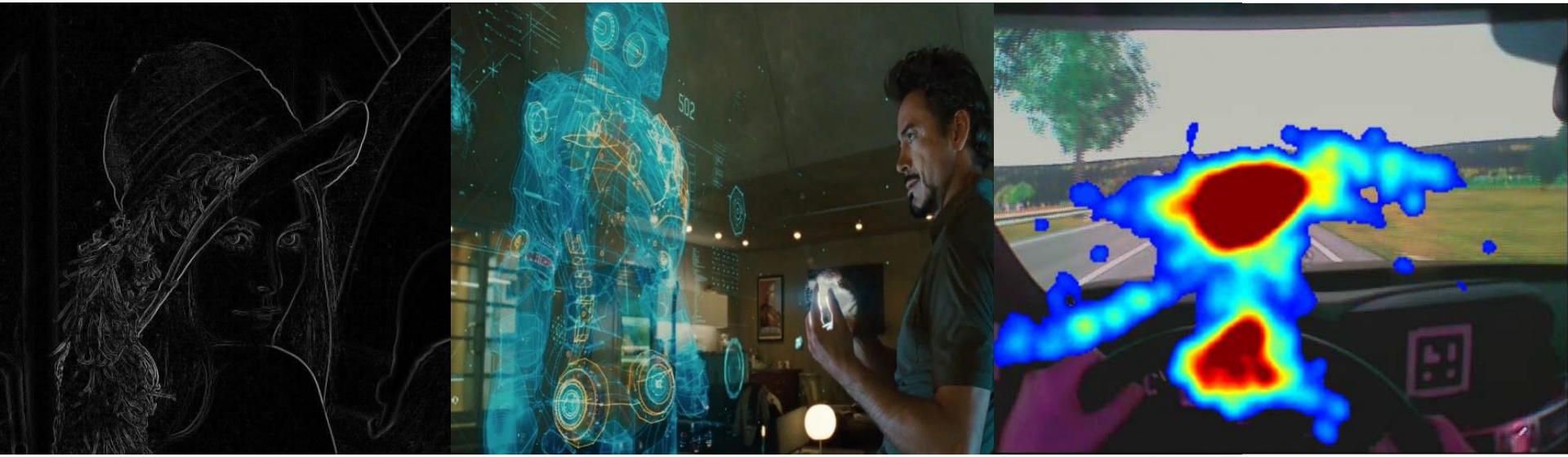
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT

Medieninformatik / Human-Computer Interaction



Grundlagen der Multimediatechnik

Gestenanalyse

28.01.2022, Prof. Dr. Enkelejda Kasneci

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Termine und Themen

22.10.2021	Einführung
29.10.2021	Menschliche Wahrnehmung – visuell, akustisch, haptisch, …
05.11.2021	Informationstheorie, Textcodierung und -komprimierung
12.11.2021	Bildverbesserung
19.11.2021	Bildanalyse
26.11.2021	Grundlagen der Signalverarbeitung
03.12.2021	Bildkomprimierung
10.12.2021	Videokomprimierung
17.12.2022	Audiokomprimierung
14.01.2022	Videoanalyse
21.01.2022	Dynamic Time Warping
28.01.2022	Gestenanalyse
04.02.2022	Tiefendatengenerierung
11.02.2022	FAQ mit den Tutoren
17.02.2022	Klausur, 14-16 Uhr, N10+N11

Gesten

Definition: spontane oder bewusst eingesetzte Bewegung **des Körpers** besonders **der Hände** und **des Kopfes**, die jemandes Worte begleitet oder ersetzt und eine bestimmte innere Haltung ausdrückt.[1]

Linguistische Typologie

- **Deiktische Gesten** werden häufig als ein abstraktes Zeigen auf nicht vorhandene Gegenstände, Orte oder Ideen genutzt
- **Ikonische Gesten** bilden ein Ikon die Wirklichkeit in übertragener Form ab
- **Metaphorische Gesten** beschreiben die Konzepte, die keine physikalische Form haben
- **Rhythmische Gesten** sind kleine rhythmische Bewegungen, die etwas betonen oder korrigieren sollen

Gestenerkennung

- Verfahren, um menschliche Gesten zur Interaktion mit technischen Geräten zu verwenden

- Gesten müssen anhand von Bildsequenzen erkannt und interpretiert werden

- Zu berücksichtigende Aspekte bei der Gestenerkennung:
  - Bewegung
  - Position
  - Geschwindigkeit
  - Richtung

Sensorik zur Gestenerkennung

- **Zwei Möglichkeiten**
  - Erkennung mittels **am Körper befindlicher Sensorik** (z.B. Datenhandschuh, Wii-Controller)
  - Erkennung mittels **externer Sensorik** (z.B. Kameras, oft gekoppelt mit Tiefendaten, Kinect)

- **Gesten** sind nun als **Bewegung eines** oder **mehrerer Merkmalspunkte bestimmbar**

Anwendungen von Gestenerkennung

- Automotive

- Smart Home [12]

- Gaming [11]

- Smartphone [10]

- Sign Language Translation[13]

Anwendungen von Gestenerkennung

Gestenerkennung

- Vorteile der Gestenerkennung:
  - Intuitive und direkte Anwendung
  - Berührungslose Bedienung möglich
  - Bequeme Bedienung

- Herausforderungen:
  - Hohe Stabilität
  - Hohe Genauigkeit
  - Möglichst geringe Reaktionsverzögerung

Gestenerkennung mittels Körpersensoren

- **Vorteile:**
  - **Direkter Zugang zu den Merkmalsvektoren** durch Auslesen der Daten aus den Beschleunigungssensoren des Geräts
  - **Hohe Präzision** bei der Erkennung

- **Nachteil:**
  - **Nutzer muss Gerät am Körper tragen**
    - Vor allem bei Handschuhen/Controllern oft unbequem
  - Im Hinblick auf echte Gestenerkennung durch Maschinen, ist diese Vorgehensweise noch ein **Kompromiss**

Gestenerkennung durch optische Sensorik

- **Vorteile:**
  - Der **Benutzer** kann sich **relativ frei** im **Raum bewegen** und ist auch nicht eingeschränkt durch das Tragen von Controllern

- **Nachteile:**
  - **Bewegungsvektoren** der Merkmalspunkte sind **nicht direkt verfügbar**.
  - Es muss zunächst eine Erkennung der wesentlichen menschlichen Merkmalspunkte erfolgen
  - Dadurch ist die **Fehlerrate größer** als bei direktem Auslesen dieser Werte aus geeigneten Sensoren

