1. Kernel K-Means

$$J_{KM} = \sum_{c=1}^{k} \sum_{i=1}^{n} g_{ci} \|\phi(\mathbf{x_i}) - \mathbf{z}_c\|^2$$

- Assigns ’n’ data points into ‘k’ clusters
- The kernel transforms the data into another space to make clustering easier
- Cannot be directly applied to our problem

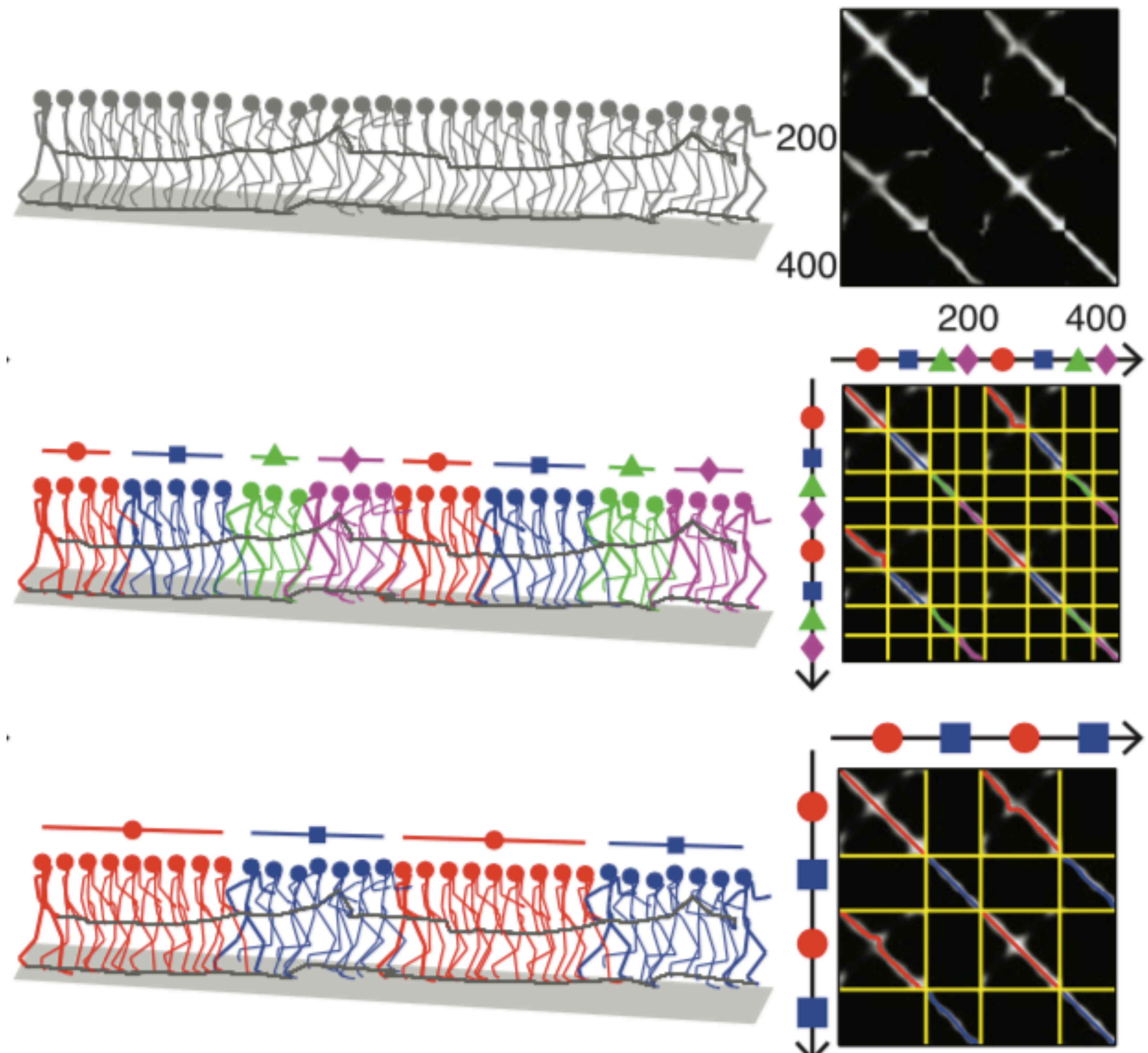
2.1 Frame Kernel Matrix

$$\mathbf{K} = \phi(\mathbf{X})^T \phi(\mathbf{X})$$

$$\mathbf{X} \in R^{d \times n}, \mathbf{K} \in R^{n \times n}$$

- Defines the similarity between the individual frames of the time series
- Gaussian Kernel is usually used
- Its structure reveals information about the dynamics of the motion which we look to exploit
- Parameter required for period ambiguity, n_{max}

