
Render the Possibilities
SIGGRAPH2016
THE 43RD INTERNATIONAL
CONFERENCE AND EXHIBITION ON
&
Computer Graphics
Interactive Techniques
24-28 JULY
ANAHEIM, CALIFORNIA

Render the Possibilities
SIGGRAPH2016
THE 43RD INTERNATIONAL CONFERENCE AND EXHIBITION ON Computer Graphics & Interactive Techniques

A Deep Learning Framework for Character Motion Synthesis and Editing

Daniel Holden *, Jun Saito †, Taku Komura *,

*The University of Edinburgh
†Marza Animation Planet

Outline

Motivation

Synthesis

Editing

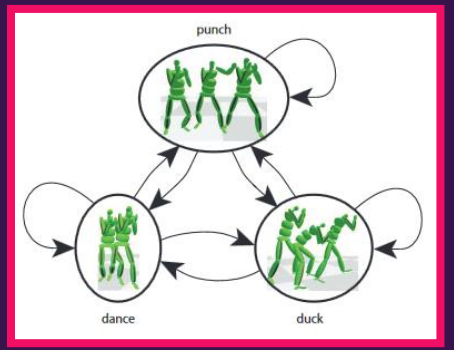
Discussion

Goal

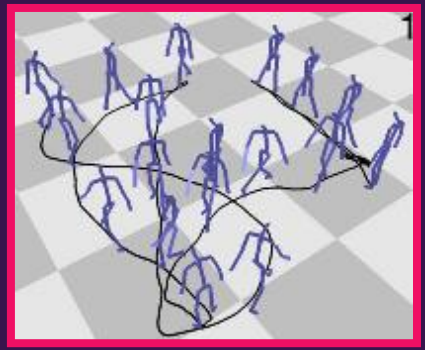
Data driven synthesis of motion
from high level controls with
no manual preprocessing

Previous Work

- Lots of manual processing (Graphs, Trees)
  - Segmentation
  - Alignment
  - Classification



[Heck et al. 2007]

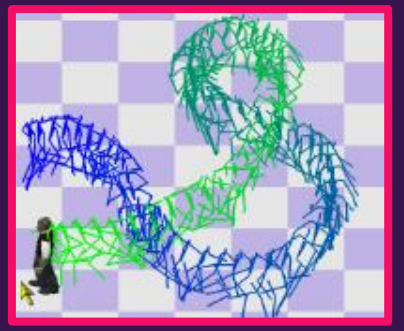


[Kovar et al. 2002]

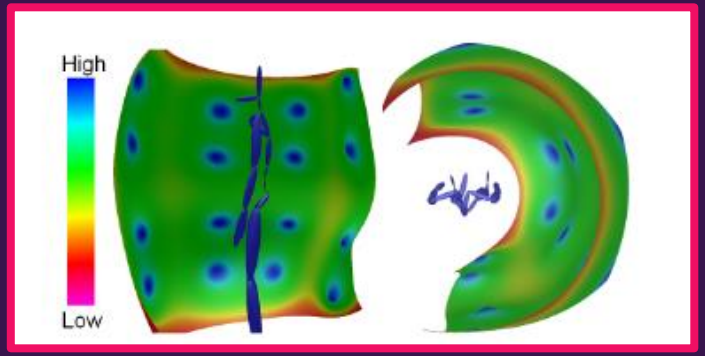
Previous Work

- Scalability Issues (RBF, GP, GPLVM, kNN)
  - Must store whole database in memory
  - Grows $O(n^2)$ with number of data points
  - Requires expensive acceleration structures

[Lee et al. 2010]



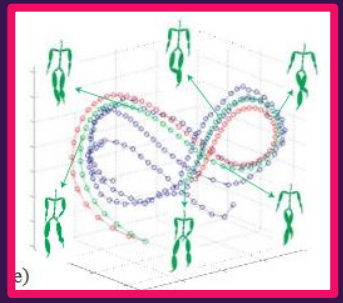
[Park et al. 2002]



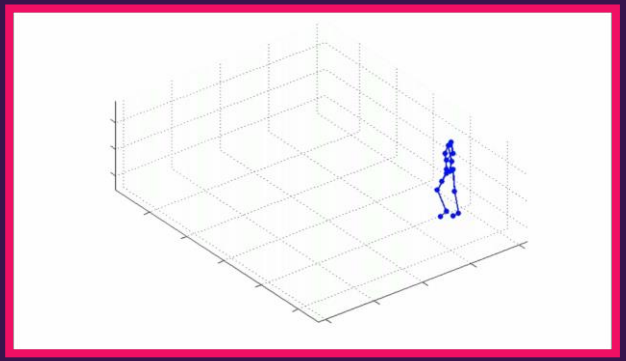
[Mukai and Kuriyama 2005]

Previous Work

- Instability issues (GPDM, CRBM, RNN)
  - Limited to some classes of motion
  - Can suffer high frequency noise or “dying out”



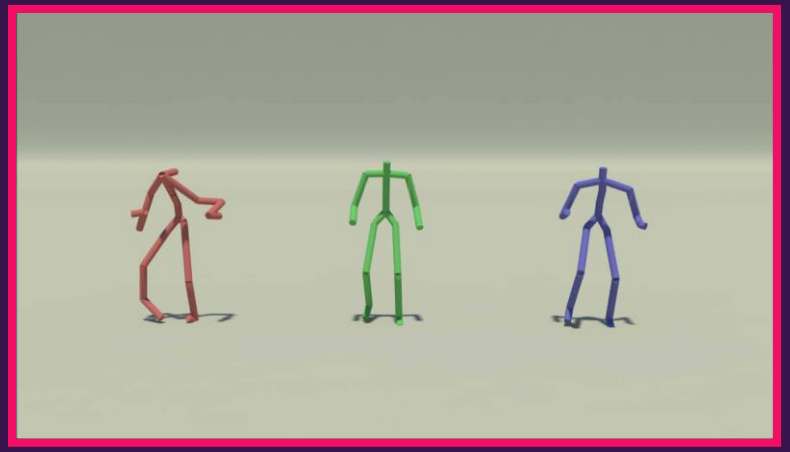
[Levine et al. 2012] [Wang et al. 2005]



[Taylor et al. 2011]

Previous Work

- Deep Neural Network *Hidden Units* used to represent motion
  - Denoising
  - Retrieval
  - Interpolation



[Holden et al. 2015]

Previous Work

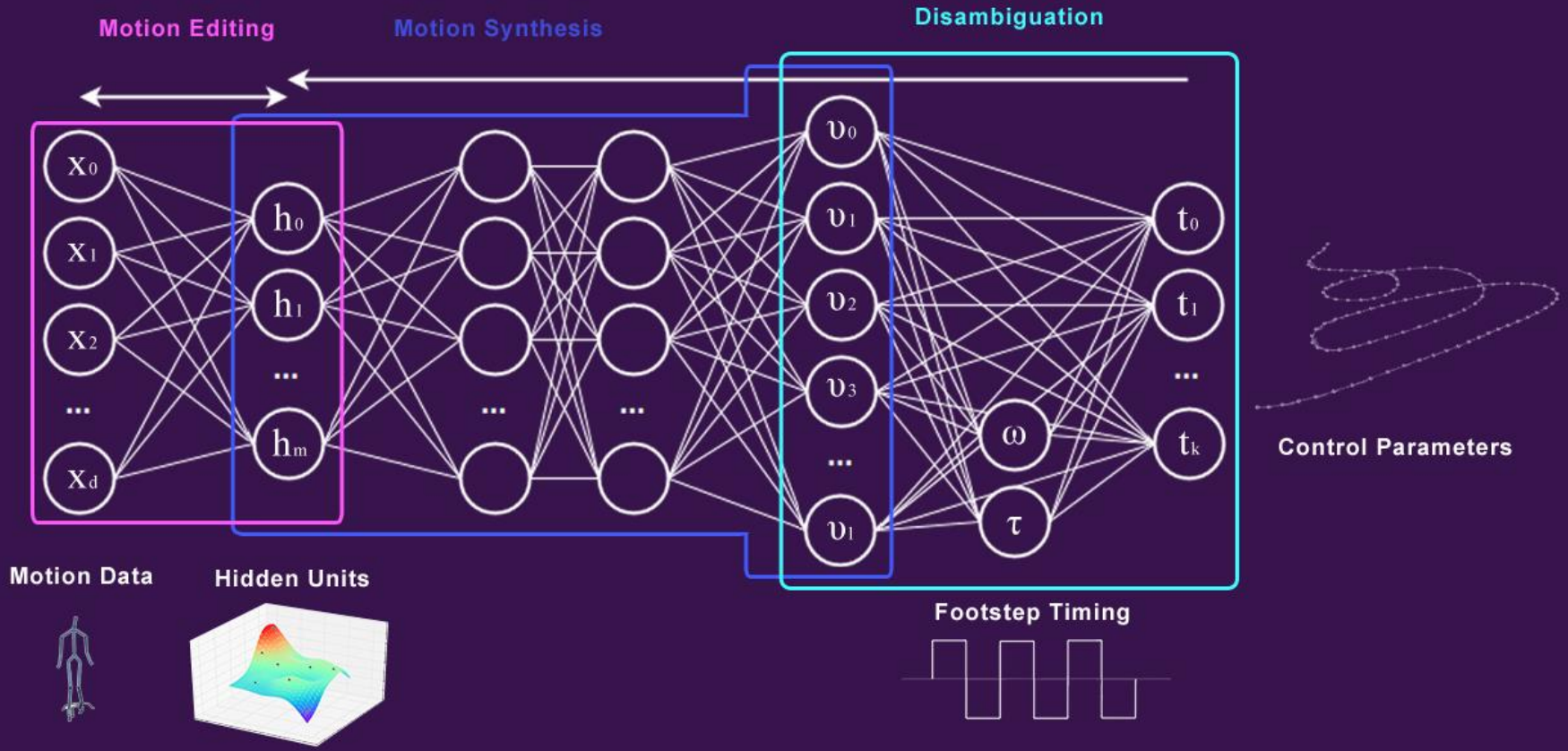
- Deep Learning not always ready for production
- Results can look strange

[Radford et al. 2015]

