Analysis and Retrieval Techniques for Motion and Music Data

Meinard Müller
Saarland University and MPI Informatik
meinard@mpi-inf.mpg.de

Eurographics 2009

Overview

Part 0

Music Data

Various interpretations – Beethoven's Fifth

- Bernstein
- Karajan
- Scherbakov (piano)
- MIDI (piano)

Motion Capture Data

- Digital 3D representations of motions
- Computer animation
- Sport sciences
- Computer vision
General Tasks

- Automated data organization
- Handling object deformations
- Handling multimodality
- Synchronization (alignment)
- Efficiency

Overview

Part I: Music Synchronization
Part II: Audio Structure Analysis
Part III: Audio Matching
Part IV: Motion Retrieval

Bonn University
- Prof. Dr. Michael Clausen
- PD Dr. Frank Kurth
- Dipl.-Inform. Christian Fremerey
- Dipl.-Inform. David Damm
- Dipl.-Inform. Sebastian Ewert
- Dr. Tido Röder

Habilitation

Part I
Music Synchronization

Score Representation

PhD students
- Dipl.-Inform. Andreas Baak (DFG)
- Dipl.-Math. Verena Konz (MMCI)
- Dipl.-Ing. Peter Grosche (MMCI)
- Dipl.-Inform. Thomas Helten (DFG)

Dec. 2007
Score Representation: Scanned Image

Score Representation: MusicXML

Audio Representation: Waveform

Audio Representation: Waveform

Audio Representation: Waveform

MIDI Representation

Bernstein (orchestra)  Glen Gould (piano)
MIDI Representation: Piano Roll

General Goals

- Automated organization of complex and inhomogeneous music collections
- Generation of annotations and cross-links
- Tools and methods for multimodal search, navigation and interaction

Music Information Retrieval (MIR)

Music Synchronization

Schematic view of various synchronization tasks

Music Synchronization

- Turetsky/Ellis (ISMIR 2003)
- Soulaz/Rodet/Schwarz (ISMIR 2003)
- Arli/Clausen/Kurth/Müller (ISMIR 2003)
- Hu/Dannenbergs/Tzanetakis (WASPAA 2003)
- Müller/Kurth/Röder (ISMIR 2004)
- Raphael (ISMIR 2004)
- Dixon/Widmer (ISMIR 2005)
- Müller/Matas/Kurth (ISMIR 2006)
- Dannenbergs/Raphael (Special Issue ACM 2006)
- Kurth/Müller/Fremerey/Clausen (ISMIR 2007)
- Fujihara/Goto (ICASSP 2008)
Music Synchronization: Audio-Audio

**Given:** Two different audio recordings of the same underlying piece of music.

**Goal:** Find for each position in one audio recording the musically corresponding position in the other audio recording.

---

**Music Synchronization: Audio-Audio**

**Beethoven’s Fifth**

- **Karajan**
- **Scherbakov**

Synchronization: Karajan → Scherbakov

---

**Music Synchronization: Audio-Audio**

**Bach Toccata**

- **Koopman**
- **Ruebsam**

Synchronization: Koopman → Ruebsam

---

**Music Synchronization: Audio-Audio**

- Transformation of audio recordings into sequences of feature vectors
  - $V := (v^1, v^2, \ldots, v^N)$
  - $W := (w^1, w^2, \ldots, w^M)$
- Fix cost measure $c$ on the feature space
- Compute $N \times M$ cost matrix $C(n, m) := c(v^n, w^m)$
- Compute cost-minimizing warping path from $C$
Chroma Features
Example: C-Major Scale

Chroma Features
Example: Bach Toccata
Koopman
Ruebsam

Feature resolution: 10 Hz

Chroma Features
Example: Bach Toccata
Koopman
Ruebsam

Feature resolution: 1 Hz

Chroma Features
Example: Bach Toccata
Koopman
Ruebsam

Feature resolution: 0.33 Hz

Chroma Features

WAV Chroma (10 Hz) CENS (1 Hz)

???

???