2.1 Frame Kernel Matrix Visualisation

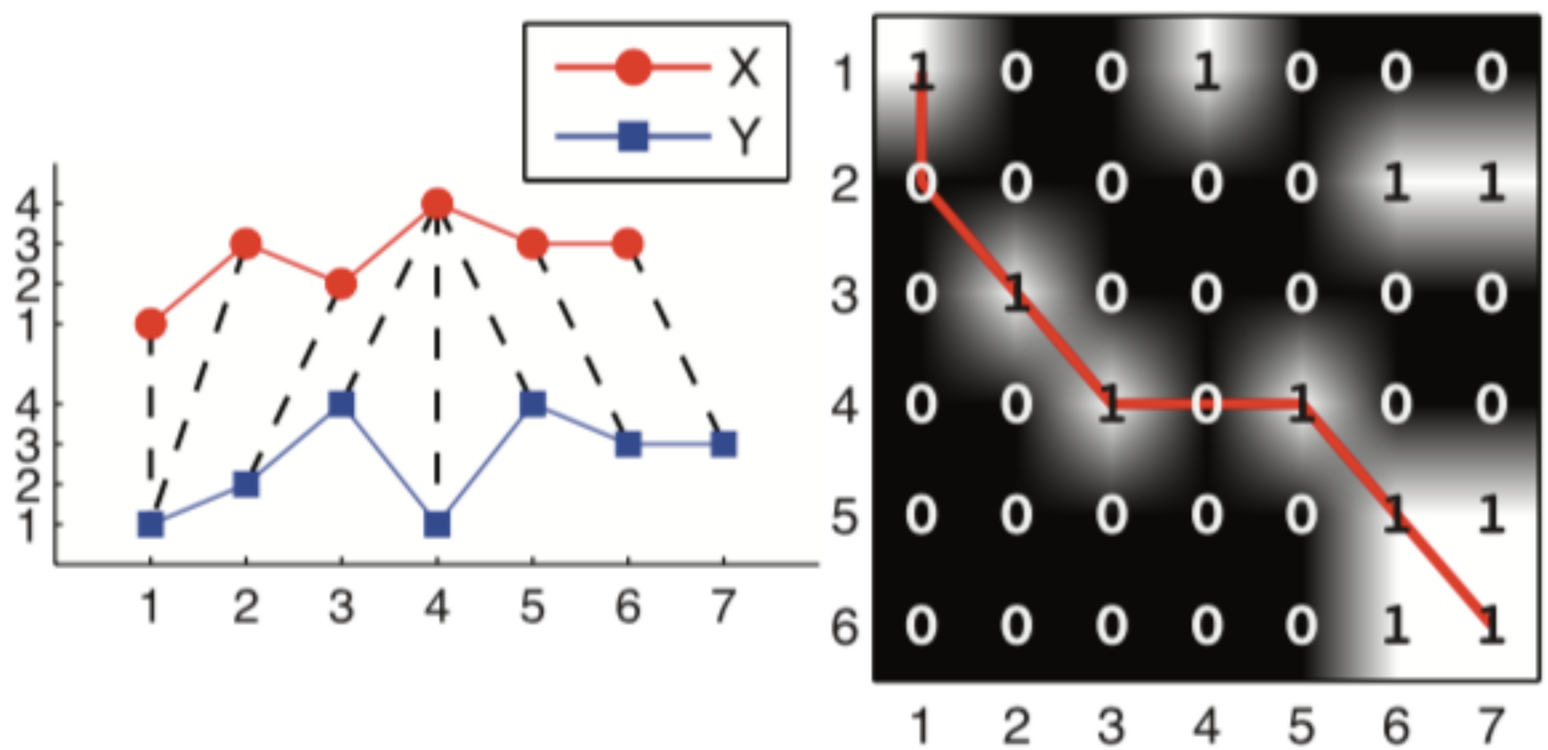


2.2 Dynamic Time Alignment Kernel (DTAK)

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

- We need a way to calculate the distance between data segments of different temporal lengths
- Dynamic Time Warping (DTW) is a concept from data mining, it aligns the two time series data
- It is invariant to temporal distortions, for example to the speed of human action
- Computed in a recursive fashion

2.2 DTAK Visualization



2.3 ACA Energy Function

$$J_{ACA}(\mathbf{G}, \mathbf{s}) = \sum_{c=1}^{k} \sum_{i=1}^{m} g_{ci} \|\psi(\mathbf{X}_{[\mathbf{s_i}, \mathbf{s_{i+1}}]}) - \mathbf{z}_c\|^2$$

$$s.t. \quad \mathbf{G}^T \mathbf{1}_k = \mathbf{1}_m, \quad s_{i+1} - s_i \in [1, n_{max}]$$

- The algorithm can be thought of as assigning samples to segments (**s**), and segments to clusters (**G**)
- Temporal ordering of frames is taken into account

3. Optimization

- Optimizing over **G, s** is NP-Hard
- Solve the problem iteratively

$$\mathbf{G}, \mathbf{s} = \arg\min_{\mathbf{G},\mathbf{s}} J_{ACA}(\mathbf{G}, \mathbf{s}) = \arg\min_{\mathbf{G},\mathbf{s}} \sum_{c=1}^{k} \sum_{i=1}^{m} g_{ci} \|\psi(\mathbf{X}_{[\mathbf{s_i},\mathbf{s_{i+1}}]}) - \mathbf{z}_c\|^2$$

- A brute force search for **s** is infeasible, authors use a DP based solution
- Introduce an auxiliary function to optimize instead

3. Optimization

$$J(v) = \min_{\mathbf{G},\mathbf{s}} J_{ACA}(\mathbf{G}, \mathbf{s})\,|_{\mathbf{X}_{[1:v]}}$$

- The above function satisfies the principle of optimality, i.e. the optimal decomposition of a subsequence $\mathbf{X}_{[1:v]}$ is achieved when subsequences on both sides $\mathbf{X}_{[1:i-1]}$ $\mathbf{X}_{[i:v]}$ are optimal

- This allows us to minimize using Bellman's equation

$$J(v) = \min_{v-n_{max}<i<v} \left(J(i-1) + \min_{g} \sum_{c=1}^{k} g_c \|\psi(\mathbf{X}_{[i:v]}) - \mathbf{z}_c\|^2 \right)$$