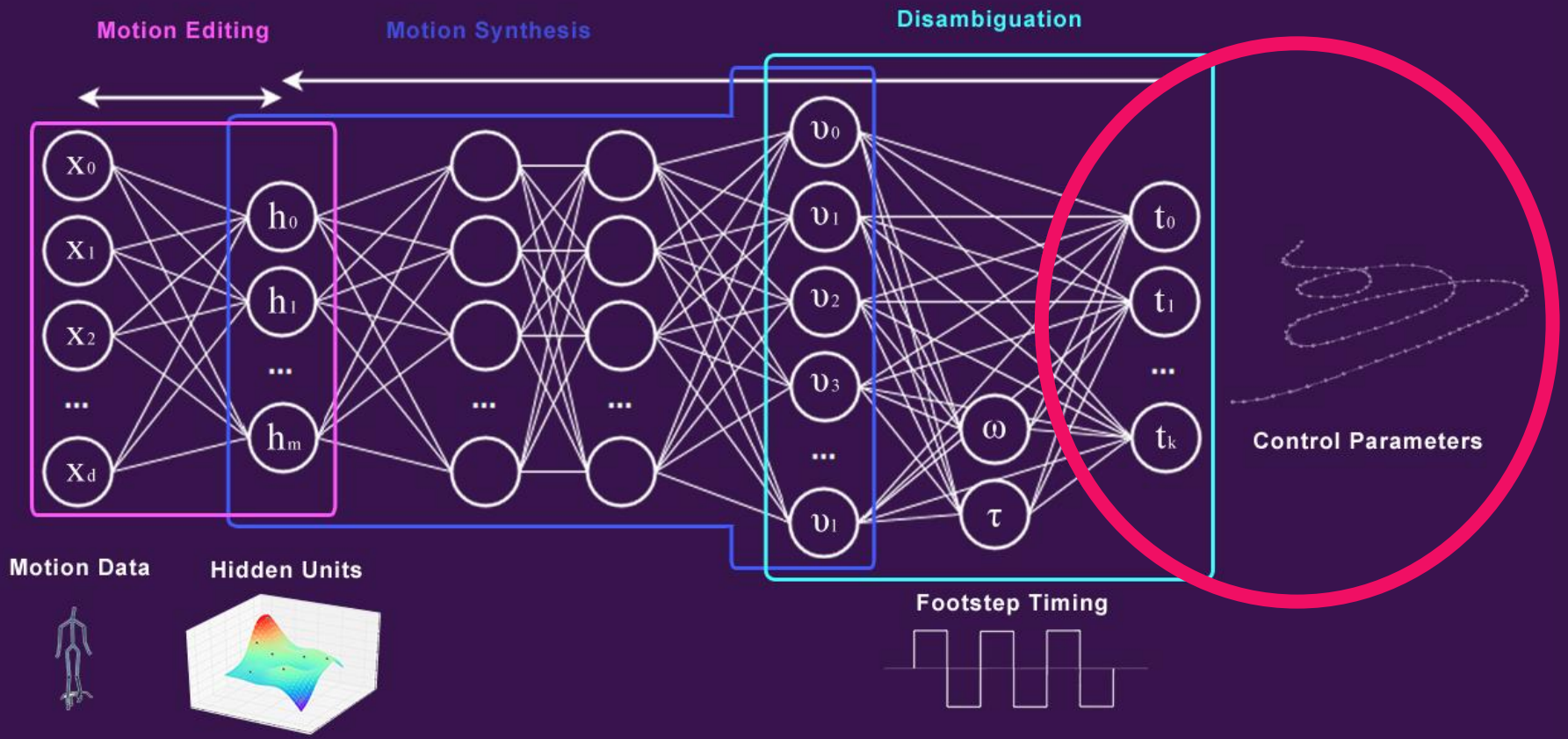
Contribution

- **High quality synthesis** with **no manual preprocessing**
- Motion synthesis and editing in **unified framework**
- **Procedural**, **parallel** technique

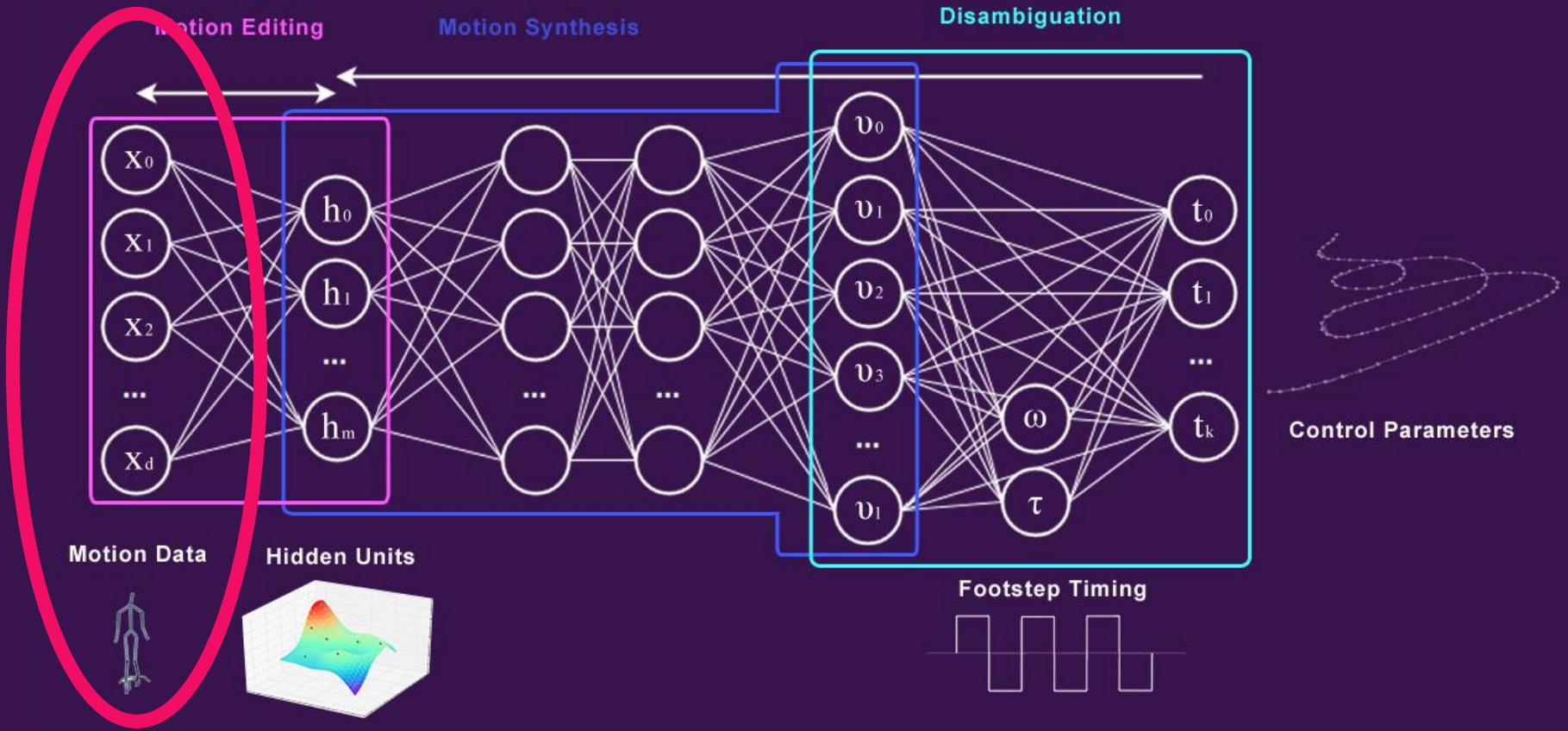
Overview


Motion Editing
Motion Synthesis
Disambiguation
x_0
x_1
x_2
...
x_d
h_0
h_1
...
h_m
...
...
υ_0
υ_1
υ_2
υ_3
...
υ_l
ω
τ
t_0
t_1
...
t_k
Control Parameters
Motion Data
Hidden Units
Footstep Timing

Overview


Motion Editing
Motion Synthesis
Disambiguation
x_0
x_1
x_2
...
x_d
h_0
h_1
...
h_m
...
...
υ_0
υ_1
υ_2
υ_3
...
υ_l
ω
τ
t_0
t_1
...
t_k
Control Parameters
Motion Data
Hidden Units
Footstep Timing

Overview


Motion Editing
Motion Synthesis
Disambiguation
x0
x1
x2
...
xd
h0
h1
...
hm
...
...
υ0
υ1
υ2
υ3
...
υl
ω
τ
t0
t1
...
tk
Control Parameters
Motion Data
Hidden Units
Footstep Timing