???
<table>
<thead>
<tr>
<th>Chroma Features</th>
<th>Chroma Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Beethoven’s Fifth (Bernstein)</strong></td>
<td><strong>Beethoven’s Fifth (Bernstein)</strong></td>
</tr>
<tr>
<td>WAV</td>
<td>Chroma</td>
</tr>
<tr>
<td>(10 Hz)</td>
<td>(1 Hz)</td>
</tr>
<tr>
<td>??</td>
<td>??</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chroma Features</th>
<th>Music Synchronization: Audio-Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Beethoven’s Fifth (Bernstein)</strong></td>
<td><strong>Koopman</strong> $\leadsto V := (v^1, v^2, \ldots, v^N)$ $N = 12$</td>
</tr>
<tr>
<td><strong>Beethoven’s Fifth (Piano/Sherbakov)</strong></td>
<td><strong>Ruebsam</strong> $\leadsto W := (w^1, w^2, \ldots, w^M)$ $M = 18$</td>
</tr>
<tr>
<td><strong>Brahms Hungarian Dance No. 5</strong></td>
<td>$v^{\text{norm}}, w^{\text{norm}} = 12$-dimensional normalized chroma vectors</td>
</tr>
<tr>
<td><strong>Local cost measure</strong> $c : \mathbb{R}^{12} \times \mathbb{R}^{12} \rightarrow \mathbb{R}$</td>
<td></td>
</tr>
<tr>
<td>$c(v^n, w^m) := 1 - \langle v^n, w^m \rangle$</td>
<td>$N \times M$ cost matrix $C(n, m) := c(v^n, w^m)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Music Synchronization: Audio-Audio</th>
<th>Music Synchronization: Audio-Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost-minimizing warping path</strong></td>
<td><strong>Cost-minimizing warping path</strong></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
V := (v^1, v^2, \ldots, v^N) \\
W := (w^1, w^2, \ldots, w^M)
\end{align*}
\]
Cost-Minimizing Warping Path

- Computation via dynamic programming
  - Dynamic Time Warping (DTW)
- Memory requirements and running time: $O(NM)$
- Problem: Infeasible for large $N$ and $M$
- Example: Feature resolution 10 Hz, pieces 15 min
  \[ N, M \approx 10,000 \]
  \[ N \cdot M \approx 100,000,000 \]

Strategy: Global Constraints

- Sakoe-Chiba band
- Itakura parallelogram

Problem: Optimal warping path not in constraint region

Strategy: Multiscale Approach

- Compute optimal warping path on coarse level
- Project on fine level
- Specify constraint region
Strategy: Multiscale Approach

Compute constrained optimal warping path

Strategy: Multiscale Approach

- Suitable features?
- Suitable resolution levels?
- Size of constraint regions?

Good trade-off between efficiency and robustness?

Strategy: Multiscale Approach

Resolution 4 Hz          Resolution 2 Hz            Resolution 1 Hz

Problem: Cost matrix may degenerate

~ useless warping path

Strategy: Multiscale Approach

Improve robustness by enhancing cost matrix

Resolution 4 Hz          Resolution 2 Hz            Resolution 1 Hz

Enhanced                 Original

Strategy: Multiscale Approach

Improve robustness by enhancing cost matrix

Resolution 4 Hz          Resolution 2 Hz            Resolution 1 Hz

Enhanced                 Original
Strategy: Multiscale Approach

Chroma features at three levels: 0.33 Hz / 1 Hz / 10 Hz

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RaetfiRam</td>
<td>1144.9</td>
<td>RaetfiKar</td>
<td>1154.8</td>
<td>31.15</td>
<td>1.08</td>
<td>3.4R</td>
</tr>
</tbody>
</table>

Number of matrix entries needed for DTW and MsDTW:

<table>
<thead>
<tr>
<th>Level</th>
<th>DTW</th>
<th>MsDTW</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120,808,050</td>
<td>2,117,929</td>
<td>1.75</td>
</tr>
<tr>
<td>2</td>
<td>1.77e+07</td>
<td>17,757</td>
<td>1.34</td>
</tr>
<tr>
<td>3</td>
<td>124,404</td>
<td>124,404</td>
<td>100</td>
</tr>
</tbody>
</table>