Gestenerkennung durch optische Sensorik

- **Ablauf:**

Bildaufnahme

Vorverarbeitung

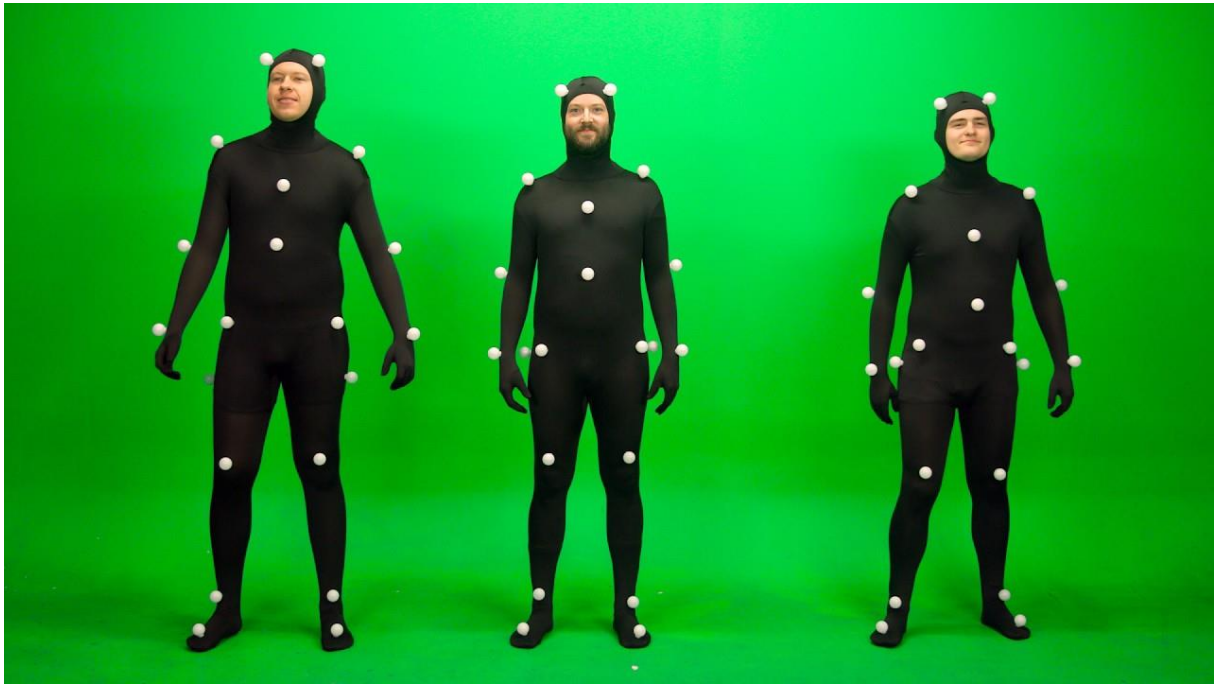
Segmentierung

Merkmalsextraktion

Klassifikation

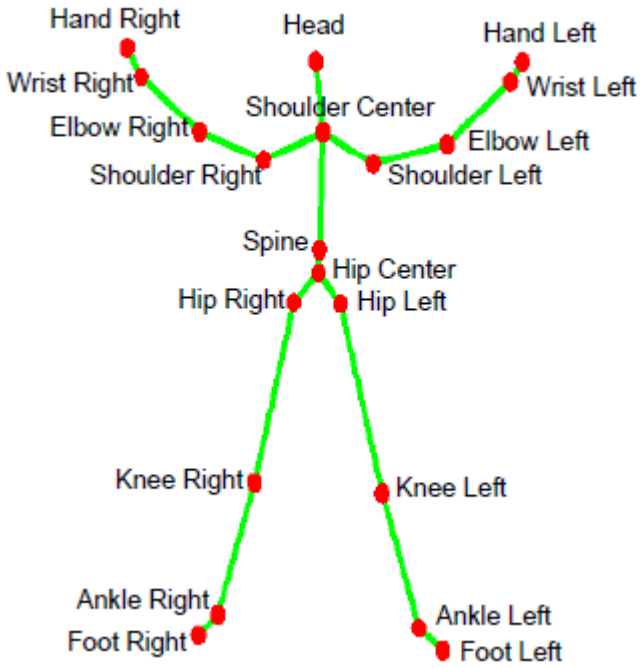
Motion Capture

- Das erste kamerabasierte Verfahren zur Aufzeichnung von menschlichen Bewegungen in einem automatisiert verarbeitbaren Format
- Zum Beispiel eingesetzt bei der Erstellung von Computeranimationen in Filmen
- Verwendung der Bewegungen zur Steuerung → Gestenerkennung



https://medium.com/@patricia.holloway80/3d-motion-capture-market-and-its-key-opportunities-and-challenges-c738ca87bcf

Gestenerkennung – Skelett mit Merkmalspunkten

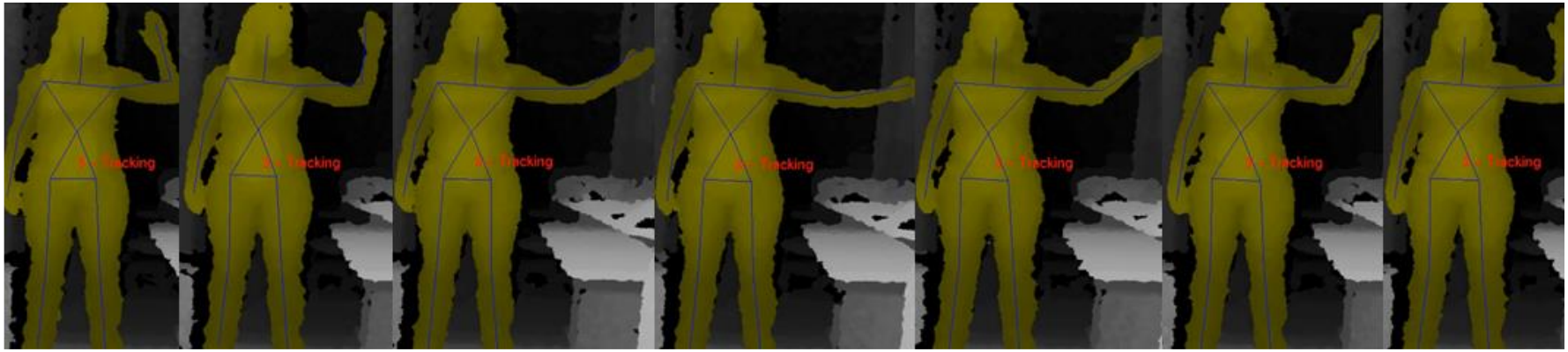


Gestenarten

- Teilkörpergesten
  - Gesten, die nur einen begrenzten Teil des Körpers betreffen, z.B. Handgesten
- Ganzkörpergesten
  - Gesten, die mit dem gesamten Körper ausgeführt werden, z.B. Kniebeuge
- Deiktische Gesten
  - Zeigegesten
- Manipulative Gesten
  - Interaktionsgeste mit einem Objekt
- Semaphorische Geste
  - Kommunikationsgeste, durch die Geste wird eine Nachricht codiert

Erkennen des Menschen

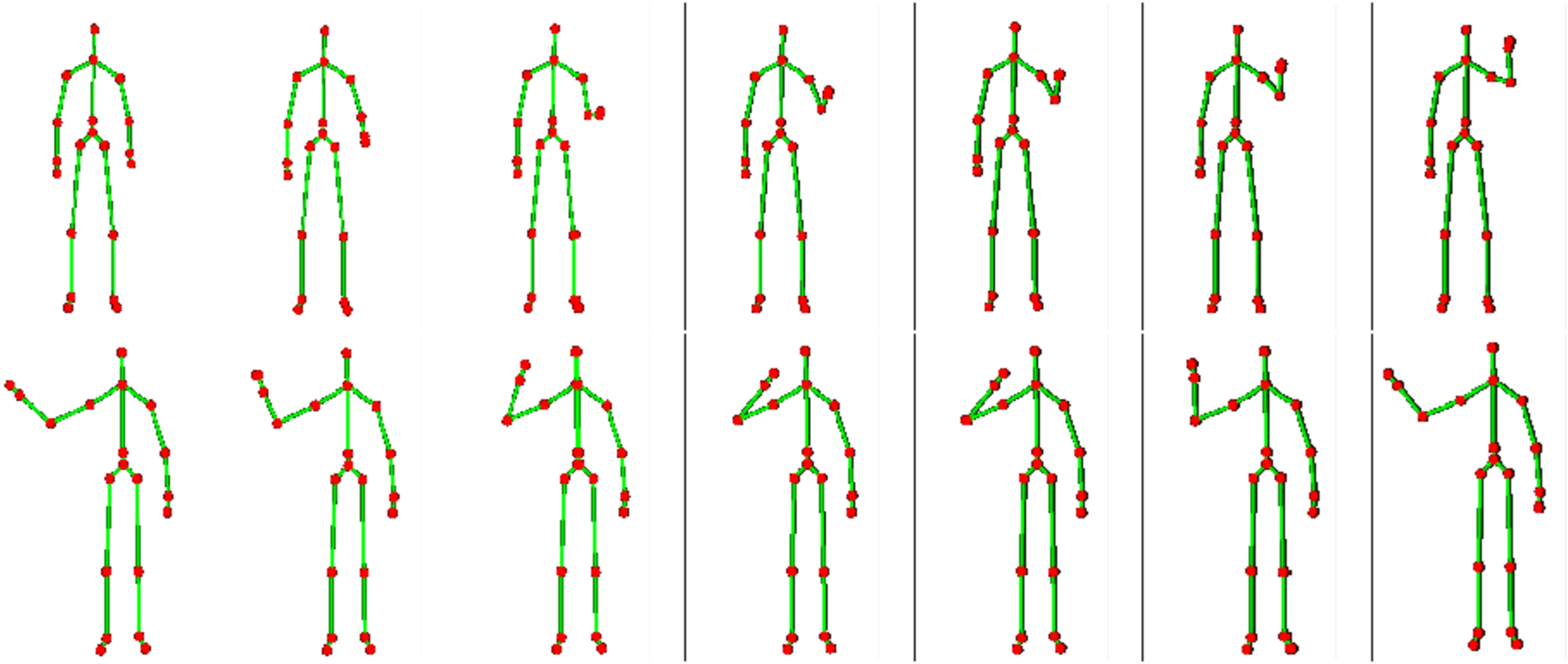
- Zunächst muss eine Erkennung des Menschen stattfinden
- Danach müssen auf Grundlage dieser Kontur die Gelenkpositionen zugeordnet werden
- Hierfür gibt es bereits unterstützende Software
- Mittels der relativen Gelenkbewegungen können dann Gesten erkannt und klassifiziert werden