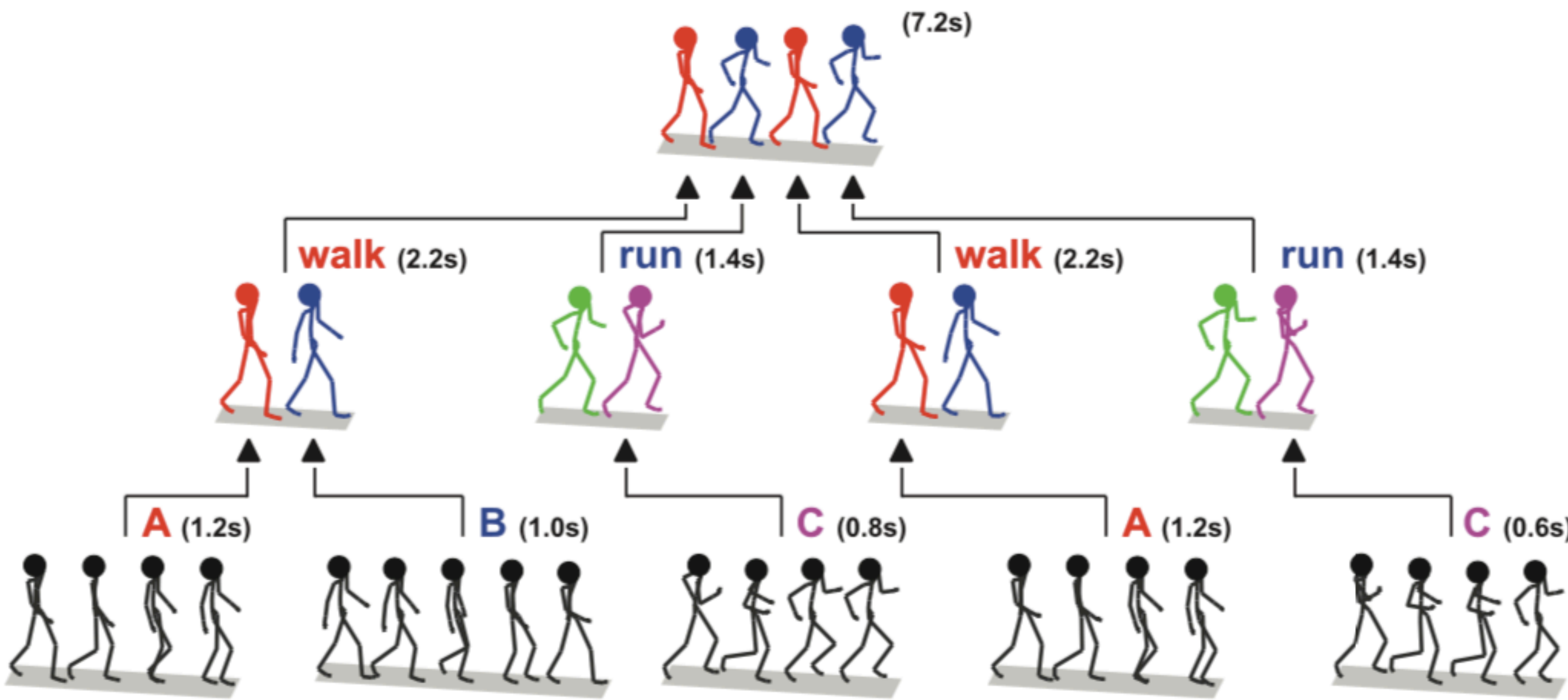
- Computation of ψ is expensive due to its recursive nature

- Due to the formulation, we can compute it part by part by maintaining a list

4. HACA

- Extend ACA to perform hierarchical decomposition of data
- This is done by extending the definition of the DTAK, this propagates solution to multiple levels
- At each hierarchy, frame kernel matrix is computed and ACA is performed with temporal length parameter $n_{max}^{(i)}$
- The temporal length parameter is higher initially and smaller later on [*]

4. HACA



Experiments

- Temporal Segmentation performed on the following data:
  - Synthetic Data
  - CMU Motion Capture data
  - Human Video Data
  - Honey Bees Dancing Video Data

Results - CMU MoCap

Results - KTH


Temporal Segmentation of Human Behavior
www.f-zhou.com
KTH Action Dataset

Results - Weizmann

Discussion

- More specialised kernel for a better embedding
- Pros
  - Robust method, proved via through experimentation
  - Can operate on Video Data
  - Creative solution to a difficult problem
- Cons
  - Computationally impractical to run on large amounts of motion capture data
  - Complicated Formulation
  - Temporal segmentation of video data occurs in 2D

Graph Based Method

- Vögele, Anna, Björn Krüger, and Reinhard Klein. "Efficient unsupervised temporal segmentation of human motion." Proceedings of the ACM SIGGRAPH/ Eurographics Symposium on Computer Animation. Eurographics Association, 2014.

Graph Based Method - Overview

- Creates a special graph called a ‘neighborhood graph’ out of motion capture data

- Formulates the problem of segmentation into activities and finding motion primitives as operations on the graph

- The method is created keeping motion synthesis using Motion Graphs in mind

Graph Based Method - Agenda

1. Neighborhood graph
2. Segmentation into distinct Activities
3. Subdividing Activities into Motion Primitives
4. Motion Synthesis using motion graphs
5. Experiments and Results

1. Creation of Neighborhood Graph

- Input Data:
  - The motion sequence, **M** is represented as a set of ’n’ frames/poses
  - Each frame is represented as a 15-dim vector representing the skeleton
- The motion sequence is organised into a kd-tree
- A search is performed on each pose to find the set of nearest neighbours using the Euclidean distance