Music Synchronization: Audio-Audio

Conclusions

- Chroma features
  - suited for harmony-based music
- Relatively coarse but good global alignments
- Multiscale approach: simple, robust, fast

Applications

- Efficient music browsing
- Blending from one interpretation to another one
- Mixing and morphing different interpretations
- Tempo studies

System: Match (Dixon)

System: SyncPlayer/AudioSwitcher
Music Synchronization: MIDI-Audio

MIDI = metadata
Automated annotation
Audio recording

Sonification of annotations

Music Synchronization: MIDI-Audio

Applications

- Automated audio annotation
- Accurate audio access after MIDI-based retrieval
- Automated tracking of MIDI note parameters during audio playback

Music Synchronization: Scan-Audio

Scanned Sheet Music
Symbolic Note Events

OMR

Correspondence
Music Synchronization: Scan-Audio

Scanned Sheet Music → Symbolic Note Events

OMR [Audio Recording]

Correspondence

High Quality

“Dirty” but hidden

System: SyncPlayer/SheetMusic

Music Synchronization: Lyrics-Audio

Lyrics-Audio → Lyrics-MIDI + MIDI-Audio

System: SyncPlayer/LyricsSeeker
Conclusions: Music Synchronization

Various requirements
- Efficiency
- Robustness
- Accuracy
- Variability of music

Conclusions: Music Synchronization

Combination of various strategies
- Feature level
- Local cost measure level
- Global alignment level
- Evidence pooling using competing strategies

Example: MIDI-Audio synchronization
- Chroma-Chroma:
- Chroma-Chroma + onset-bonus:

Conclusions: Music Synchronization

Combination of various strategies
- Feature level
- Local cost measure level
- Global alignment level
- Evidence pooling using competing strategies

Offline vs. Online
- Online version: Dixon/Widmer (ISMIR 2005)
- Score-following
- Automatic accompaniment

Conclusions: Music Synchronization

Presence of variations
- Instrumentation
- Musical structure
- Polyphony
- Musical key
- ...

Part II

Audio Structure Analysis
Music Structure Analysis

- Music segmentation
  - pitch content (e.g., melody, harmony)
  - music texture (e.g., timbre, instrumentation, sound)
  - rhythm

- Detection of repeating sections, phrases, motives
  - song structure (e.g., intro, versus, chorus)
  - musical form (e.g., sonata, symphony, concerto)

- Detection of other hidden relationships

Audio Structure Analysis

**Given**: CD recording

**Goal**: Automatic extraction of the repetitive structure (or of the musical form)

**Example**: Brahms Hungarian Dance No. 5 (Ormandy)

Audio Structure Analysis

- Audio features
- Cost measure and cost matrix
  - self-similarity matrix
- Path extraction (pairwise similarity of segments)
- Global structure (clustering, grouping)

Audio Structure Analysis

- Dannenberg/Hu (ISMIR 2002)
- Paeters/Burtscha/Robot (ISMIR 2002)
- Cooper/Foote (ISMIR 2002)
- Goto (ICASSP 2003)
- Chai/Vercoe (ACM Multimedia 2003)
- Müller/Kurth (EURASIP 2007)
- Rhodes/Casey (ISMIR 2007)
- Peeters (ISMIR 2007)

Audio Structure Analysis

- Audio \( \sim V := (v^1, v^2, \ldots, v^N) \)
- \( v^n \) = 12-dimensional normalized chroma vector
- Local cost measure \( c : \mathbb{R}^{12} \times \mathbb{R}^{12} \to \mathbb{R} \)
  \( c(v^n, w^m) := 1 - (v^n, w^m) \)
- \( N \times N \) cost matrix \( C(n, m) := c(v^n, w^m) \)
  \( \sim \) quadratic self-similarity matrix
Audio Structure Analysis

**Self-similarity matrix**

Matrix Enhancement

**Challenge:** Presence of musical variations

- Fragmented paths and gaps
- Paths of poor quality
- Regions of constant (low) cost
- Curved paths

Idea: Enhancement of path structure

Matrix Enhancement (Shostakovich)