Reyes et. Al., Feature Weighting in Dynamic Time Warping for Gesture Recognition in Depth Data

Beispielgesten

Rechte Hand wird hochgehoben:



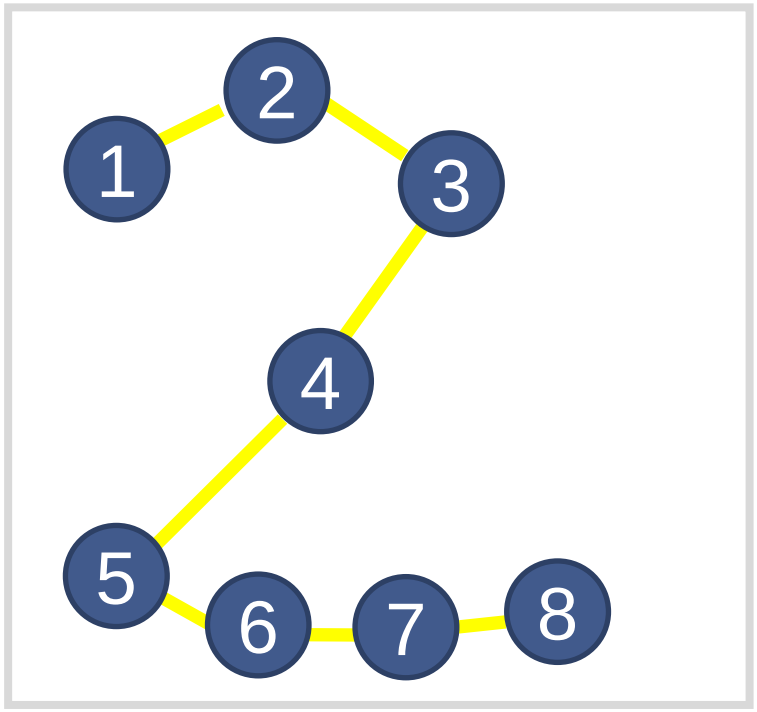
Linke Hand winkt:

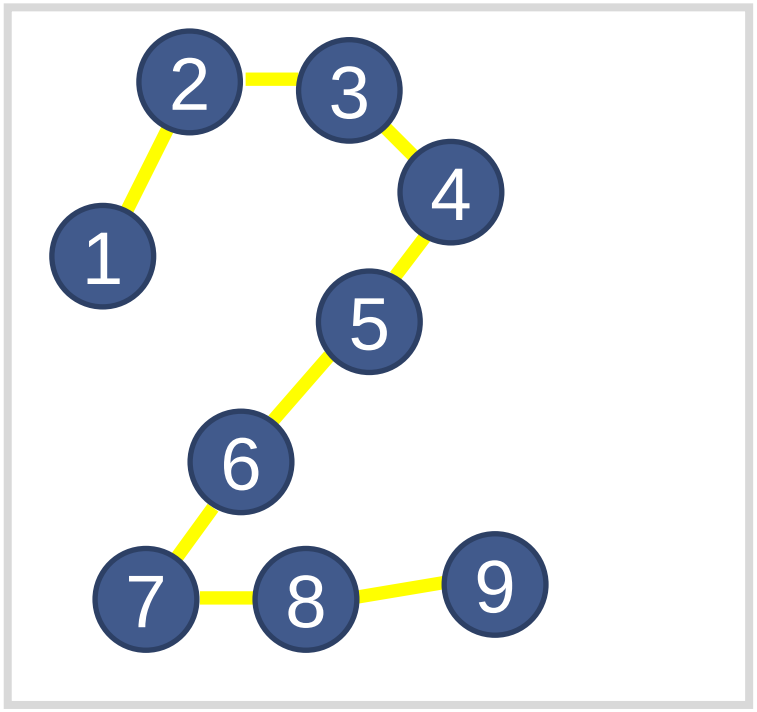
Problemstellung

- **Frage:** Warum funktioniert einfaches Matching der einzelnen Bildframes nicht zur Gestenerkennung?

  - Unterschiedliche **Samplingraten**
  - Unterschiedliche **Ausrichtung** des **Menschen** zur Kamera
  - Unterschiedliche **zeitliche Bewegungsausführung**