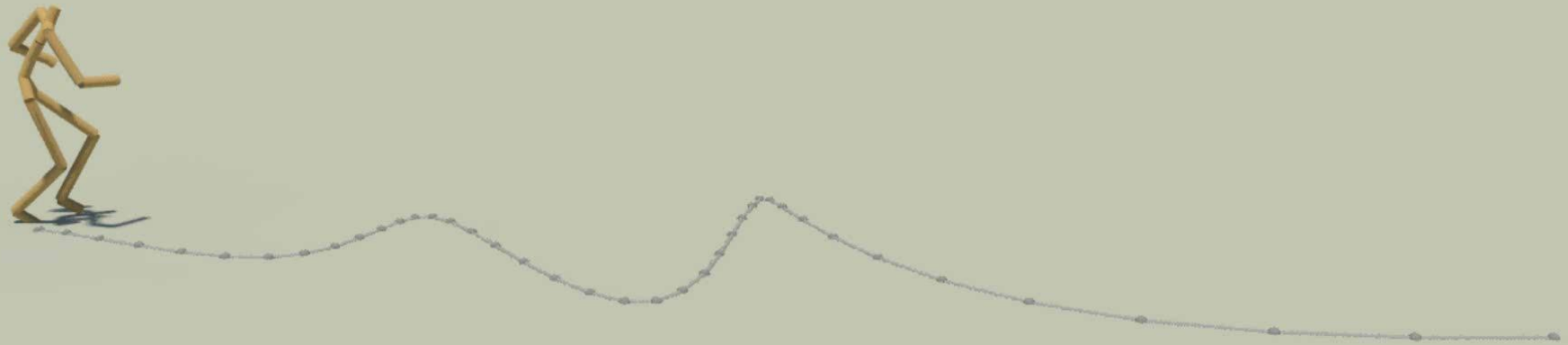
Outline

Motivation

Synthesis

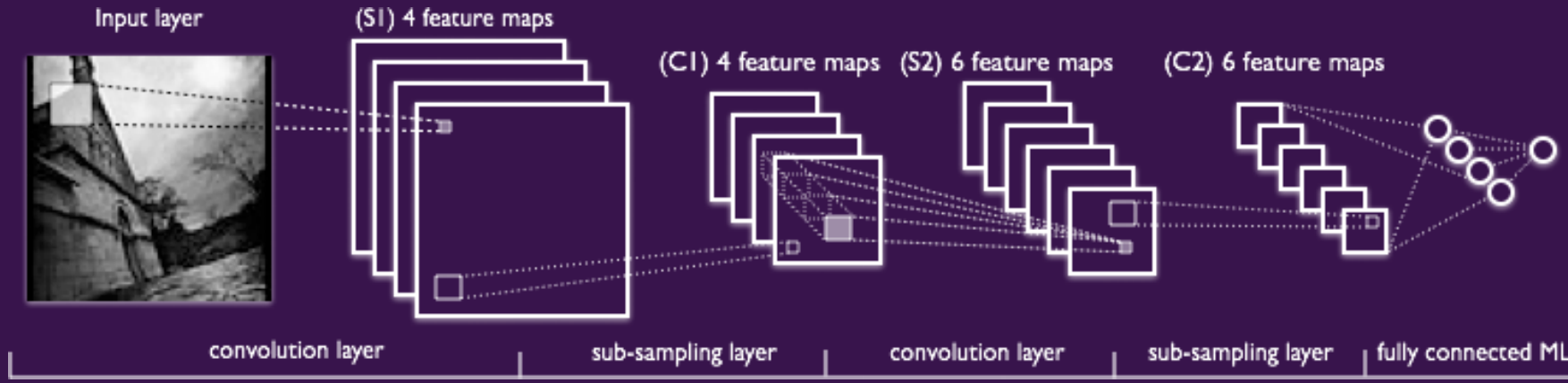
Editing

Discussion



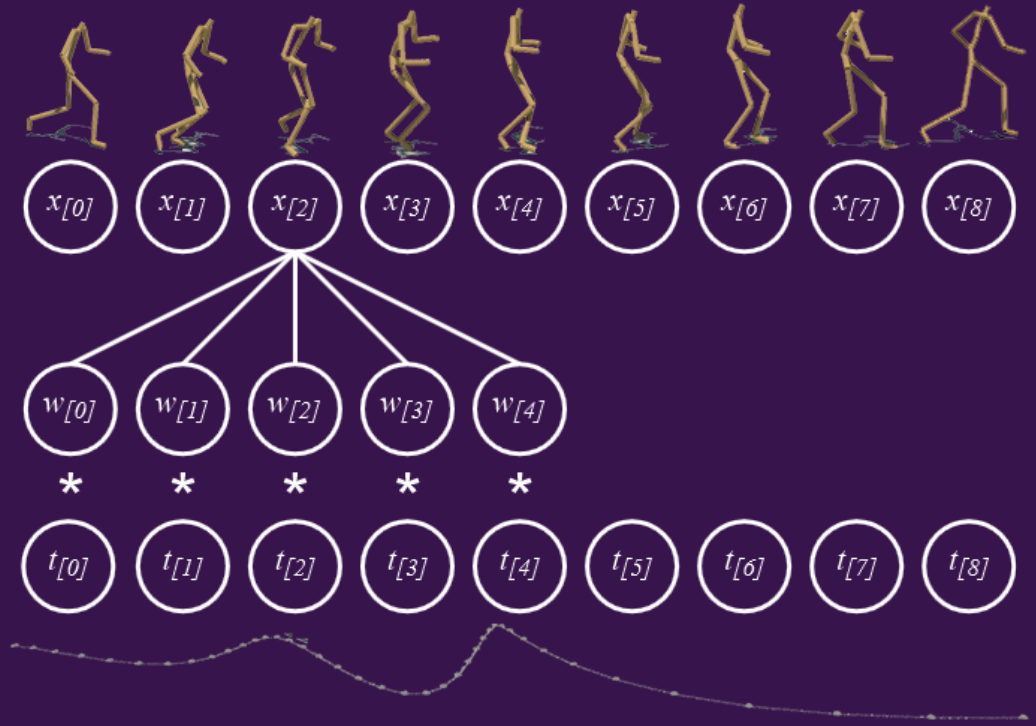
Convolutional Neural Networks

- Great success in classification and segmentation for images, video, sound
- We can use CNN on motion data too