**Idea:** Usage of contextual information (Foote 1999)

\[
C_L(n, m) := \frac{1}{L} \sum_{t=0}^{L-1} c(v_n(t), v_m(t))
\]

- Comparison of entire sequences
- \(L = \) length of sequences
- \(C_L = \) enhanced cost matrix

\(\sim\) smoothing effect

Cost matrix \(C\)
Matrix Enhancement (Shostakovich)

Enhanced cost matrix $C_L$

Matrix Enhancement (Brahms)

Cost matrix $C$

Matrix Enhancement (Brahms)

Enhanced cost matrix $C_L$

Problem: Relative tempo differences are smoothed out

Matrix Enhancement

Idea: Smoothing along various directions and minimizing over all directions

$$C_{L}^{\min}(n,m) := \min_{k} C_{L}^{\text{slope}_k}(n,m)$$

- $\text{slope}_k = k$th direction of smoothing
- $C_{L}^{\text{slope}_k}$ - enhanced cost matrix w.r.t. $\text{slope}_k$
- Usage of eight slope values
  - tempo changes of -30 to +40 percent

Matrix Enhancement

Cost matrix $C$

Matrix Enhancement
Matrix Enhancement

Cost matrix $C_L$ with $L = 20$
Filtering along main diagonal

Matrix Enhancement

Cost matrix $C_L^{\text{min}}$ with $L = 20$
Filtering along 8 different directions and minimizing

Path Extraction

• Start with initial point
• Extend path in greedy fashion
• Remove path neighborhood

Path Extraction

Cost matrix $C$

Path Extraction

Enhanced cost matrix $C_L$
Path Extraction

Thresholded $C^{\text{min}}_2$

Path Extraction

Thresholded $C^{\text{min}}_2$, upper left

Path Extraction

Path removal

Path Extraction

Path removal

Path Extraction

Path removal

Path Extraction

Extracted paths
How can one derive the global structure from pairwise relations?

- Task: Computation of similarity clusters
- Problem: Missing and inconsistent path relations
- Strategy: Approximate "transitive hull"
Global Structure

Path relations

Transposition Invariance

Example: Zager & Evans “In The Year 2525”

Goto (ICASSP 2003)
- Cyclically shift chroma vectors in one sequence
- Compare shifted sequence with original sequence
- Perform for each of the twelve shifts a separate structure analysis
- Combine the results
Transposition Invariance

Goto (ICASSP 2003)
- Cyclically shift chroma vectors in one sequence
- Compare shifted sequence with original sequence
- Perform for each of the twelve shifts a separate structure analysis
- Combine the results

Müller/Clausen (ISMIR 2007)
- Integrate all cyclic information in one transposition-invariant self-similarity matrix
- Perform one joint structure analysis

Example: Zager & Evans “In The Year 2525”

Original: \((v^1, \ldots, v^N)\)
Shifted: \((\sigma(v^1), \ldots, \sigma(v^N))\)
Transposition Invariance

Minimize over all twelve matrices

Thresholded self-similarity matrix

Path extraction

Self-similarity matrix (thresholded)
Transposition Invariance
Stabilizing effect

Self-similarity matrix (thresholded)

Transposition Invariance
Stabilizing effect

Transposition-invariant self-similarity matrix (thresholded)

Transposition Invariance

Transposition-invariant matrix
Minimizing shift index

Transposition Invariance

Transposition-invariant matrix
Minimizing shift index

Transposition Invariance

Transposition-invariant matrix
Minimizing shift index = 0

Transposition Invariance

Transposition-invariant matrix
Minimizing shift index = 1
Transposition Invariance

Transposition-invariant matrix
Minimizing shift index = 2

Serra/Gomez (ICASSP 2008): Used for Cover Song ID
Discrete structure \( \rightarrow \) suitable for indexing?