1. Creation of Neighborhood Graph

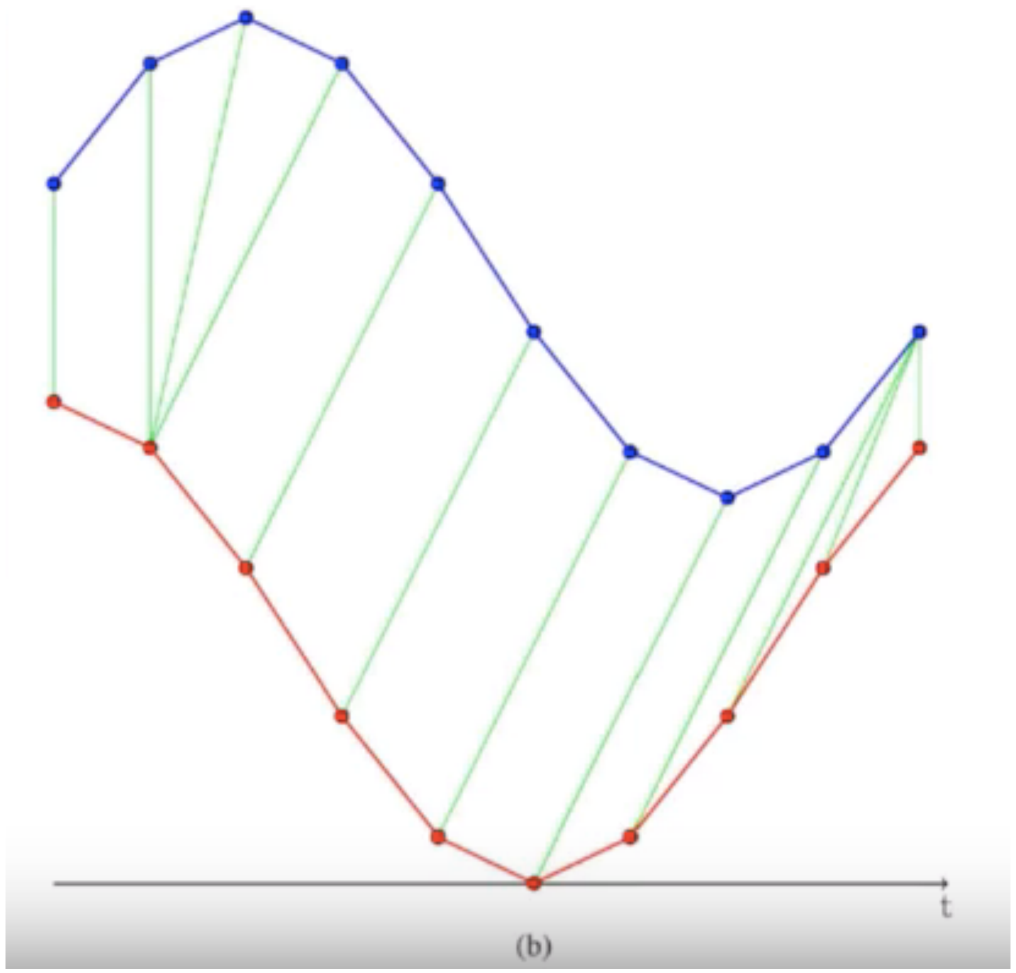
$$S_i = \{p_j\}_{j=1}^{k}$$

$$S = \{S_i, i \in 1,\ldots,n\}$$

- Each of poses in the above sets are nodes of the neighborhood graph

- Edges are added between nodes if they are ‘sufficiently similar’, this is characterised using Dynamic Time Warping (DTW)

1. Creation of Neighborhood Graph



- DTW finds optimal alignment between time series data, used to find optimal connection between poses