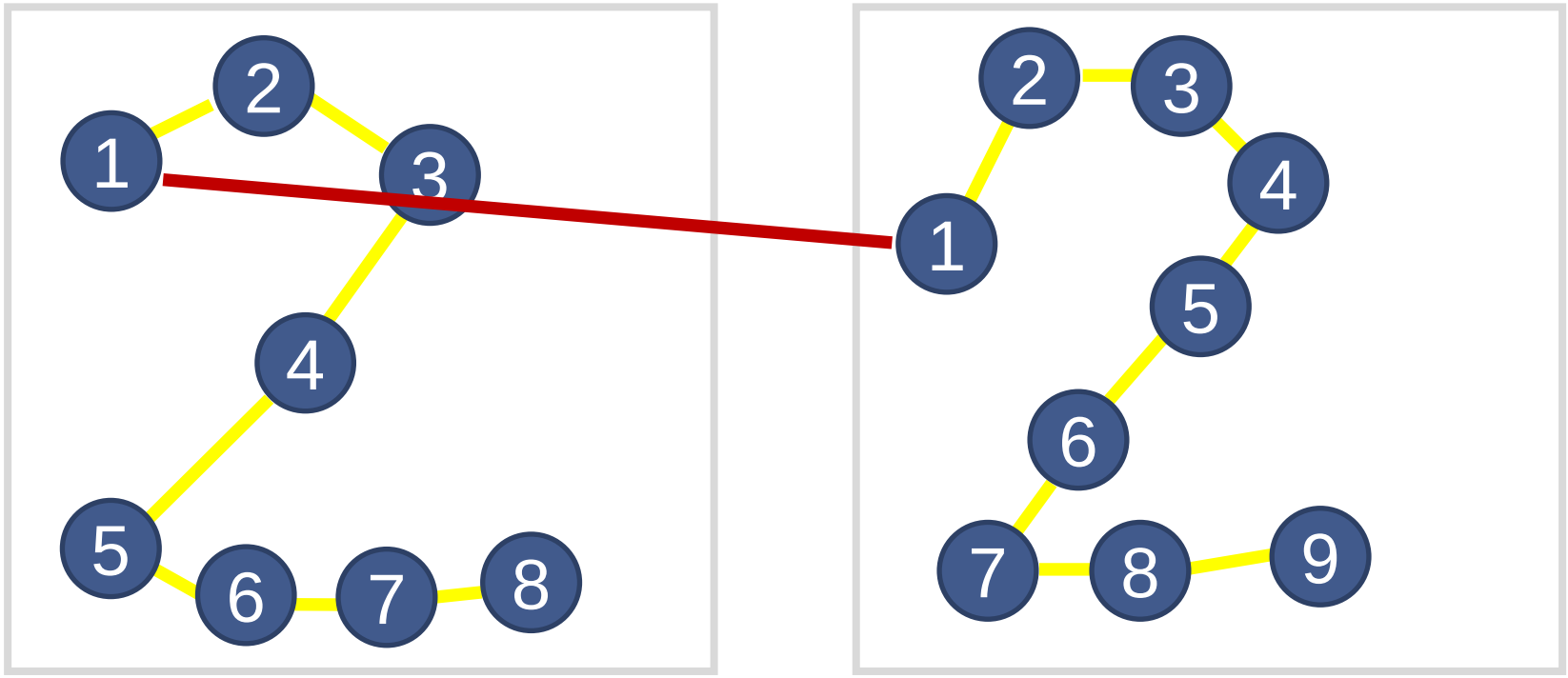
Vergleich von Bewegungen





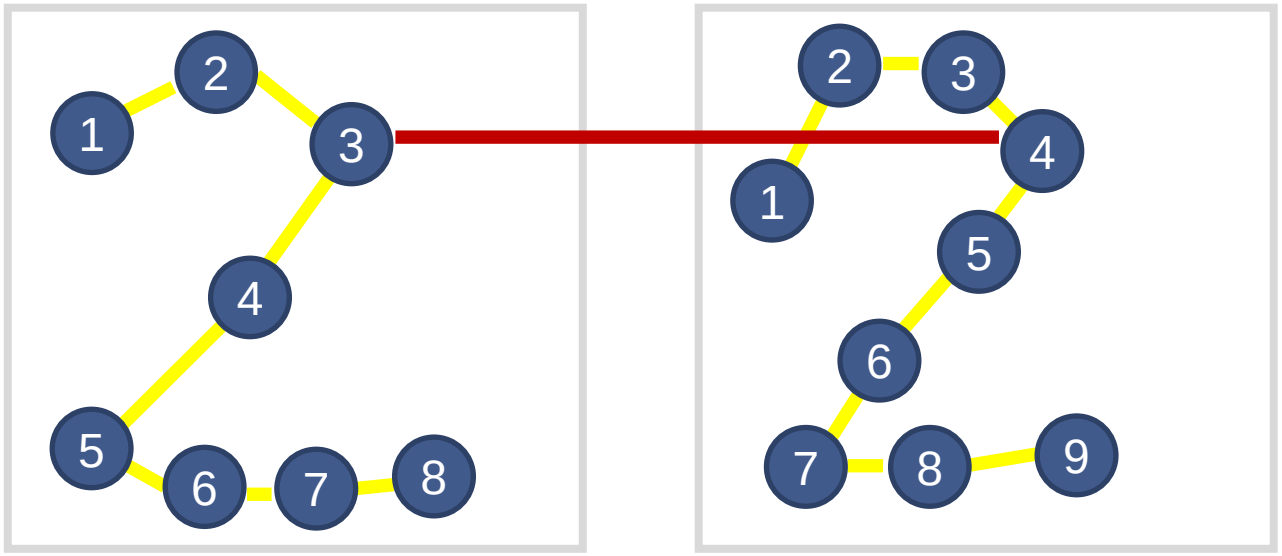
- Wir können Bewegung aufgrund der Handposition in einzelnen Frames erkennen
- Wie kann man diese vergleichen?

Vergleich von Bewegungen



- **Abgleich:**
  - $((1,1),(2,2),(2,3),(3,4),(4,5),(4,6),(5,7),(6,7),(7,8),(8,9))$
  - $((s_1,t_1),(s_2,t_2),\ldots,(s_p,t_p))$

Vergleich von Bewegungen



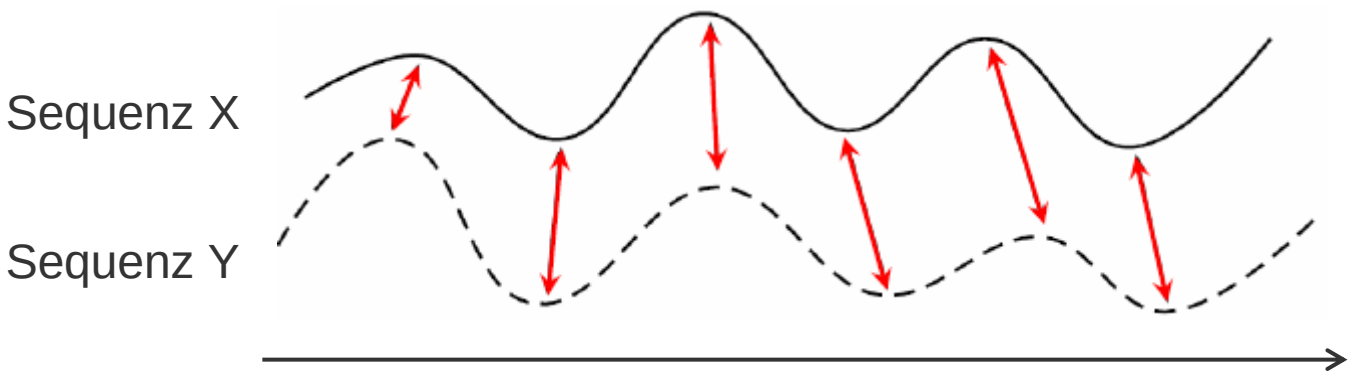
- **Abgleich:**
  - $((1,1),(2,2),(2,3),(3,4),(4,5),(4,6),(5,7),(6,7),(7,8),(8,9))$
  - $((s_1,t_1),(s_2,t_2),\dots,(s_p,t_p))$
- **Kosten für die Ausrichtung:**
  - $cost(s_1,t_1)+cost(s_2,t_2)+\cdots+cost(s_m,t_n)$
- **Beispiel:** $cost(si,t_i)$ → Euklidischer Abstand zwischen den Positionen
  - $cost(3,4)$ → Euklidischer Abstand zwischen M_3 und Q_4

Möglichkeiten zur Gestenerkennung

- Hidden Markov Models (HMM)
- Graphbasierte Ansätze
- Standard Dynamic Time Warping (DTW)
  - **Feature−Vektoren** $s(i), i = 1, 2, \dots I$ und $t(j), j = 1, 2, \dots, J$
  - **Warping-Pfad** $(i_0, j_0), (\mathrm{i}_1, \mathrm{j}_1), \dots, (\mathrm{i}_\mathrm{f}, \mathrm{j}_\mathrm{f})$
  - **Abstandsfunktion** $d(i, j)$**, z.B. euklidische Distanz**
  - **Gesamtkosten** $D = \sum_{k=0}^{f} d(i_k, j_k)$
- Gewichtetes Dynamic Time Warping
  - Gewichtete Abstandsfunktion, je nach Einfluss eines Merkmalspunktes auf die Geste