Transposition Invariance

Example: Beethoven “Tempest”

Self-similarity matrix

Transposition-invariant self-similarity matrix

Conclusions: Audio Structure Analysis

Challenge: Musical variations

- Timbre, dynamics, tempo
- Musical key - cyclic chroma shifts
- Major/minor
- Differences at note level / improvisations

Strategy: Matrix enhancement

- Filtering techniques / contextual information
  - Cooper/Foote (ISMIR 2002)
  - Müller/Kurth (ICASSP 2006)
- Transposition-invariant similarity matrices
  - Goto (ICASSP 2003)
  - Müller/Clausen (ISMIR 2007)
- Higher-order similarity matrices
  - Peeters (ISMIR 2007)
Conclusions: Audio Structure Analysis

Challenge: Hierarchical structure of music

System: SmartMusicKiosk (Goto)

System: SyncPlayer/AudioStructure

Part III

Audio Matching

Given:
- Large music database containing several
  - recordings of the same piece of music
  - interpretations by various musicians
  - arrangements in different instrumentations

Goal:
- Given a short query audio clip, identify all corresponding audio clips of similar musical content
  - irrespective of the specific interpretation and instrumentation
  - automatically and efficiently

Query-by-Example paradigm

Audio Matching

- Müller/Kurth/Clausen (ISMIR 2005)
- Kurth/Müller (IEEE T-ASLP 2008)
- Allamanche et al. (AES 2001)
- Cano et al. (IEEE MMSP 2002)
- Kurth/Clausen/Ribbrock (AES 2002)
- Wang (ISMIR 2003)
- Shrestha/Kalker (ISMIR 2004)

Related problems

Audio identification
- Müller/Kurth/Clausen (ISMIR 2005)
- Cano et al. (IEEE MMSP 2002)
- Kurth/Clausen/Ribbrock (AES 2002)
- Wang (ISMIR 2003)
- Shrestha/Kalker (ISMIR 2004)

Audio synchronization

Audio structure analysis
Audio Matching

General strategy

- Normalized and smoothed chroma features
  - correlates to harmonic progression
  - robust to variations in dynamics, timbre, articulation, local tempo
- Robust matching procedure
  - efficient
  - robust to global tempo variations
  - scalable using index structure

Feature Design

Beethoven’s Fifth: Bernstein

Resolution: 10 features/second
Feature window size: 200 milliseconds

Feature Design

Beethoven’s Fifth: Bernstein vs. Sawallisch

Resolution: 10 features/second
Feature window size: 200 milliseconds

Feature Design

Beethoven’s Fifth: Bernstein

Resolution: 1 features/second
Feature window size: 4000 milliseconds

Feature Design

Two stages:

Stage 1: Local chroma energy distribution features
Stage 2: Normalized short-time statistics

\[ \text{CENS} = \text{Chroma Energy Normalized Statistics} \]
Feature Design

Beethoven’s Fifth: Bernstein vs. Sawallisch

Resolution: 1 features/second
Feature window size: 4000 milliseconds

Matching Procedure

Compute CENS feature sequences
- Database $D \sim F[D] = (v^1, v^2, \ldots, v^N)$
- Query $Q \sim F[Q] = (w^1, w^2, \ldots, w^M)$
- $N \approx 500000$, $M \approx 20$

$$\Delta(i) = \text{local distance}(v^i, v^{i+1}, \ldots, v^{i+M-1}, w^1, w^2, \ldots, w^M)$$

$$\rightsquigarrow \text{Global distance function } \Delta : [1 : N] \rightarrow [0, 1]$$

Matching Procedure

Query: Beethoven’s Fifth / Bernstein, first 20 seconds

Best audio matches: 1

Matching Procedure

Query: Beethoven’s Fifth / Bernstein, first 20 seconds

Best audio matches: 2

Matching Procedure

Query: Beethoven’s Fifth / Bernstein, first 20 seconds

Best audio matches: 3
Matching Procedure

Query: Beethoven’s Fifth / Bernstein, first 20 seconds

Best audio matches: 4

Matching Procedure

Query: Beethoven’s Fifth / Bernstein, first 20 seconds

Best audio matches: 5

Matching Procedure

Query: Beethoven’s Fifth / Bernstein, first 20 seconds

Best audio matches: 6

Matching Procedure

Query: Beethoven’s Fifth / Bernstein, first 20 seconds

Best audio matches: 7

Global Tempo Variations

Query: Beethoven’s Fifth / Bernstein, first 20 seconds
Problem: Karajan is much faster useless

Solution?