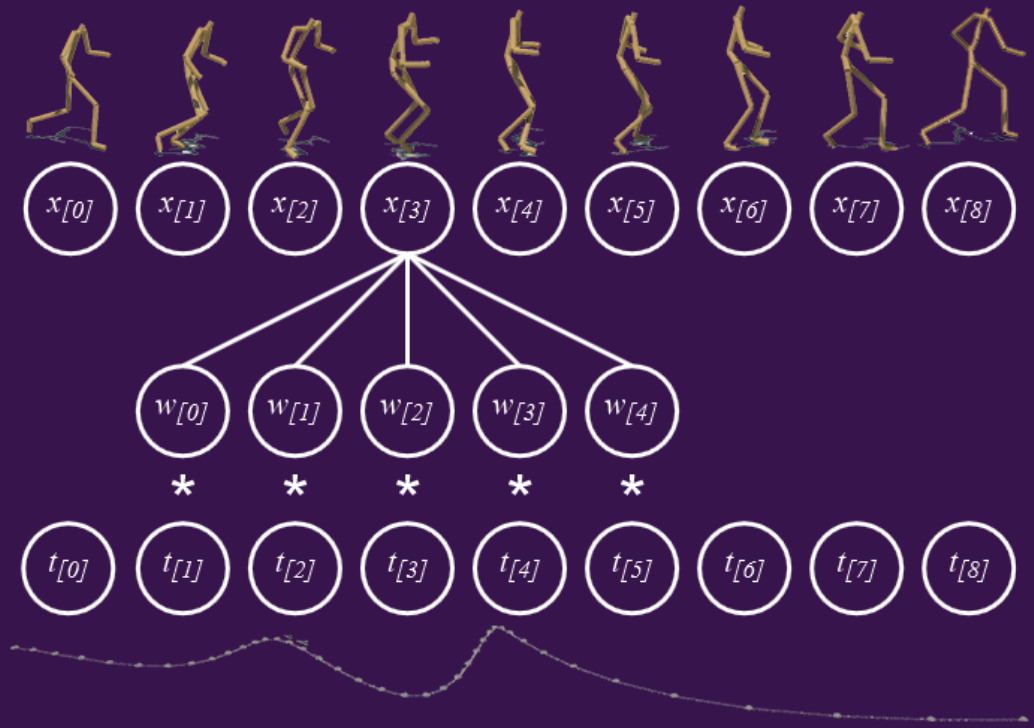
Convolution

Filters convolve over temporal dimension



Convolution

Filters convolve over temporal dimension

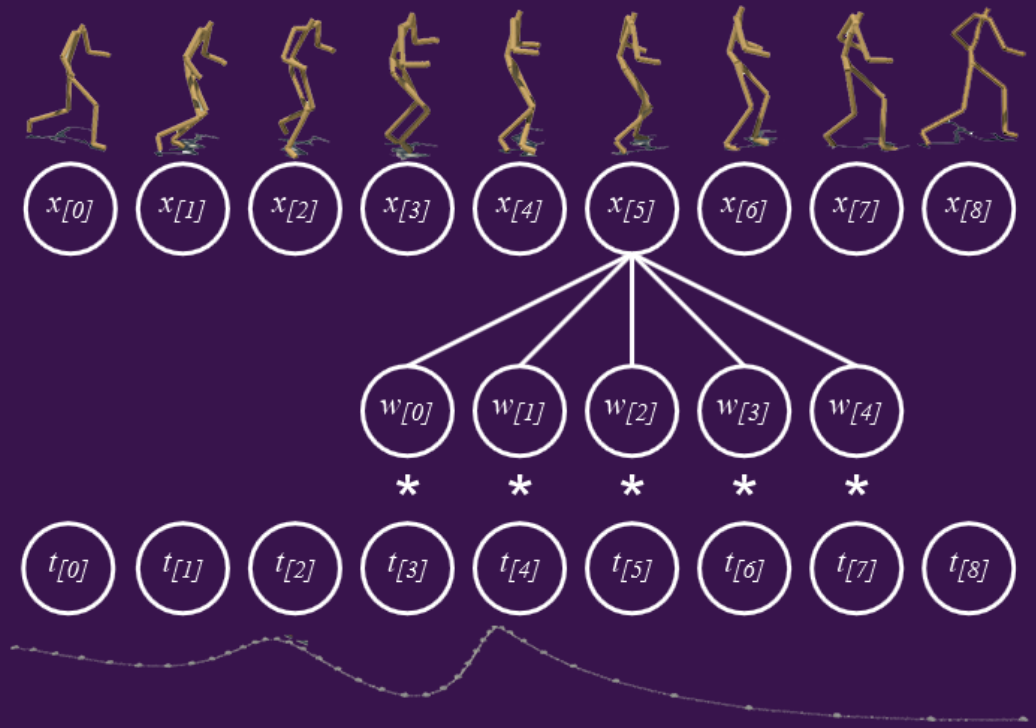


Convolution

Filters convolve over temporal dimension

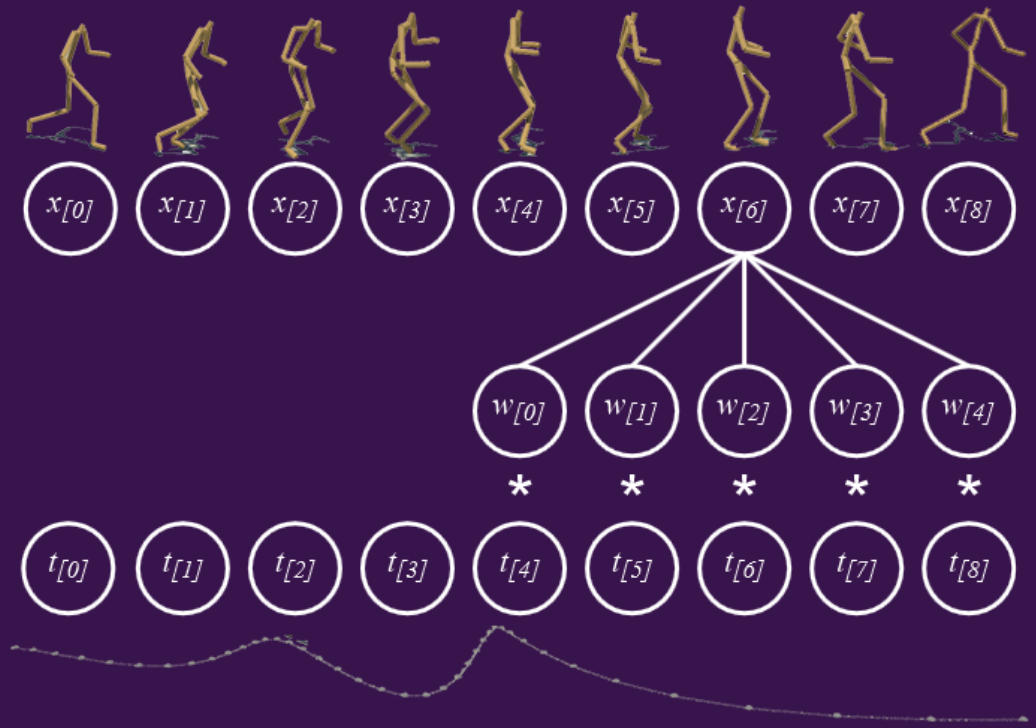
Convolution

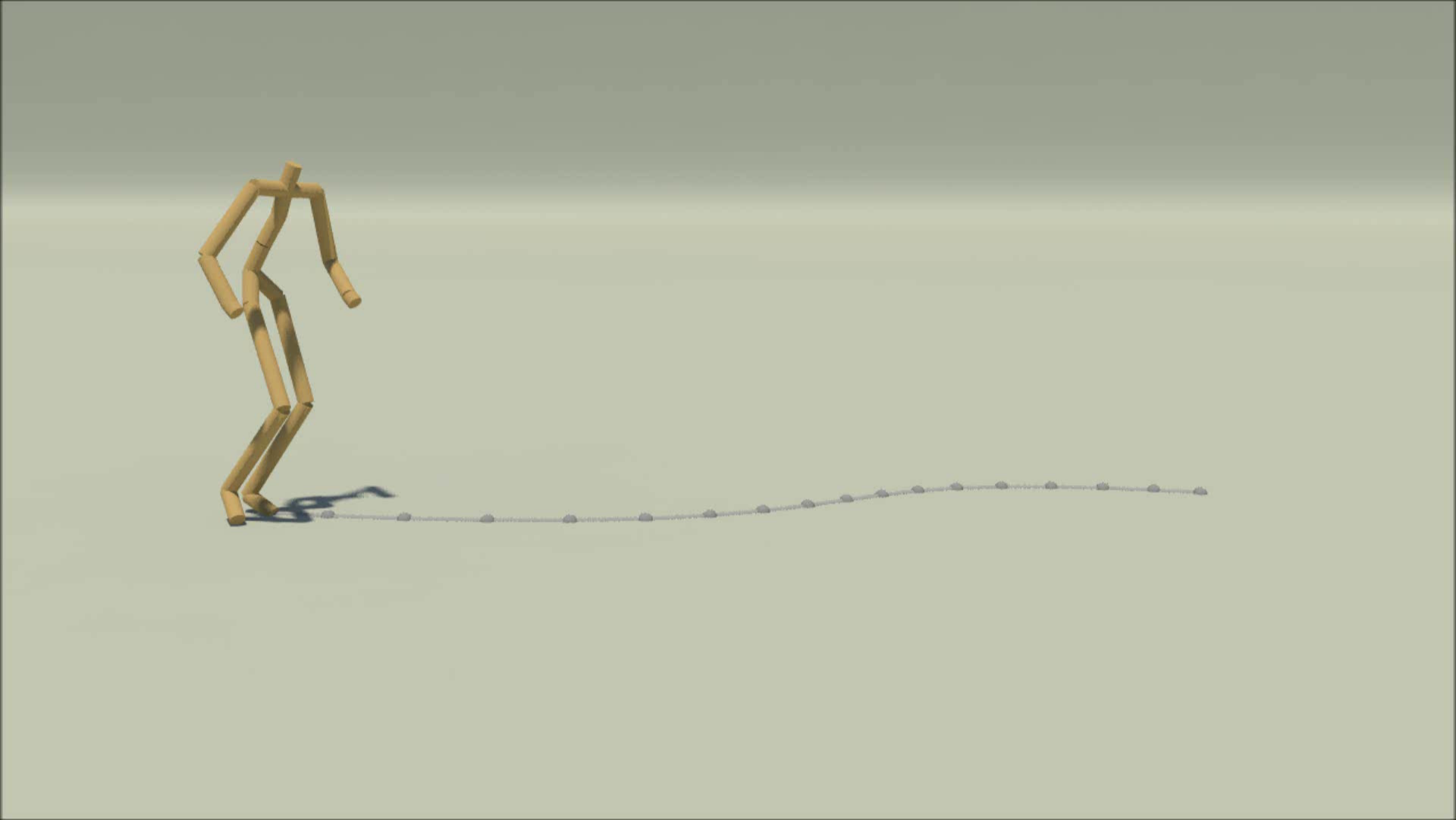
Filters convolve over temporal dimension



Convolution

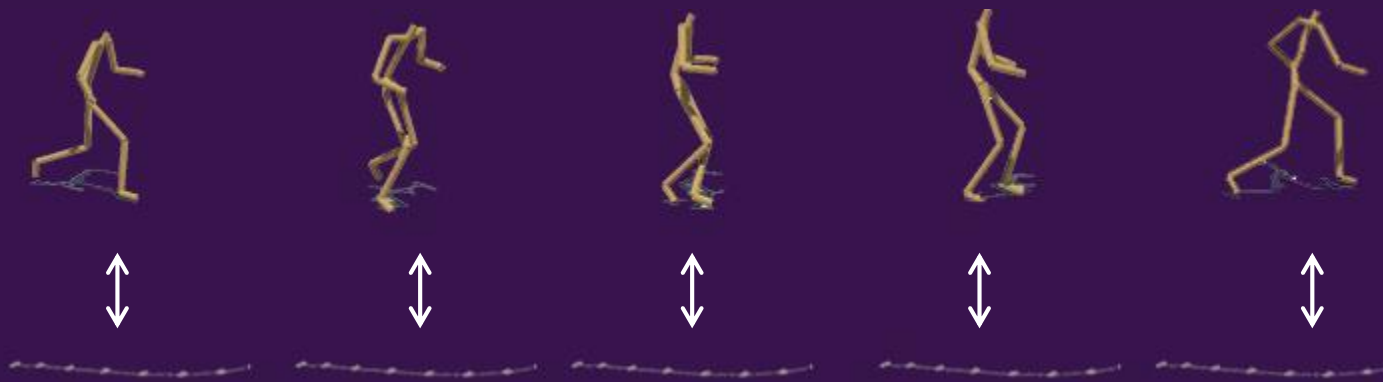
Filters convolve over temporal dimension





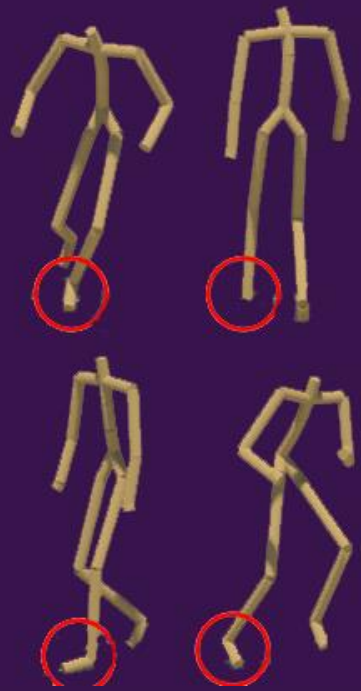
What Happened?

- **Ambiguity**: the same control signal maps to multiple motions
- These motions are averaged in the output



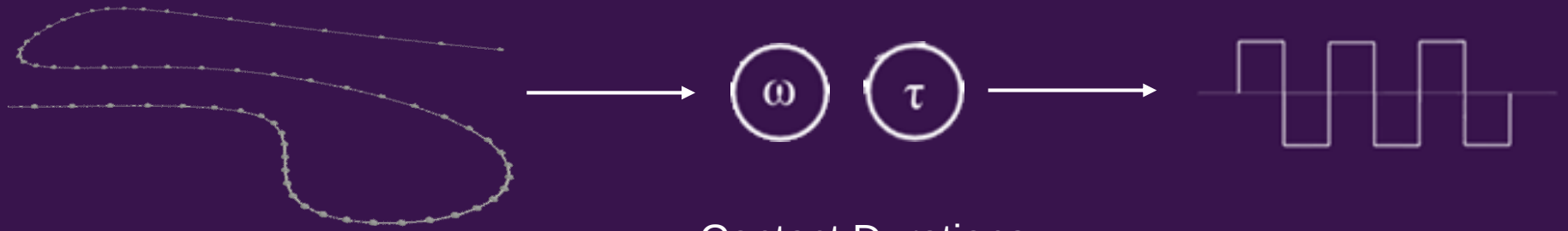
Foot Contact

- Contact times resolve ambiguity
- Automatically label using foot speed and height
- Learn model that generates contact times from trajectory



Foot Contact

- Use small neural network to map trajectories to contact durations and frequencies
- Produce timings from durations and frequencies