Image Courtesy: https://github.com/tkorting/youtube/blob/master/how-dtw-works.m

1. Neighborhood Graph

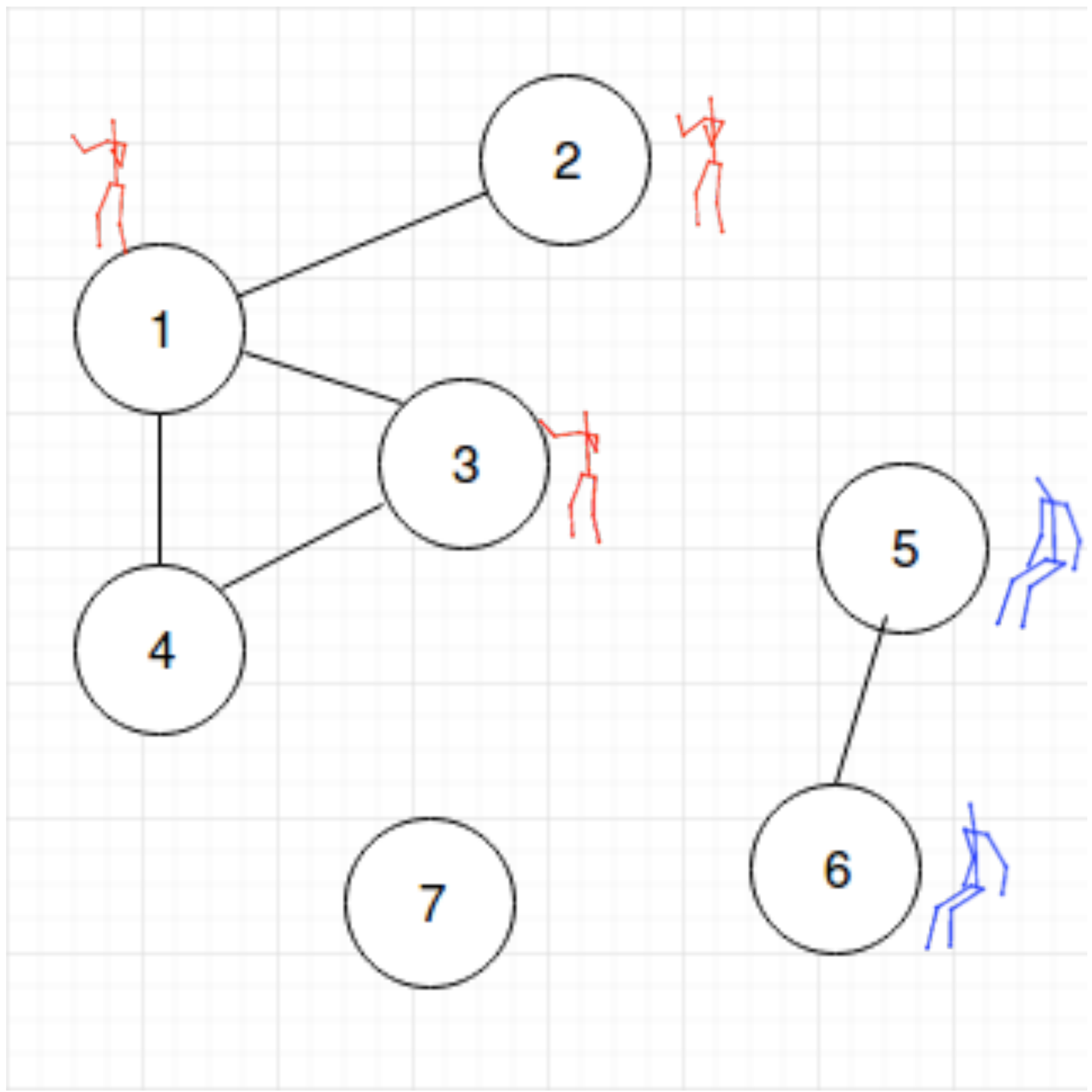
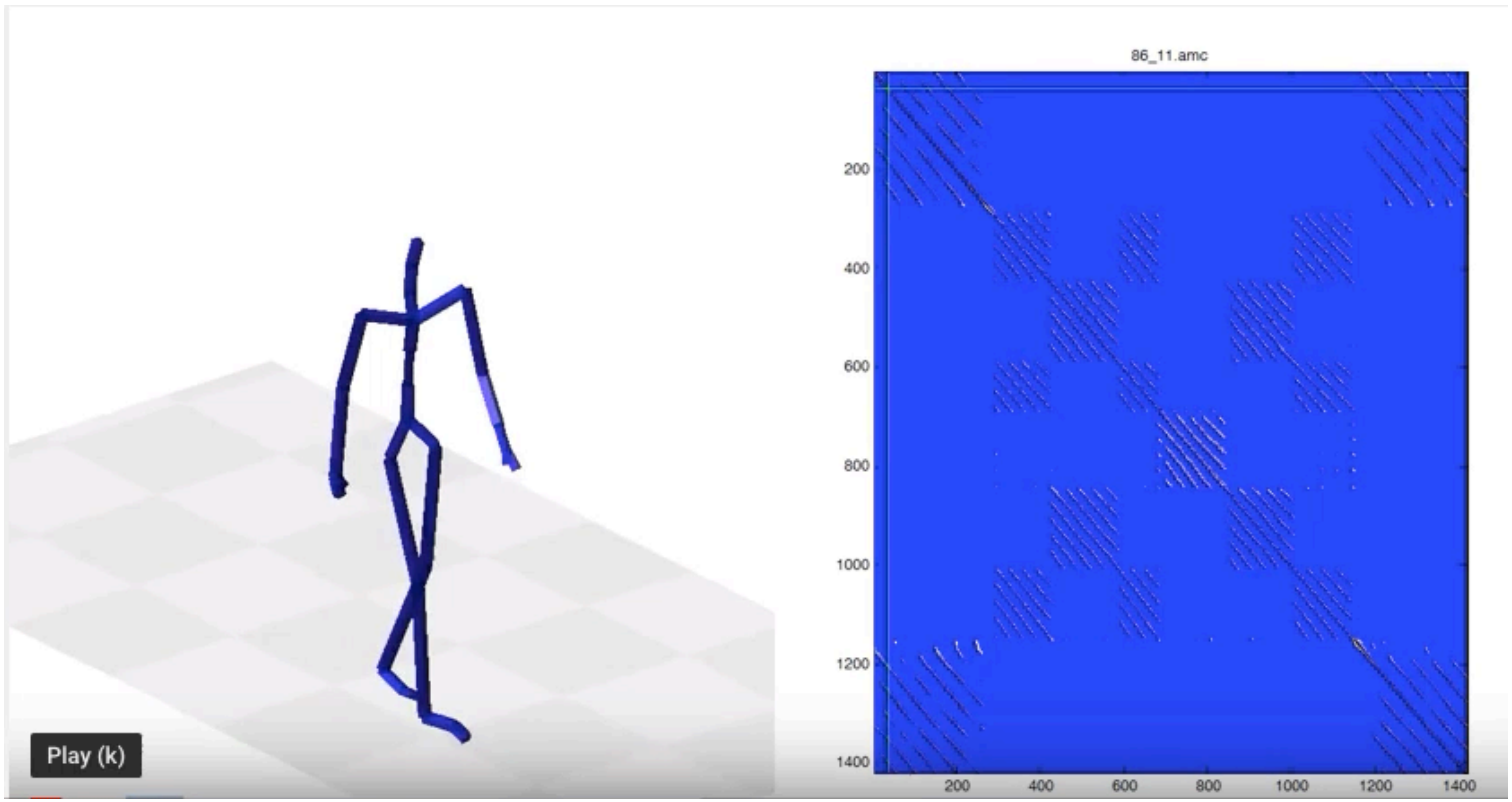


Image Courtesy : http://sleepincode.blogspot.com/2017/07/finding-connected-components-using-dfs.html