Global Tempo Variations

Query: Beethoven’s Fifth / Bernstein, first 20 seconds
Problem: Karajan is much faster useless

Solution: Make Bernstein query faster and compute new
Global Tempo Variations

Query: Beethoven's Fifth / Bernstein, first 20 seconds
Problem: Karajan is much faster $\rightsquigarrow$ useless $\triangle$
Solution: Compute $\Delta$ for various tempi

Experiments

- Audio database $> 110$ hours, 16.5 GB
- Preprocessing $\rightsquigarrow$ CENS features, 40.3 MB
- Query clip $\approx$ 20 seconds
- Query response time $< 10$ seconds

Experiments

<table>
<thead>
<tr>
<th>Place</th>
<th>Name</th>
<th>Tempo</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beethoven's Fifth/Bernstein</td>
<td>0 - 21</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Beethoven's Fifth/Bernstein</td>
<td>101 - 122</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Beethoven's Fifth/Karajan</td>
<td>86 - 103</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Beethoven's Fifth/Karajan</td>
<td>252 - 271</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Beethoven (Last)/Scherbakov</td>
<td>0 - 18</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Beethoven's Fifth/Scherbakov</td>
<td>270 - 290</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Beethoven's Fifth (Last)/Scherbakov</td>
<td>00 - 100</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Brahms op 07,1/Levine</td>
<td>28 - 43</td>
<td></td>
</tr>
</tbody>
</table>

Experiments

Query: Shostakovich, Waltz/Chailly, first 27 seconds
Experiments

Query: Shostakovich, Waltz/Chailly, first 21 seconds

<table>
<thead>
<tr>
<th>Rank</th>
<th>( \Delta \text{MSE} )</th>
<th>Piece</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0172</td>
<td>Shostakovich/Chailly</td>
<td>0 - 21</td>
</tr>
<tr>
<td>2</td>
<td>0.0505</td>
<td>Shostakovich/Chailly</td>
<td>41 - 60</td>
</tr>
<tr>
<td>3</td>
<td>0.0983</td>
<td>Shostakovich/Chailly</td>
<td>180 - 188</td>
</tr>
<tr>
<td>4</td>
<td>0.1044</td>
<td>Shostakovich/Yablonsky</td>
<td>1 - 19</td>
</tr>
<tr>
<td>5</td>
<td>0.1090</td>
<td>Shostakovich/Yablonsky</td>
<td>36 - 52</td>
</tr>
<tr>
<td>6</td>
<td>0.1401</td>
<td>Shostakovich/Yablonsky</td>
<td>156 - 174</td>
</tr>
<tr>
<td>7</td>
<td>0.1476</td>
<td>Shostakovich/Chailly</td>
<td>144 - 162</td>
</tr>
<tr>
<td>8</td>
<td>0.1626</td>
<td>Bach BWV 582/Chorzempa</td>
<td>358 - 373</td>
</tr>
<tr>
<td>9</td>
<td>0.1668</td>
<td>Beethoven op 37,1/Toscanini</td>
<td>12 - 28</td>
</tr>
<tr>
<td>10</td>
<td>0.1729</td>
<td>Beethoven op 37,1/Pollini</td>
<td>202 - 218</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Conclusions

Strategy: Absorb variations at feature level

- Chroma \( \sim \) invariance to timbre
- Normalization \( \sim \) invariance to dynamics
- Smoothing \( \sim \) invariance to local time deviations

System: SyncPlayer/AudioMatching

Global Matching Procedure

- Strategy: Exact matching and multiple scaled queries
  - simulate tempo variations by feature resampling
  - different queries correspond to different tempi
  - indexing possible

- Strategy: Dynamic Time Warping
  - subsequence variant
  - more flexible (in particular for longer queries)
  - indexing hard