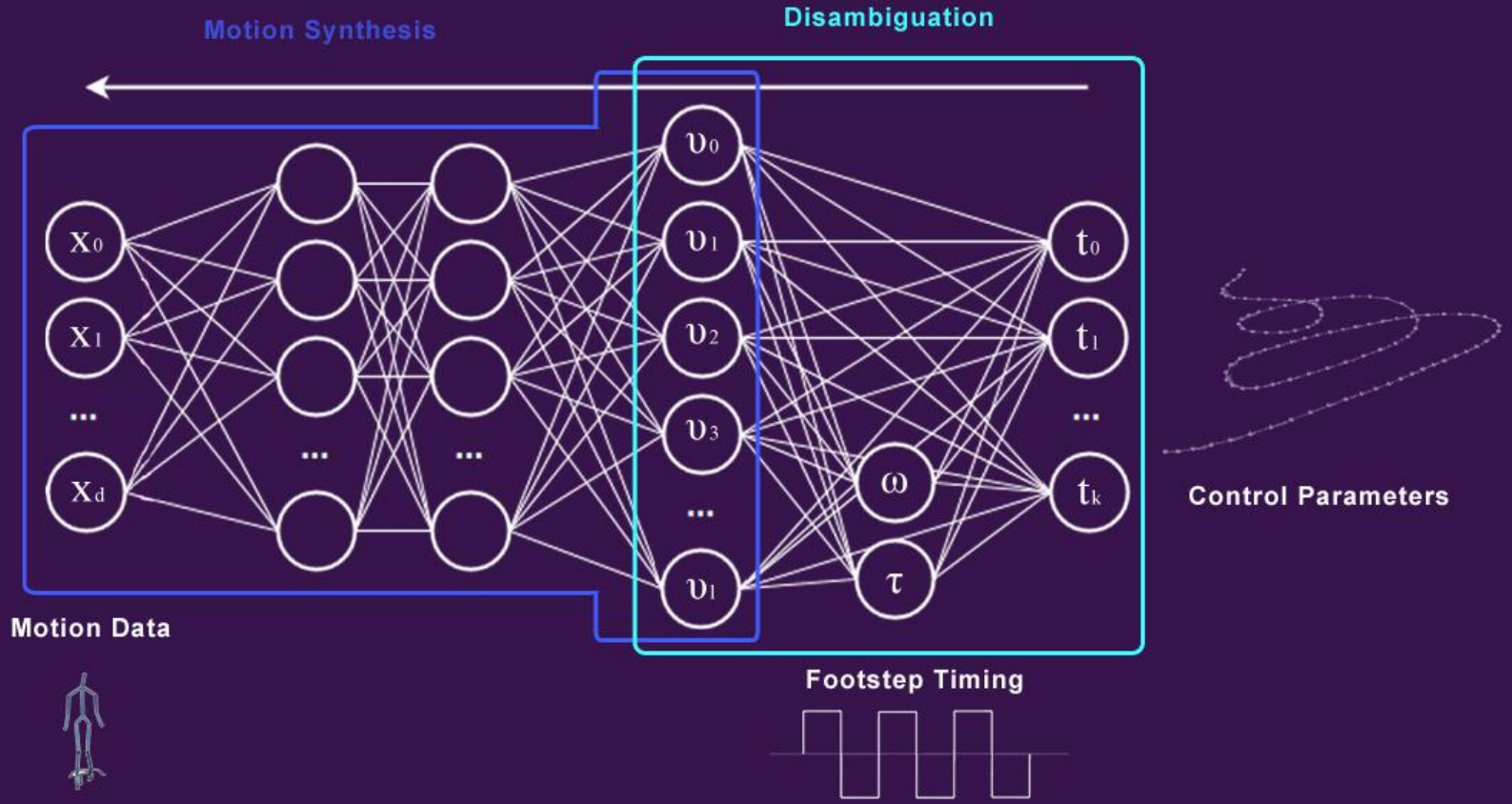
Trajectory

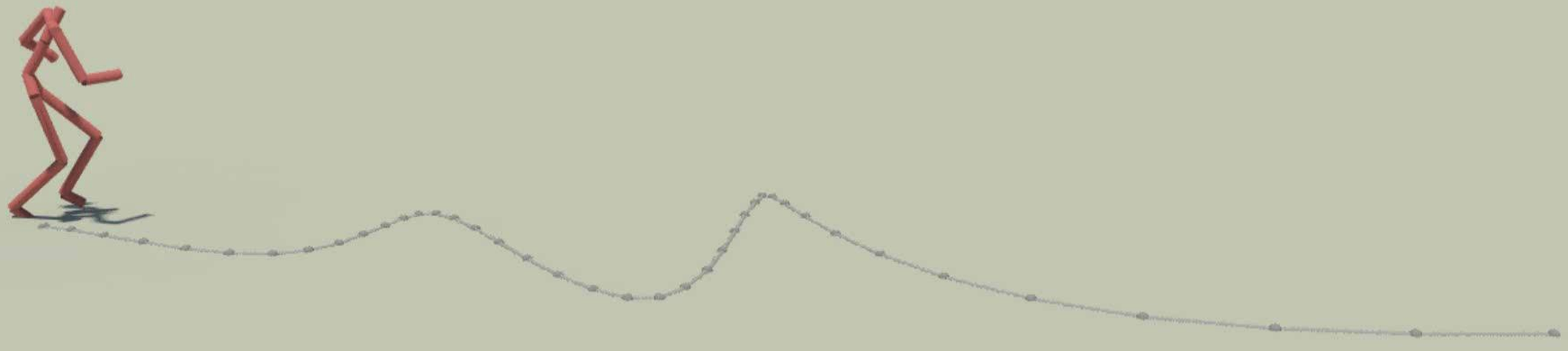
Contact Durations

Contact Frequencies

Contact Timings

Overview


Motion Synthesis
Disambiguation
x_0
x_1
x_d
υ_0
υ_1
υ_2
υ_3
υ_l
ω
τ
t_0
t_1
t_k
Control Parameters
Motion Data
Footstep Timing



Outline

Motivation

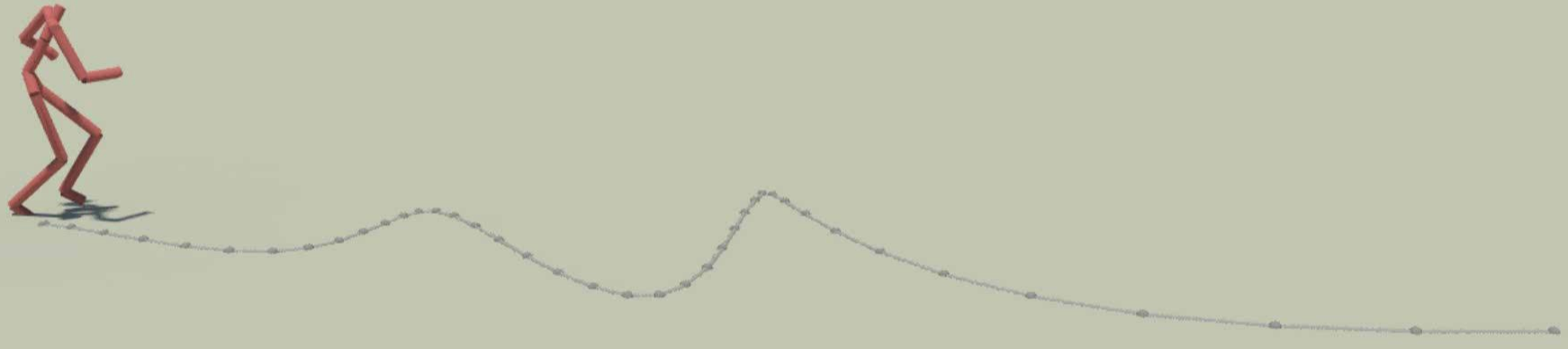
Synthesis

Editing

Discussion

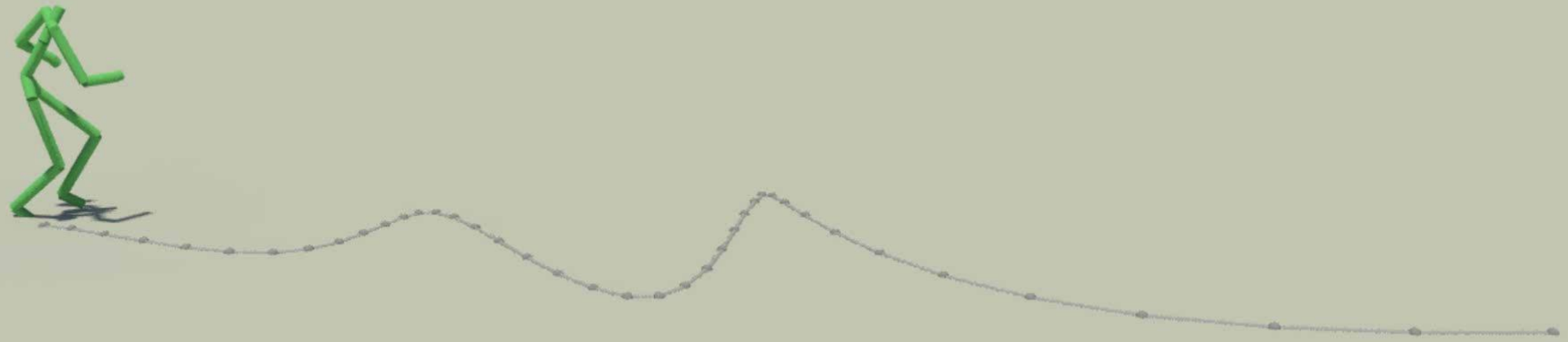
Motion Editing

Once motion is generated it must be edited



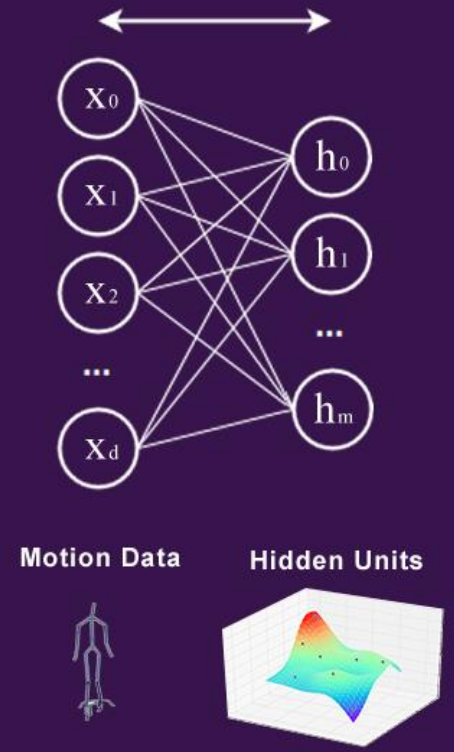
Motion Editing

Post processing may not ensure naturalness



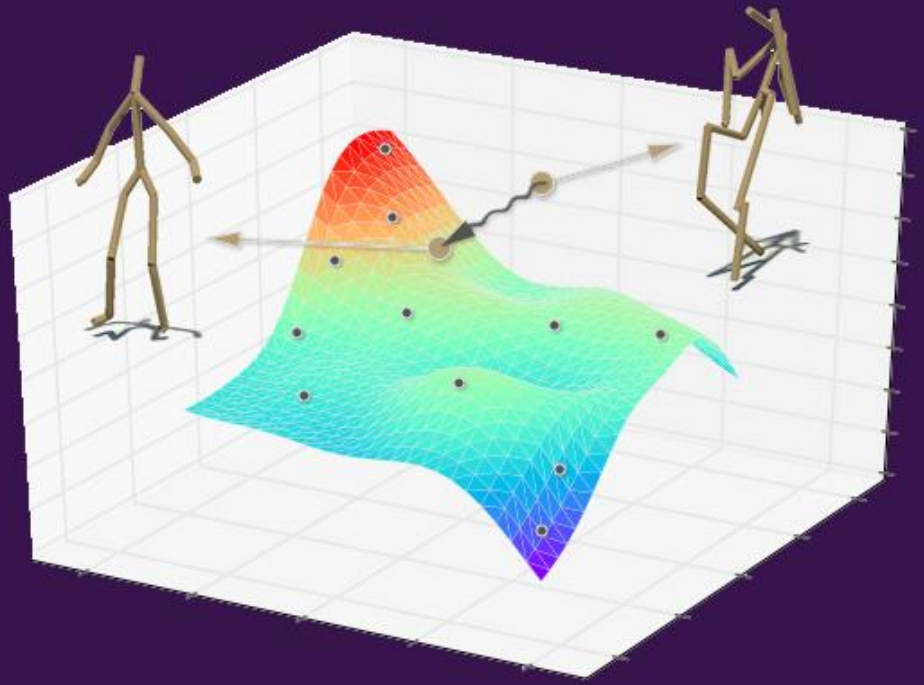
Motion Editing

- We edit using the motion manifold learned by a Convolutional Autoencoding Network [Holden et al. 2015]