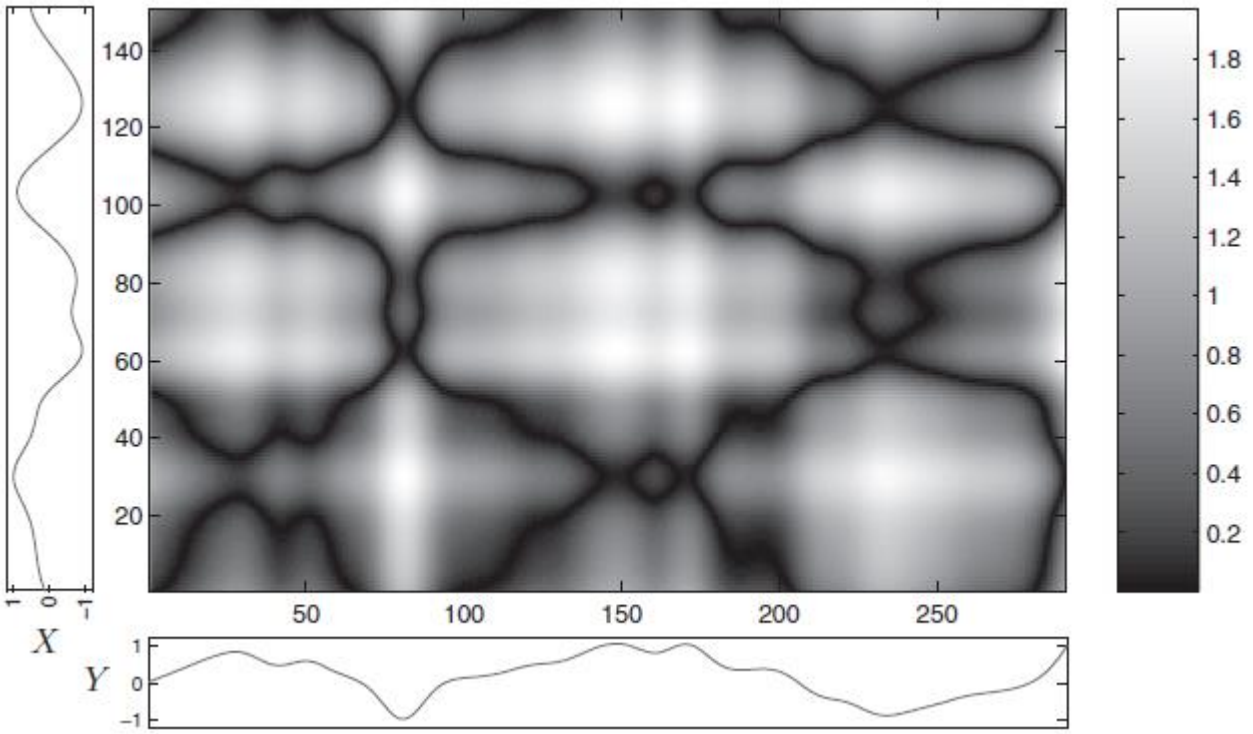
Dynamic Time Warping

- Anpassen bzw. Ausrichtung (engl. Alignment) zeitlicher oder geometrischer Sequenzen

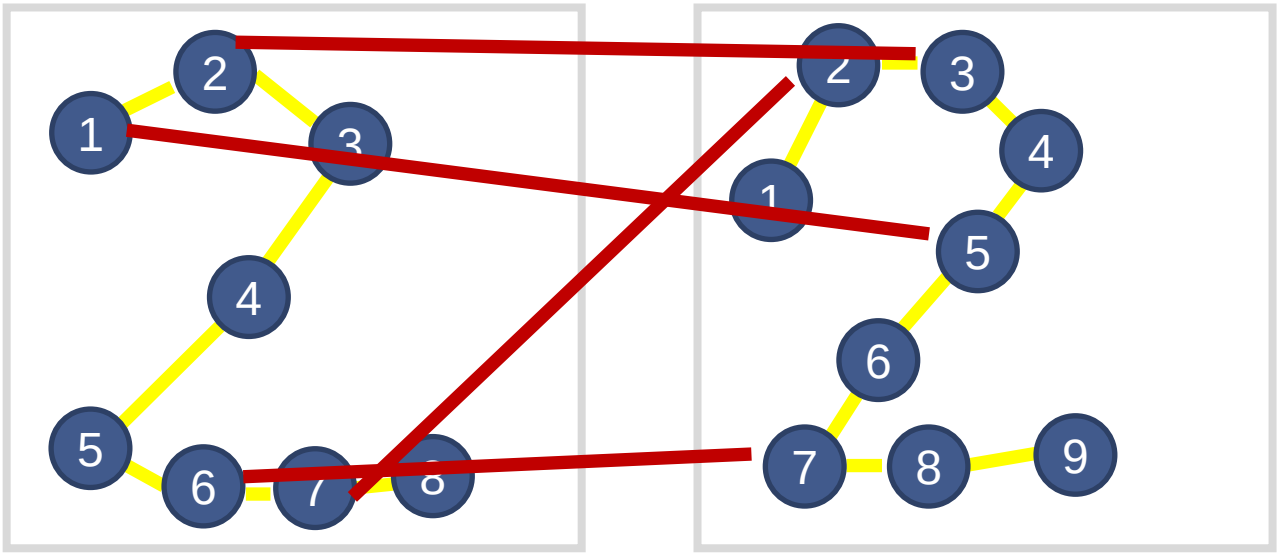


Kostenmatrix zweier reellwertiger Sequenzen X, Y

- Kostenmatrix $C^{n \times m} := c(x_i, y_j) = |x_i - y_j|$

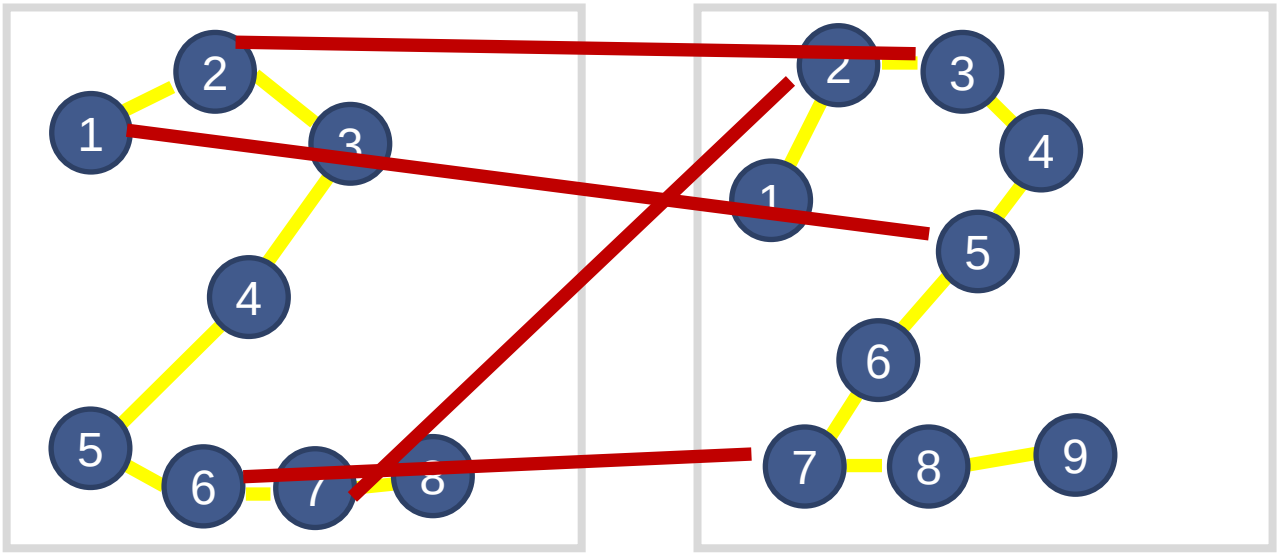


Vergleich von Bewegungen



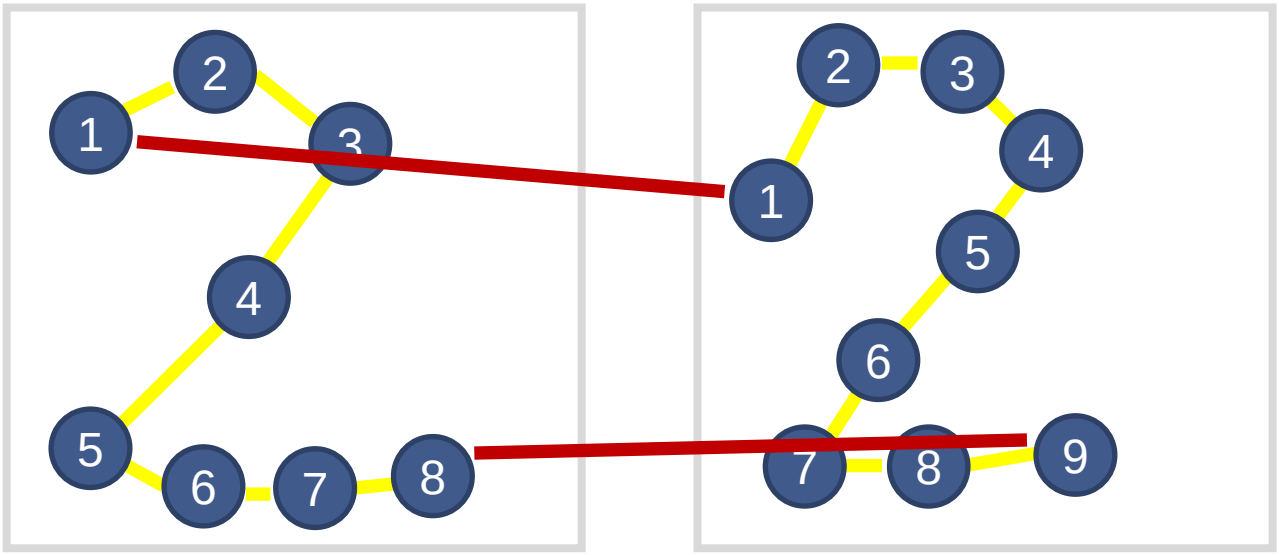
- **Abgleich:**
  - $((1,1),(2,2),(2,3),(3,4),(4,5),(4,6),(5,7),(6,7),(7,8),(8,9))$
  - $((s_1,t_1),(s_2,t_2),\ldots,(s_p,t_p))$
- **Regeln für Ausrichtung:**
  - Darf man $(1,5),(2,3),(6,7),(7,1)$ aneinander ausrichten?