Multimodal Computing and Interaction

MusicXML (Text)    Music
Sheet Music (Image) MIDI
CD / MP3 (Audio)
Music Literature (Text) Singing / Voice (Audio)
Music Film (Video) Dance / Motion (Mocap)

Part IV

Motion Retrieval
Motion Capture Data

- Digital 3D representations of motions
- Computer animation
- Sports
- Gait analysis

Motion Capture Data

Application: Motion Morphing

From Kovar/Gleicher (SIGGRAPH 2004)

Motion Capture Data

Optical System

Mechanical and magnetic systems

Motion Capture Data

Skeletal kinematic chain

Conversion: Marker → Skeleton

http://vrlab.epfl.ch/research/MC_motion_capture.html

http://www.metamotion.com/gypsy/gypsy-motion-capture-system.html

http://www.metamotion.com/gypsy/gypsy-motion-capture-system.html

http://www.metamotion.com/gypsy/gypsy-motion-capture-system.html

http://www.metamotion.com/gypsy/gypsy-motion-capture-system.html
Motion Capture Data

Motion Retrieval

\[ \mathcal{D} = \text{MoCap database} \]
\[ \mathcal{Q} = \text{query motion clip} \]

**Goal**: find all motion clips in \( \mathcal{D} \) similar to \( \mathcal{Q} \)

Motion Retrieval

Motion Similarity

- **Numerical similarity** vs. **logical similarity**

  Logically related motions may exhibit significant spatio-temporal variations

Motion Similarity

Global Transforms
- Translation
- Spatial scaling
- Rotation
- Reflection
- Temporal Scaling

Motion Styles
- Cheerful walking
- Furious walking
- Limping
- Tiptoeing
- Marching
Motion Similarity

Spatio-Temporal Deformations

\[
e^{3D}\left(D(n), D(m) \right) := \min_{\theta, x, z} \left( \sum_{i=1}^{K} w_i \| p_i - T_{\theta, x, z}(p_i') \|^2 \right)
\]

Local Similarity Measure

Point cloud (Kovar & Gleicher)

\[
e^{3D}\left(D(n), D(m) \right) := \min_{\theta, x, z} \left( \sum_{i=1}^{K} w_i \| p_i - T_{\theta, x, z}(p_i') \|^2 \right)
\]

Local Similarity Measure

Point cloud (Kovar & Gleicher)

\[
e^{\text{Quat}} : \mathcal{J} \times \mathcal{J} \rightarrow [0, 1]
\]

\[
e^{\text{Quat}}(j, j') := \sum_{b \in B} w_b \cdot \frac{2}{\pi} \cos \left| \langle q_b, q_b' \rangle \right|
\]
Dynamic Time Warping (DTW)

Given: motion database (one single document)
Compute: self-similarity matrix
Self-Similarity Matrix

- Query: segment of motion database
- Consider similarity matrix over query

Some Drawbacks

- DTW-based techniques computationally expensive
  - do not scale to large databases
- Rely on numerical features
  - hard to identify logically related motions
- No user-specified "center of attention"
  - incorporation of a-priori knowledge not possible

Other Recent Approaches

- Wu et al. (IPPR 2003):
  - identify candidates for start and end frames
  - use DTW to compute actual distance from query
- Keogh et al. (VLDB 2004):
  - identify motion clips differing by global scaling
- Forbes/Fiume (SCA 2005):
  - PCA-based local features
  - substring DTW for matching

Our Approach

- Introduction of relational features
  - accounting for spatial deformations
- Introduction of adaptive temporal segmentation
  - accounting for temporal deformations
- Usage of linear time/space indexing techniques
  - scalable to large databases

Müller/Röder/Clausen (SIGGRAPH 2005)
Relational Features

Right foot
Left foot
Conjunction

Temporal Segmentation:

Induced feature sequence:

\(((), (), (), (), (), (), (), (), ()\))
Relational Features

Spatio-temporal invariance

Motion Retrieval

Feature Adaptivity

Motion Retrieval

Motion Retrieval
Motion Retrieval
Indexing with inverted lists

Preprocessing (Index)
- 3 hours of Mocap data
- 31 (manually designed) boolean features