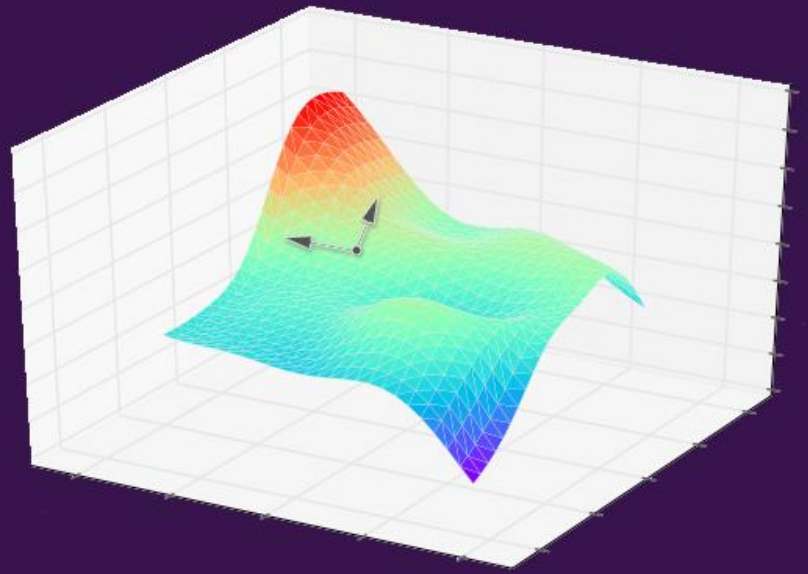
Autoencoder

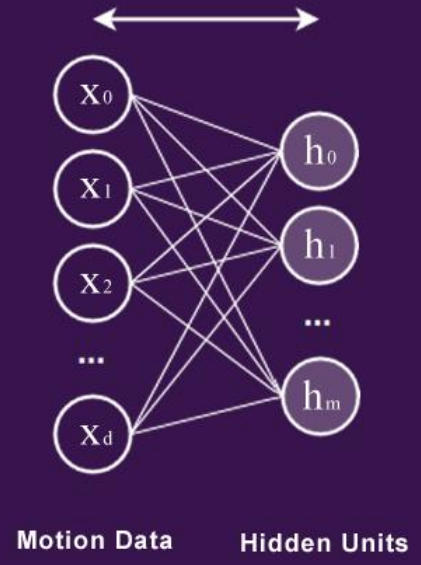
- Learns *projection operator* of motion manifold



Manifold Surface

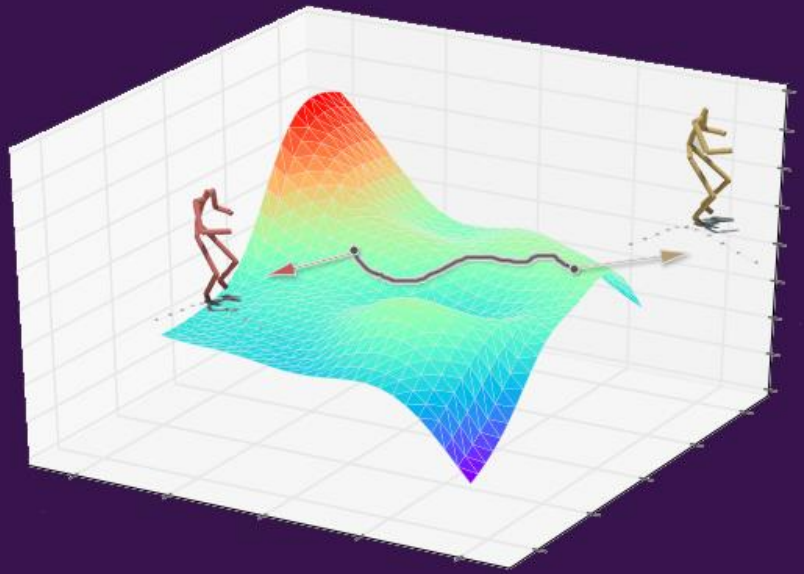
- *Hidden Unit* values parametrise manifold surface
- Adjusting them ensures motion remains natural

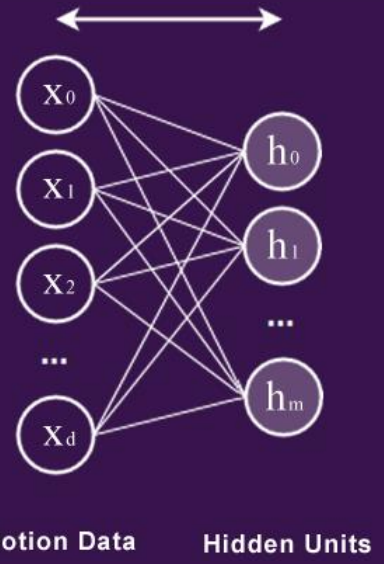




Constraint Satisfaction

- Motion editing is a *constraint satisfaction problem* over *Hidden Units*





Constraint Satisfaction

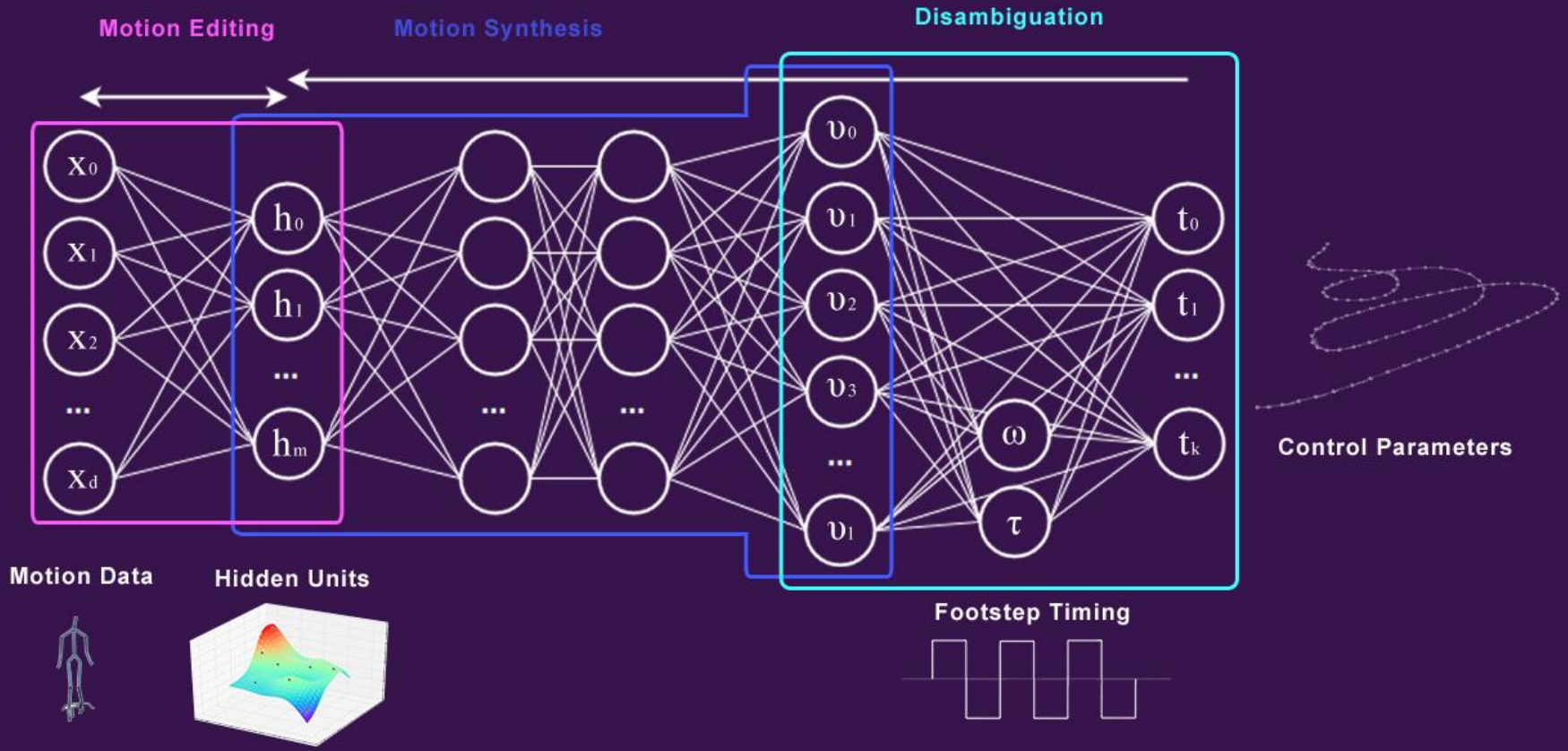
- Local foot velocity must equal global velocity

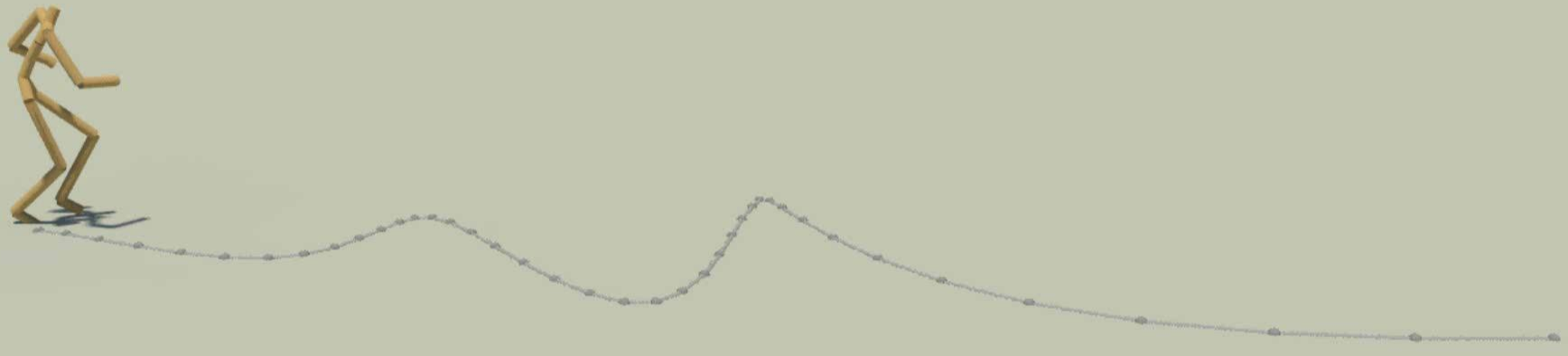
$$Pos(\mathbf{H}) = \sum_j \|\mathbf{v}_r^{\mathbf{H}} + \omega^{\mathbf{H}} \times \mathbf{p}_j^{\mathbf{H}} + \mathbf{v}_j^{\mathbf{H}} - \mathbf{v}_j'\|_2^2.$$