Vergleich von Bewegungen



- **Abgleich:**
  - $((1,1),(2,2),(2,3),(3,4),(4,5),(4,6),(5,7),(6,7),(7,8),(8,9))$
  - $((s_1,t_1),(s_2,t_2),\ldots,(s_p,t_p))$
- **Regeln für Ausrichtung:**
  - Darf man $(1,5),(2,3),(6,7),(7,1)$ aneinander ausrichten?
  - Kommt darauf an, ob es Sinn für unsere Anwendung macht.

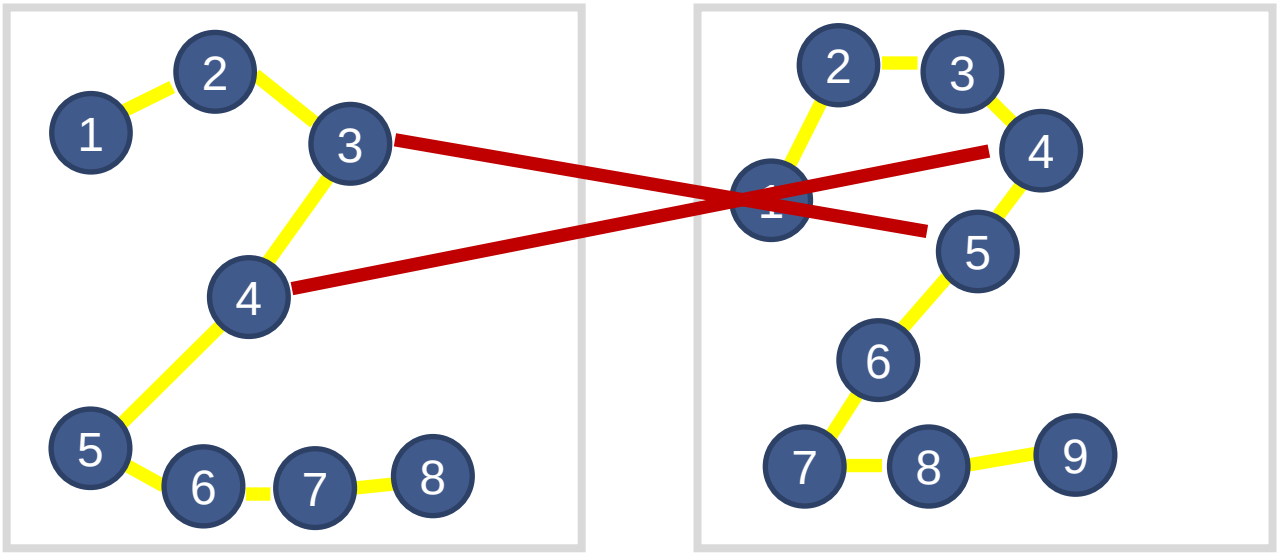
Vergleich von Bewegungen



- **Abgleich:**
  - $((1,1),(2,2),(2,3),(3,4),(4,5),(4,6),(5,7),(6,7),(7,8),(8,9))$
  - $((s_1,t_1),(s_2,t_2),\ldots,(s_p,t_p))$
- **Warping-Pfad-Regeln: Randbedingung**
  - $s_1 = 1,\ t_1 = 1$
  - $s_p = m$ → Länge der ersten Sequenz
  - $t_p = n$ → Länge der zweiten Sequenz

Erstes Element passt
Letztes Element passt

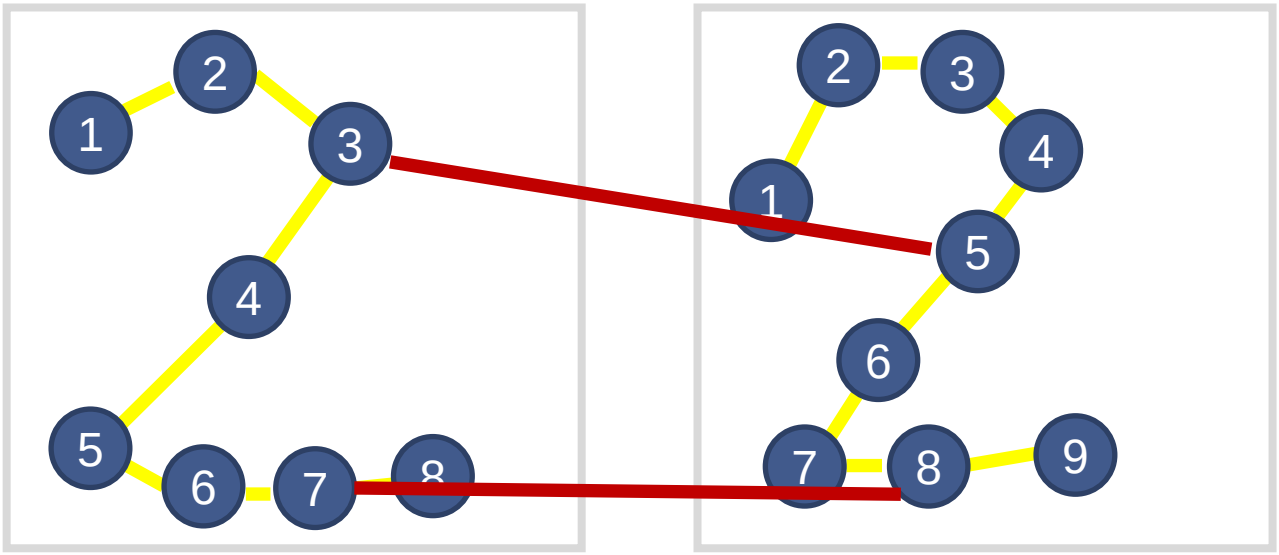
Vergleich von Bewegungen



- **Illegale Ausrichtung (verletzt Monotonität):**
  - $(\dots, (3,5), (4,3), \dots)$
  - $((s_1, t_1), (s_2, t_2), \dots, (s_p, t_p))$
- **Warping-Pfad-Regeln: Monotonität**
  - $0 \leq (s_{t+1} - s_t)$
  - $0 \leq (t_{t+1} - t_t)$