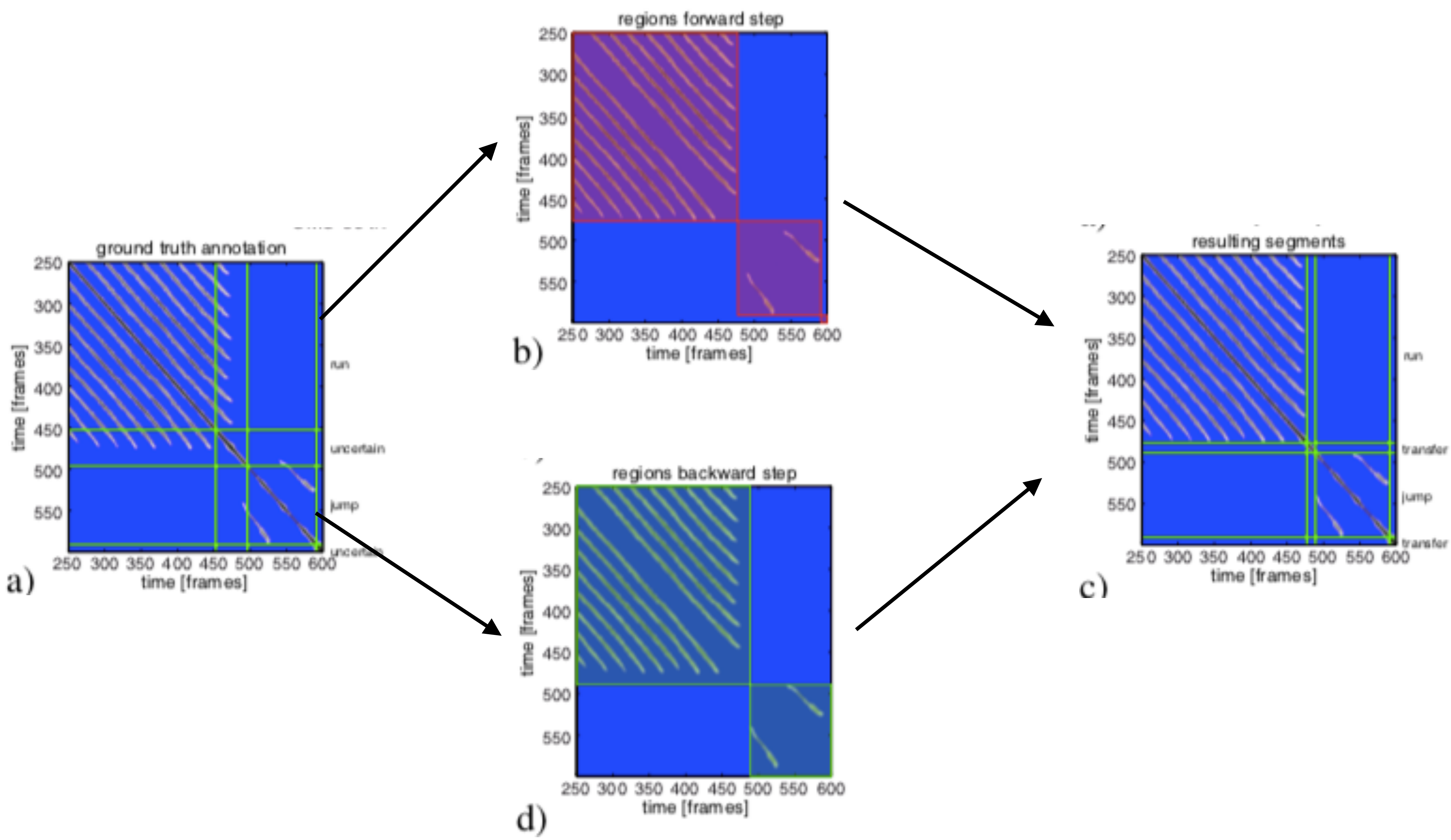
1. Visualization of graph using Similarity Matrices



2. Segmentation into distinct Activities

- Performed using a variation of ‘region growing’ algorithm
- As a preprocessing step, the connected component of the graph corresponding to the seed is removed
- Forward and backward steps performed to identify the transition of the motion
- Implementation in graph via finding connected components [*]

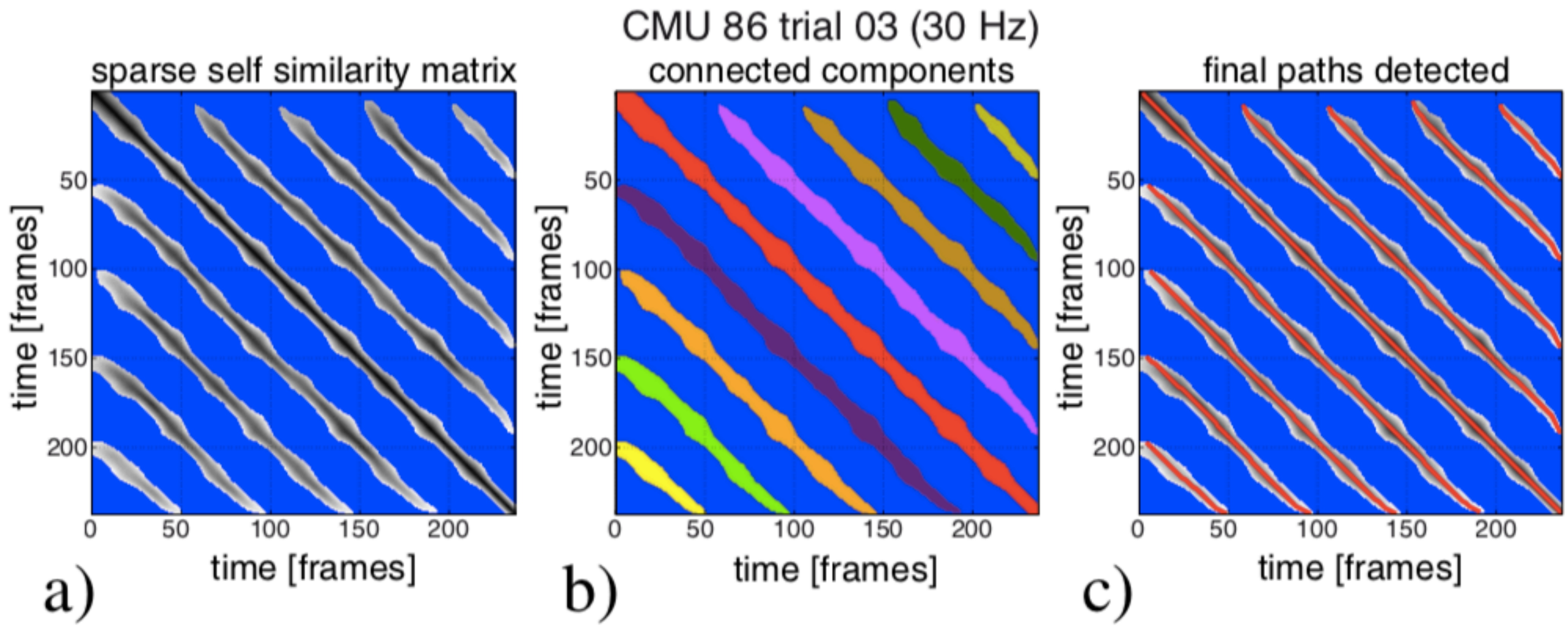
2. Segmentation into distinct Activities