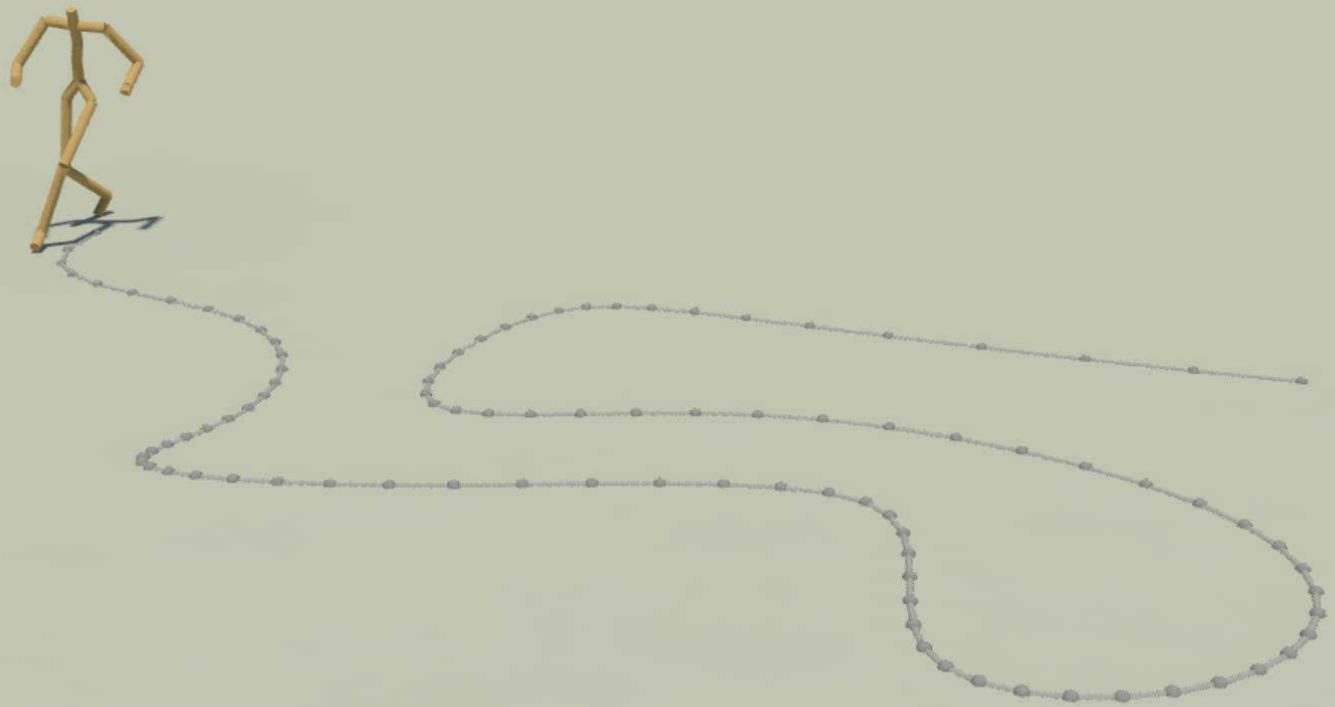
- Output trajectory must equal input trajectory

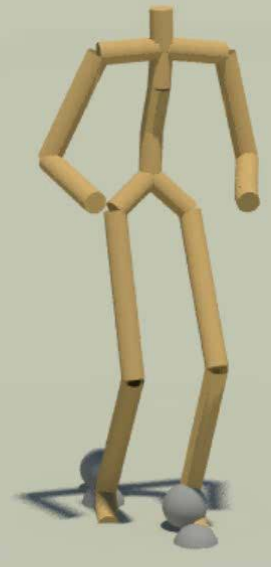
$$Traj(\mathbf{H}) \quad = \quad \|\omega^{\mathbf{H}} - \omega'\|_2^2 + \|\mathbf{v}_r^{\mathbf{H}} - \mathbf{v}_r'\|_2^2$$

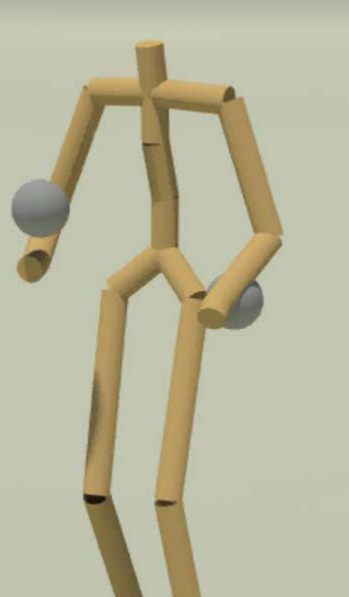
Overview

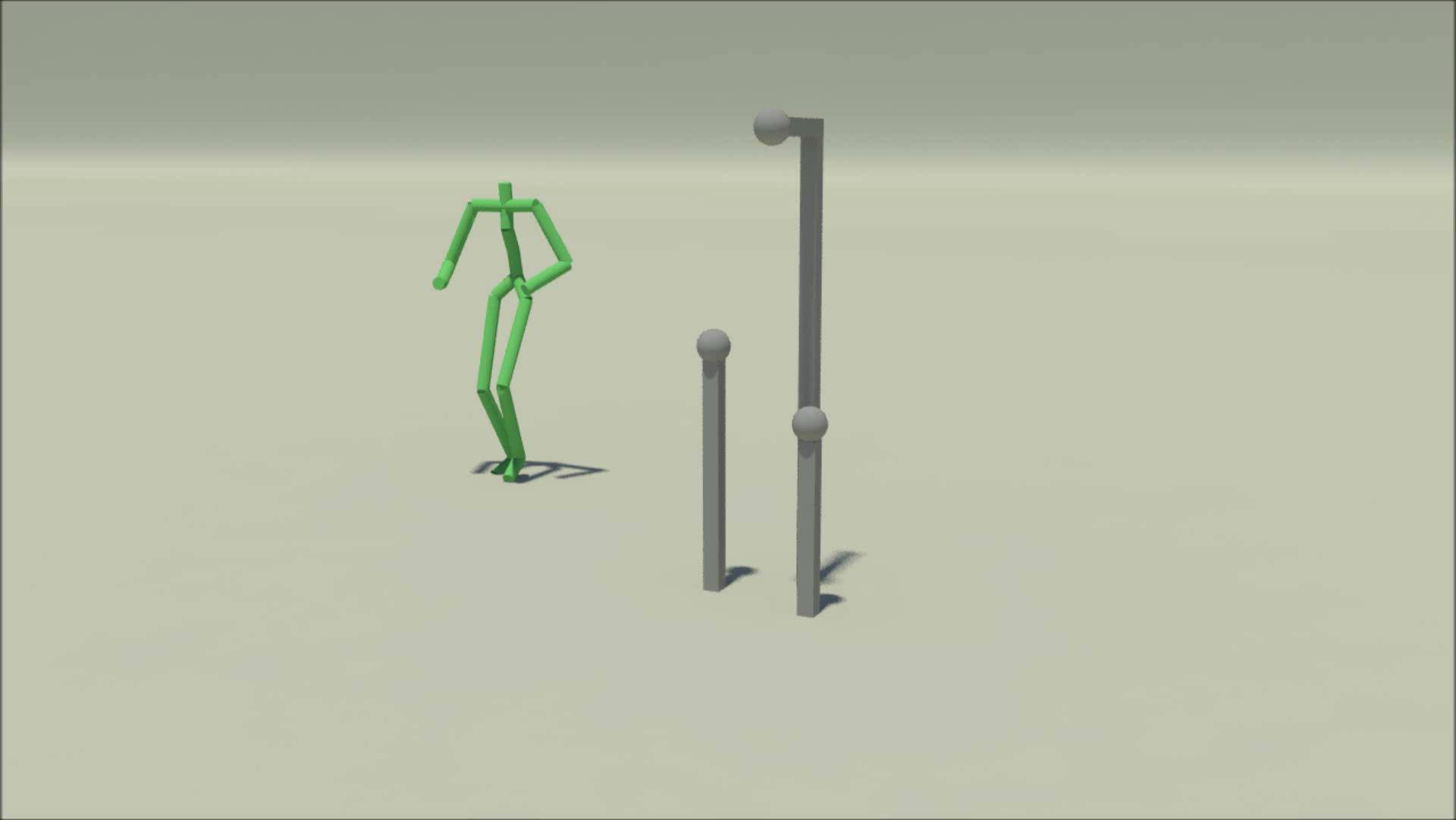

Motion Editing
Motion Synthesis
Disambiguation
x_0
x_1
x_2
...
x_d
h_0
h_1
...
h_m
...
...
υ_0
υ_1
υ_2
υ_3
...
υ_l
ω
τ
t_0
t_1
...
t_k
Control Parameters
Motion Data
Hidden Units
Footstep Timing

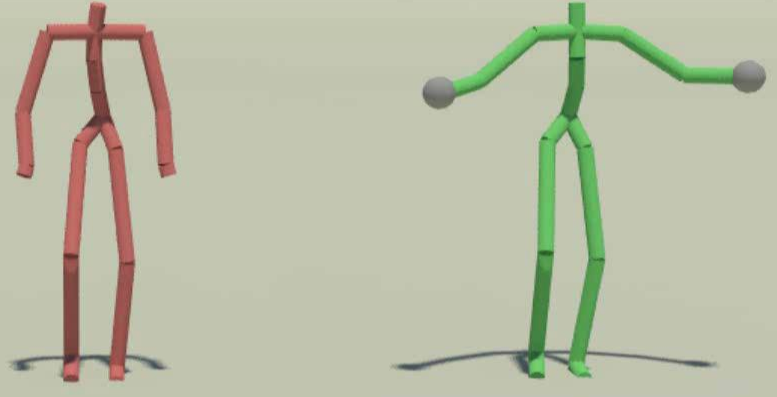






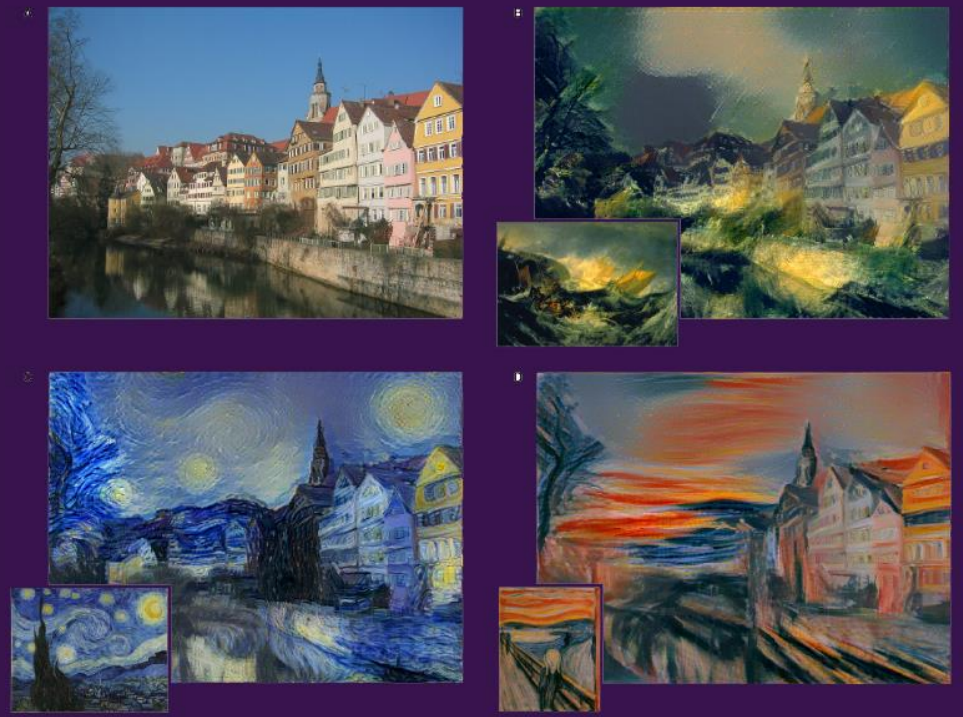






A Neural Algorithm of Artistic Style

- Combine style of one image with content of another [Gatys et al. 2015]