Keine Rückwärtsausrichtung

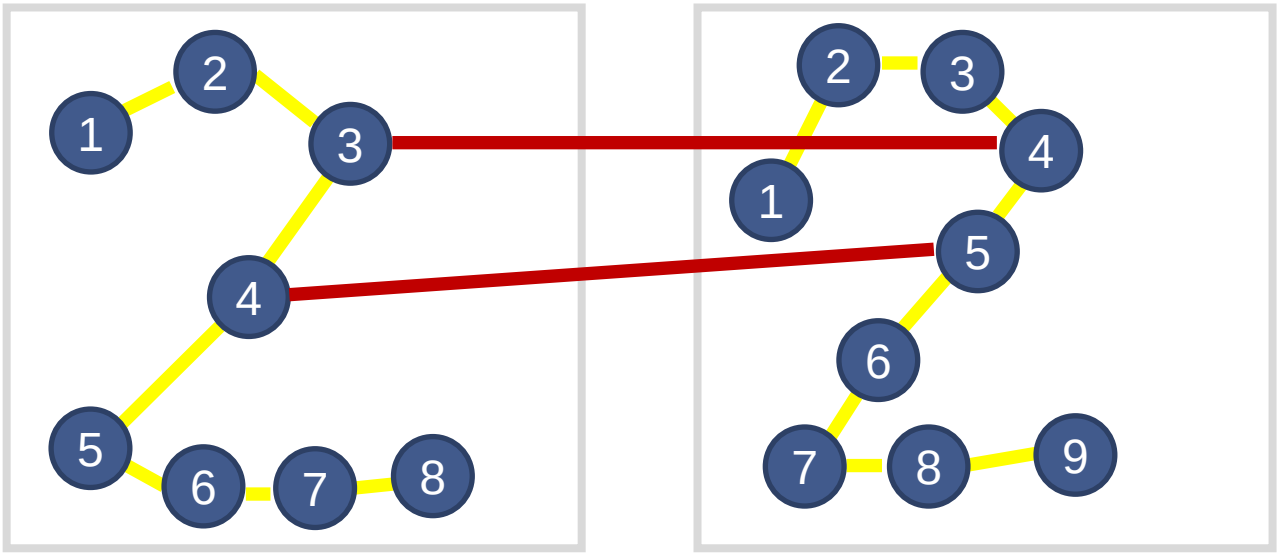
Vergleich von Bewegungen



- **Illegale Ausrichtung (Verletzt Kontinuität)**
  - $(\dots, (3,5), (6,7), \dots)$.
  - $((s_1, t_1), (s_2, t_2), \dots, (s_p, t_p))$
- **Warping-Pfad-Regeln: Kontinuität (Schrittweitenbedingung)**
  - $(s_{t+1} - s_t) \leq 1$
  - $(t_{t+1} - t_t) \leq 1$

Ausrichtung überspringt keine Elemente

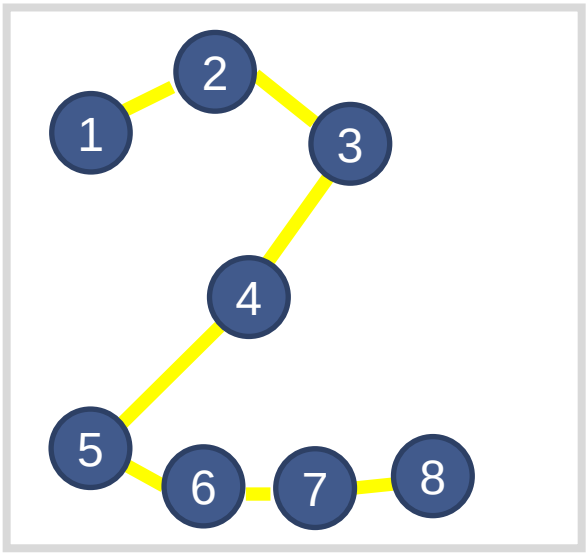
Vergleich von Bewegungen

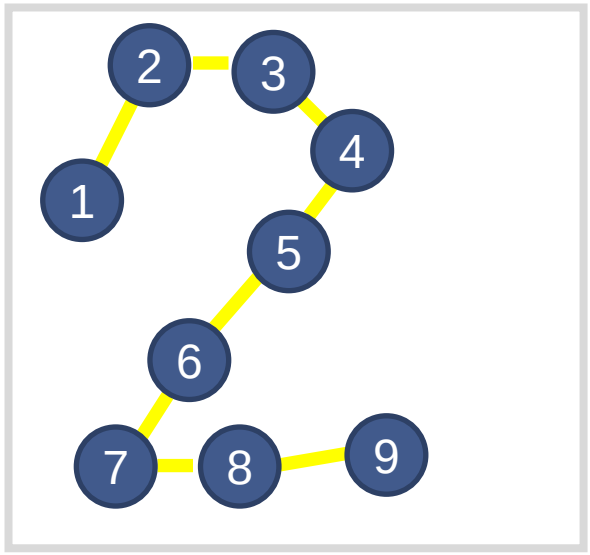


- **Abgleich:**
  - $((1,1),(2,2),(2,3),(3,4),(4,5),(4,6),(5,7),(6,7),(7,8),(8,9))$
  - $((s_1,t_1),(s_2,t_2),\ldots,(s_p,t_p))$
- **Dynamic-Time-Warping-Regeln: Monotonität, Kontinuität (Schrittweitenbed.)**
  - $0 \leq (s_{t+1} - s_t) \leq 1$
  - $0 \leq (t_{t+1} - t_t) \leq 1$