3. Subdividing Activities into Motion Primitives

- Each connected component represents a particular activity
- Most activities contain motion primitives which can be combined to obtain the activity
- Corresponds to finding minor diagonals in self-similarity matrix, these are basically the minimum cost warping paths

3. Subdividing Activities into Motion Primitives



- In the graph, implemented as follows [*]:
  - Isolate the connected components corresponding to the activity
  - Find the shortest path between the pair of nodes

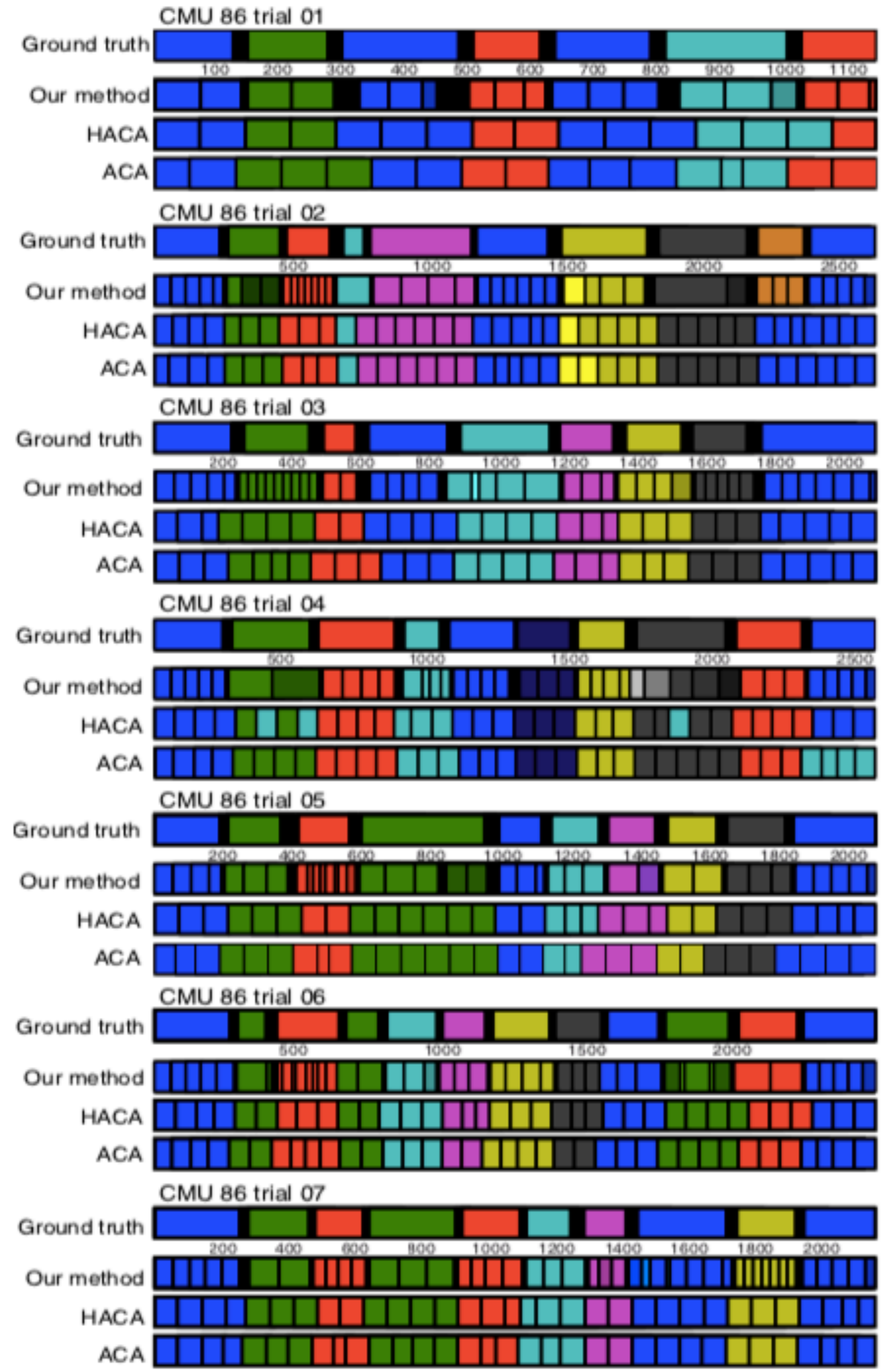
4. Motion Synthesis

- A motion graph is created from the cluster of motion primitives
- The authors claim that their segmentation and clustering results in superior motion synthesis
- More motion primitives result in more possible transitions, e.g. for CMU subject 86, 9 primitives corresponding to ‘running’ found