Style Constraint

- Gram Matrix of *Hidden Units* encode style
- Actual Values of *Hidden Units* encode content

$$Style(\mathbf{H}) = s\|G(\mathbf{\Phi}(\mathbf{S})) - G(\mathbf{H})\|_2^2 + c\|\mathbf{\Phi}(\mathbf{C}) - \mathbf{H}\|_2^2$$

$$G(\mathbf{H}) = \frac{\sum_i^n \mathbf{H}_i \mathbf{H}_i^T}{n}$$

- **No correspondence between clips required!**

Style Constraint

- Gram Matrix of *Hidden Units* encode style
- Actual Values of *Hidden Units* encode content

$$Style(\mathbf{H}) = s\|G(\mathbf{\Phi}(\mathbf{S})) - G(\mathbf{H})\|_2^2 + c\|\mathbf{\Phi}(\mathbf{C}) - \mathbf{H}\|_2^2$$

Content Term

$$G(\mathbf{H}) = \frac{\sum_i^n \mathbf{H}_i \mathbf{H}_i^T}{n}$$

- **No correspondence between clips required!**

Style Constraint

- Gram Matrix of *Hidden Units* encode style
- Actual Values of *Hidden Units* encode content

$$Style(\mathbf{H}) = s\|G(\mathbf{\Phi}(\mathbf{S})) - G(\mathbf{H})\|_2^2 + c\|\mathbf{\Phi}(\mathbf{C}) - \mathbf{H}\|_2^2$$

Style Term

$$G(\mathbf{H}) = \frac{\sum_i^n \mathbf{H}_i \mathbf{H}_i^T}{n}$$

- **No correspondence between clips required!**

Style Constraint

- Gram Matrix of *Hidden Units* encode style
- Actual Values of *Hidden Units* encode content

$$Style(\mathbf{H}) = s\|G(\mathbf{\Phi}(\mathbf{S})) - G(\mathbf{H})\|_2^2 + c\|\mathbf{\Phi}(\mathbf{C}) - \mathbf{H}\|_2^2$$

$$G(\mathbf{H}) = \frac{\sum_i^n \mathbf{H}_i \mathbf{H}_i^T}{n}$$

Gram Matrix

- **No correspondence between clips required!**

Style Constraint

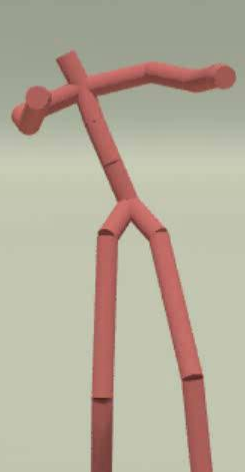
- Gram Matrix of *Hidden Units* encode style
- Actual Values of *Hidden Units* encode content

$$Style(\mathbf{H}) = s\|G(\mathbf{\Phi}(\mathbf{S})) - G(\mathbf{H})\|_2^2 + c\|\mathbf{\Phi}(\mathbf{C}) - \mathbf{H}\|_2^2$$

$$G(\mathbf{H}) = \frac{\sum_i^n \mathbf{H}_i \mathbf{H}_i^T}{n}$$

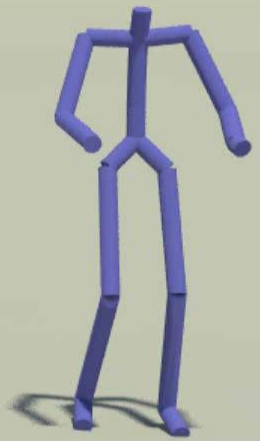
- **No correspondence between clips required!**

Style



Content

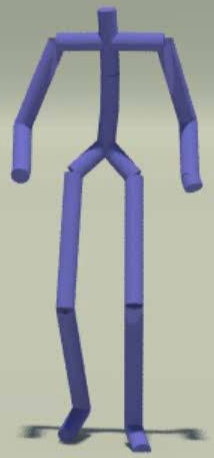
Transfer



Style

Content

Transfer



Outline

Motivation

Synthesis

Editing

Discussion

Representation

- We use joint positions local to hip
- Global rotation / translation removed using hips and shoulders direction
- Root velocity and contact times appended to representation

Training

- **Motion Manifold**
  - Several large databases (including whole CMU)
  - Training takes around 6 hours

- **Motion Synthesis**
  - Task specific data only (e.g. locomotion only)
  - Training takes around 1 hour

Procedural vs Simulated

- **Procedural**
  - Output computed at arbitrary times or in parallel
  - Ideal for precise animation
- **Simulated**
  - Output computed frame by frame in series
  - Ideal for interactive applications

Performance

- Using GPU system runs in parallel over frames
- Very fast for long motions or many characters

Task	Duration	Foot Contacts	Synthesis	Editing	Total	FPS
Walking	60s	0.025s	0.067s	1.096s	1.188s	3030
Running	60s	0.031s	0.073s	1.110s	1.214s	2965
Punching	4s	-	0.019s	0.259s	0.278s	863
Kicking	4s	-	0.020s	0.302s	0.322s	745
Style Transfer	8s	-	-	2.234s	2.234s	214
Crowd Scene	10s	0.557s	1.335s	2.252s	4.144s	28957

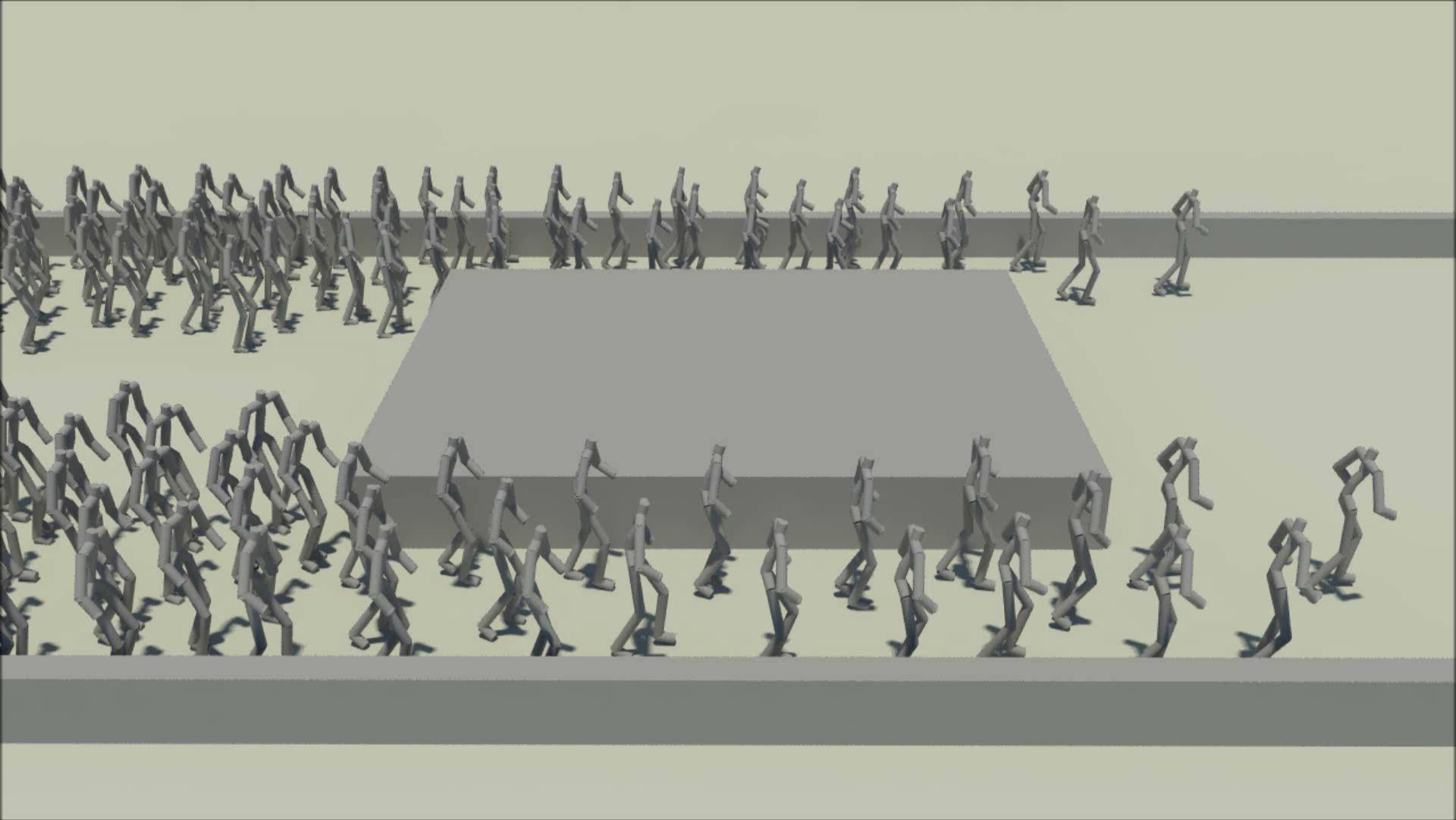
Figure 13: *Performance breakdown.*

Performance

- Using GPU system runs in parallel over frames
- Very fast for long motions or many characters

Task	Duration	Foot Contacts	Synthesis	Editing	Total	FPS
Walking	60s	0.025s	0.067s	1.096s	1.188s	3030
Running	60s	0.031s	0.073s	1.110s	1.214s	2965
Punching	4s	-	0.019s	0.259s	0.278s	863
Kicking	4s	-	0.020s	0.302s	0.322s	745
Style Transfer	8s	-	-	2.234s	2.234s	214
Crowd Scene	10s	0.557s	1.335s	2.252s	4.144s	28957

Figure 13: *Performance breakdown.*



Future Work

- Need more general solution for ambiguity issue
- Wish to use more high level features with a deeper network
- What changes are required for interactive applications?