<table>
<thead>
<tr>
<th>Database</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,200,000 frames</td>
<td>230,000 segments</td>
</tr>
<tr>
<td>370 MB</td>
<td>7.54 MB</td>
</tr>
</tbody>
</table>

- Index construction: 376 seconds
- Index time and index size linear in #(segments)
- Index is query independent

Query and retrieval stage
- Query motion clip
- Optional selection of preferences
  - feature selection
  - degree of fault tolerance
  - ranking strategy
- Automatic conversion of query into feature sequence
- Retreiving hits based on inverted lists
- Typical query response times: 10-300 ms
Motion Retrieval

Results: Punch

Motion Retrieval

Results: Kick

Motion Retrieval

Results: Squat (unranked)

Motion Retrieval

Results: Squat (top 9 ranked)

Strengths and Weaknesses

<table>
<thead>
<tr>
<th>Feature</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval</td>
<td>Efficiency</td>
<td>Rigid</td>
</tr>
<tr>
<td></td>
<td></td>
<td>False positives/negatives</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ranking?</td>
</tr>
<tr>
<td>Feature Design</td>
<td>Clear semantics</td>
<td>Ad-hoc</td>
</tr>
<tr>
<td>Feature Selection</td>
<td>A-priori knowledge</td>
<td>Critical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automation?</td>
</tr>
</tbody>
</table>

Motion Templates

Müller/Röder (SCA 2006)
Motion Templates

Motion Templates

Motion Templates

Motion Templates
MT-based Motion Retrieval: Jumping Jack

MT-based Motion Retrieval: Elbow-To-Knee

MT-based Motion Retrieval: Elbow-To-Knee

MT-based Motion Retrieval: Cartwheel

MT-based Motion Retrieval: Throw

MT-based Motion Retrieval: Throw
MT-based Motion Retrieval: Basketball

MT-based Motion Retrieval: Lie Down Floor

Problems and Future Work

- **Efficiency**: MT-based matching is linear in database size
- **Hit quality**: MT-based matching has problems with short motions with few characteristic aspects
- **Current work**: Combine MT-based matching with aspects of exact matching:
  - “Hard constraints” such as keyframes
  - Index-based preselection

Conclusions

- Automated data organization
- Handling object deformations
- Handling multimodality
- Synchronization (alignment)
- Efficiency
Conclusions

Part I: Music Synchronization

- C. Fremerey, F. Kurth, M. Müller, and M. Clausen, A demonstration of the SyncPlayer system, in Proc. ISMIR, Vienna, Austria, 2007.

Part I: Audio Structure Analysis


Part I: Audio Matching

- M. Müller, F. Kurth, D. Damm, C. Fremerey, and M. Clausen, Lyrics-based audio retrieval and multimodal navigation in music collections, in Proc. ECDL, 2007, pp. 112–123.
- N. Orio, Alignment of performances with scores aimed at content-based music access and retrieval, in Proc. ECDL, 2002, pp. 479–492.
- C. Raphael, A hybrid graphical model for aligning polyphonic audio with musical scores, in Proc. ISMIR, Barcelona, Spain, 2004.

Part IV: Motion Retrieval

Part II: Audio Structure Analysis


Part II: Audio Structure Analysis


Part III: Audio Matching


Part IV: Motion Retrieval

- F. Kurth, M. Clausen, and A. Ribbrock, Identification of highly distorted audio material for querying large scale data bases, in Proc. 112th AES Convention, Munich, Germany, 2002.

- F. Kurth, M. Clausen, and A. Ribbrock, Identification of highly distorted audio material for querying large scale data bases, in Proc. 112th AES Convention, Munich, Germany, 2002.


- F. Kurth, M. Clausen, and A. Ribbrock, Identification of highly distorted audio material for querying large scale data bases, in Proc. 112th AES Convention, Munich, Germany, 2002.

Part IV: Motion Retrieval


Book

Müller, Meinard
Information Retrieval for Music and Motion
2007, XVI. 318 pages
136 Illus. 39 in Color, Hardcover
ISBN: 978-3-540-74047-6
www.springer.com/978-3-540-74047-6/
69,50 EUR