Experiments

- CMU Motion Capture Database
- Computation Time
- Checking that the method gives Low intra-cluster variance
- Label Transfer Problem

Results - CMU MoCap



Discussion

- Pros
  - Graph based approach allows more manipulations on data
  - Transitions between activities also learnt
  - Comparatively faster
- Cons:
  - No objective function formulation
  - Lack of thorough experimentation, e.g. no video data

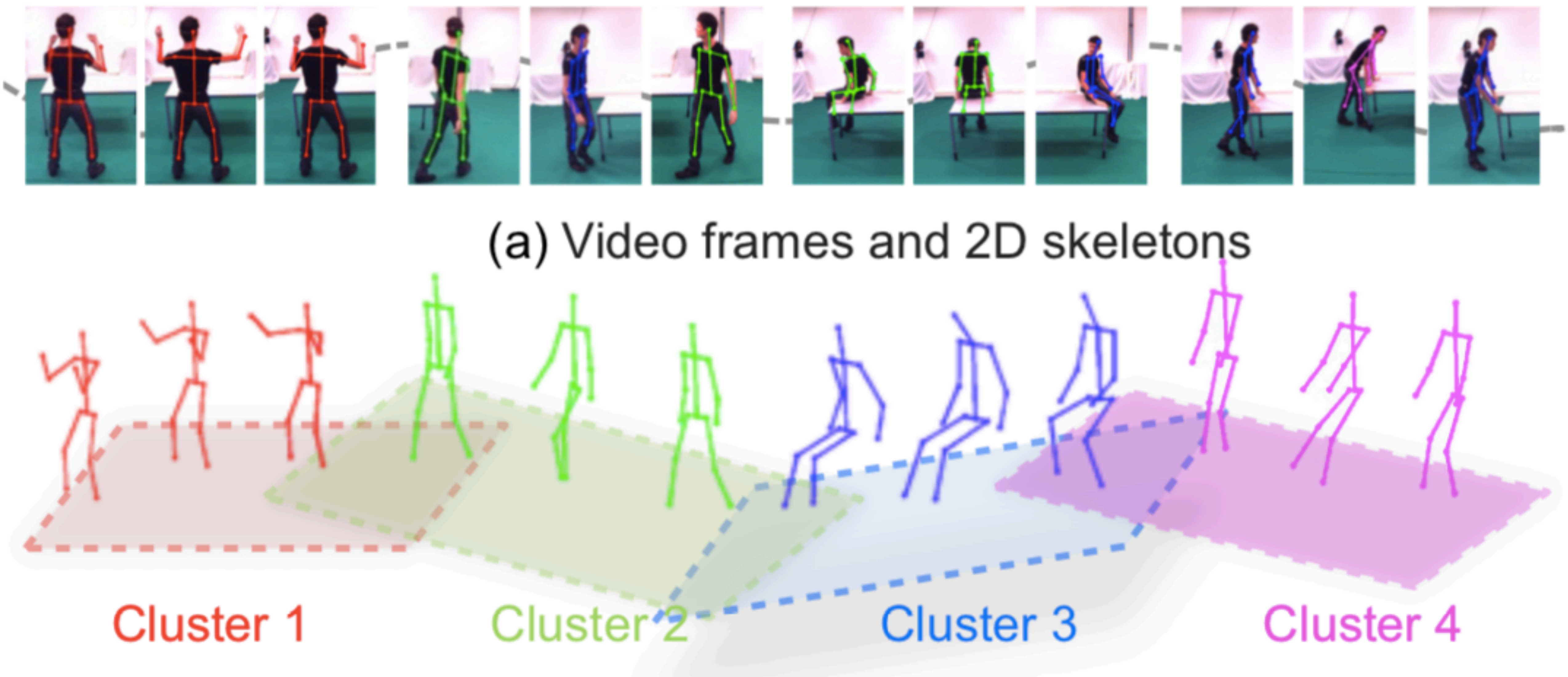
PCA Based Method

- Zhu, Yingying, et al. "Complex non-rigid motion 3d reconstruction by union of subspaces." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

PCA Based Method - Overview

- Input is video data with detected skeletons
- Performs a 3D reconstruction, obtaining the skeleton of the human along with camera matrices
- Additionally learns which cluster each point of the dataset belongs to

PCA Based Method - Overview



Formulation

$$\arg\min_{\mathbf{X},\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \|\mathbf{X}\|_* + \lambda\|\mathbf{E}\|_l$$

$$s.t. \quad X = XZ, \quad W = RX' + E$$

- **Z** - affinity matrix / similarity matrix
- **W** - 2D skeleton
- **X** - 3D skeleton
- **R** - Rotation Matrices
- **E** - noise
- Optimization done using Augmented Lagrangian Methods (ALMs)
- Basically it is the Lagrangian function with some additional terms to handle constraints

Discussion

- Operates on video
- Directly learns the similarity matrix **Z** instead of using indirect approaches
- Elegant formulation

References

- Zhou, Feng, Fernando De la Torre, and Jessica K. Hodgins. "Aligned cluster analysis for temporal segmentation of human motion." 2008 8th IEEE international conference on automatic face & gesture recognition. IEEE, 2008.
- Zhou, Feng, Fernando De la Torre, and Jessica K. Hodgins. "Hierarchical aligned cluster analysis for temporal clustering of human motion." *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Vögele, Anna, Björn Krüger, and Reinhard Klein. "Efficient unsupervised temporal segmentation of human motion." Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation. Eurographics Association, 2014.
- Zhu, Yingying, et al. "Complex non-rigid motion 3d reconstruction by union of subspaces." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

Thanks! Questions?