Ausrichtung überspringt keine Elemente und geht nicht rückwärts

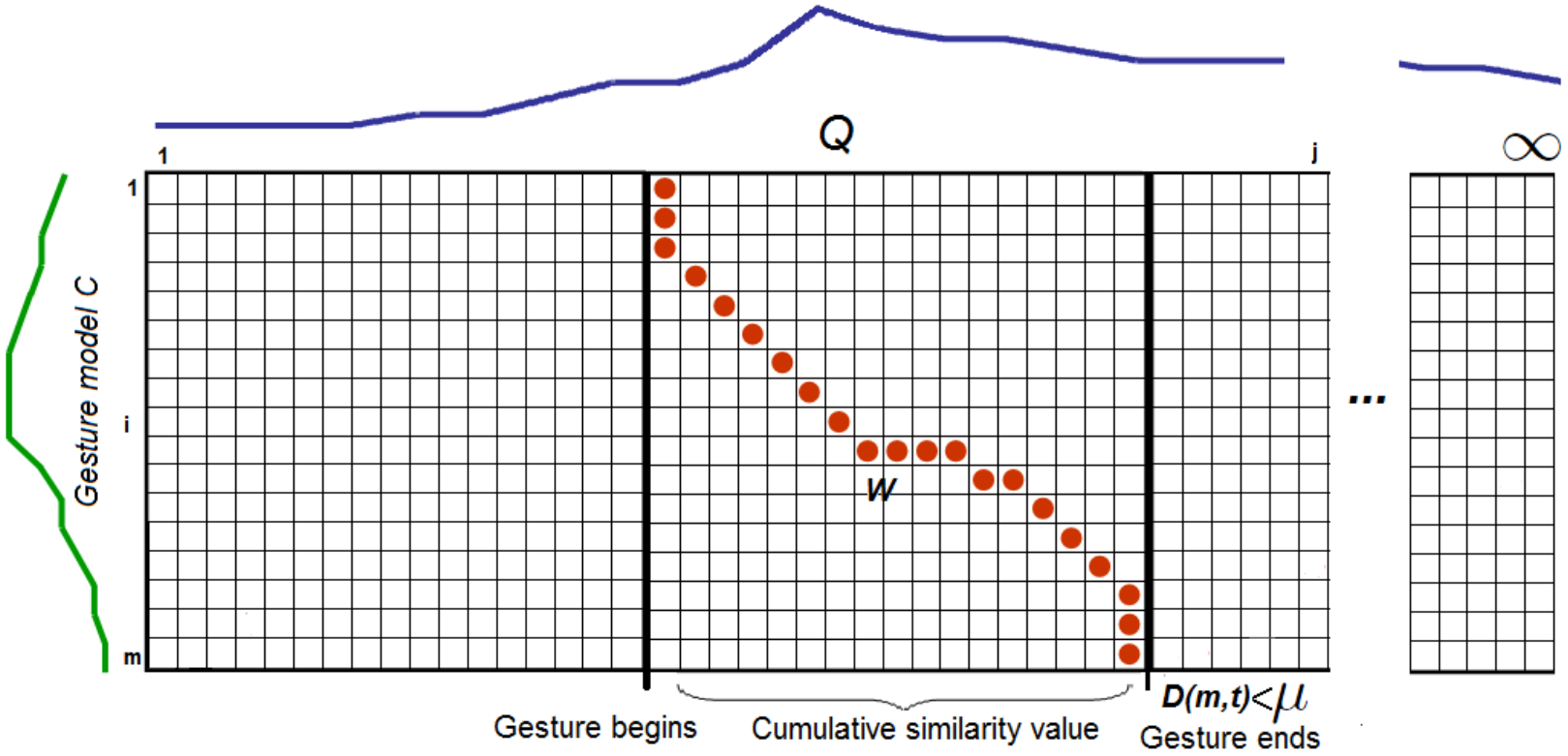
Dynamic Time Warping





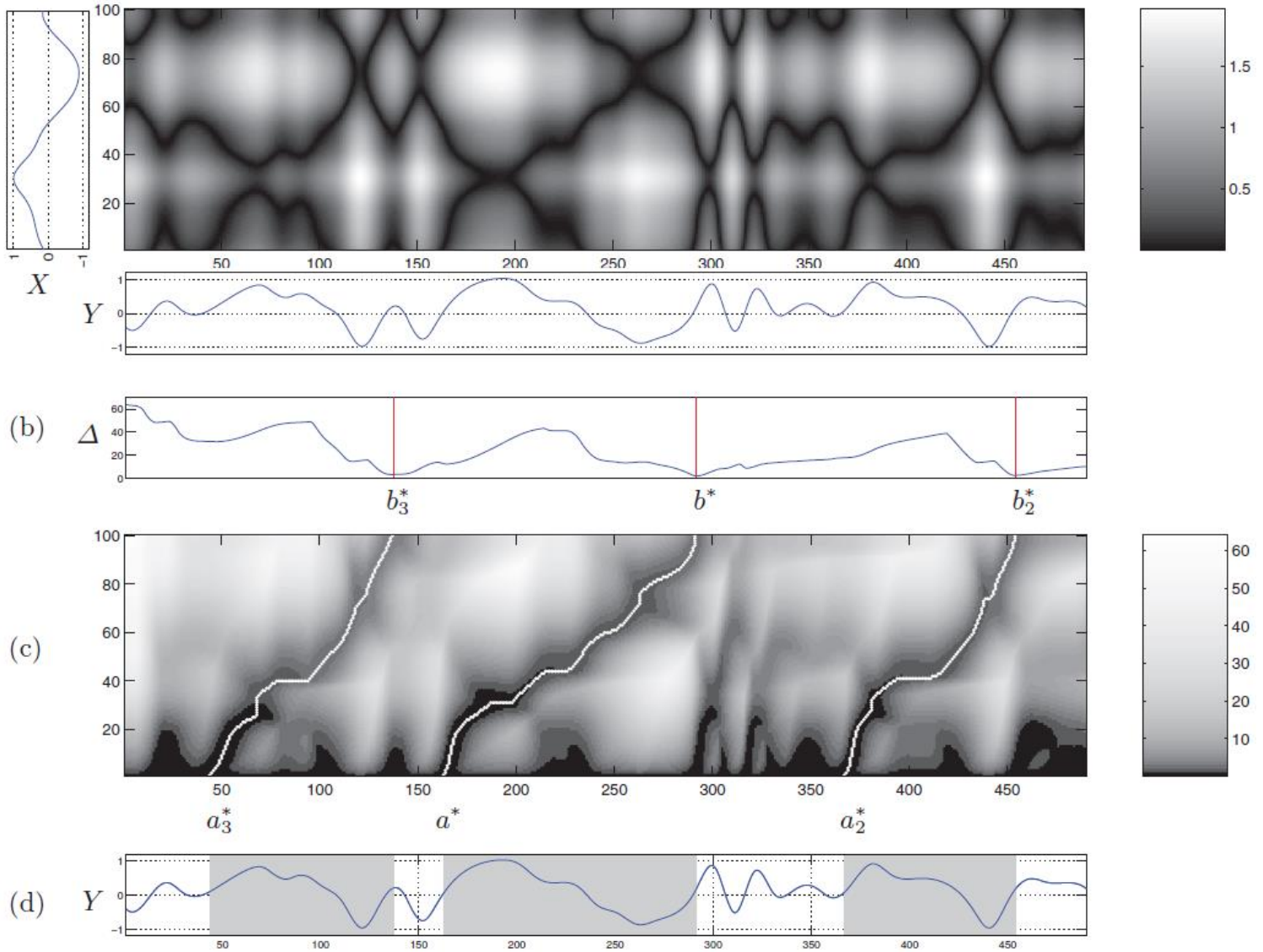
- Dynamic Time Warping (DTW) ist ein Distanzmass zwischen Sequenzen von Punkten
- Die DTW Distanz sind die Kosten für eine **optimale** Ausrichtung zwischen zwei Bewegungen
  - Die Ausrichtung muss die DTW Regeln aus den vorherigen Folien berücksichtigen

Gestenerkennung (Beispiel)



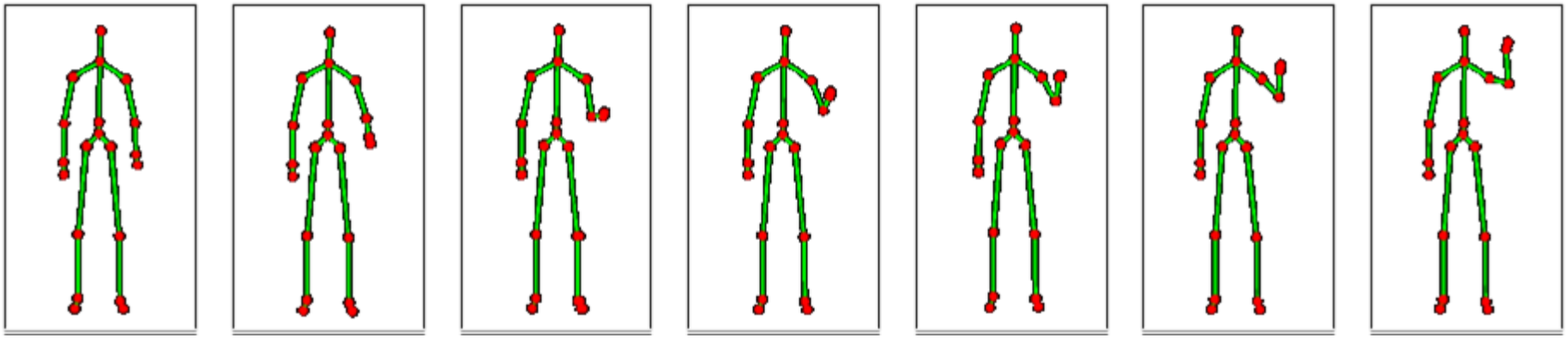
Reyes et. al., Feature Weighting in Dynamic Time Warping for Gesture Recognition in Depth Data

DTW über Teilsequenzen



Gewichtete DTW

- Für jede Geste spielen die einzelnen Merkmalspunkte unterschiedlich große Rollen
- Daher: Definiere die Kostenfunktion so, dass eine Gewichtung enthalten ist
- Z.B. $w_j = \sum_{n=2}^{N} Dist^j(s_n, s_{n-1})$ für j = Merkmalspunkt und n = Framenummer.
- Dies wird dann für jede Geste einzeln ermittelt.



Beispiel: Gestenerkennung mit gewichteten DTW-Verfahren

	R push up	L push up	R pull down	L pull down	R swipe L	L swipe R
R push up	65	0	0	30	5	0
L push up	15	40	0	0	45	0
R pull down	0	0	85	15	0	0
L pull down	15	0	0	75	10	0
R swipe L	0	0	0	30	70	0
L swipe R	15	0	0	5	55	25

Trefferrate mit ungewichteten DTW

	R push up	L push up	R pull down	L pull down	R swipe L	L swipe R
R push up	100	0	0	0	0	0
L push up	0	100	0	0	0	0
R pull down	0	0	100	0	0	0
L pull down	0	0	0	85	15	0
R swipe L	0	0	0	0	100	0
L swipe R	0	0	0	0	5	95

Trefferrate mit gewichteten DTW

Klassifikation

- Gehört ein aufgenommener Merkmalsvektor zu einer Gestenklasse?
- Kostenfunktion einer Klasse (z.B. „Hand hoch“)

$$d(i_k, j_k) = \sum_j Dist^j\left(r_{i_k}, t_{j_k}\right) w_j$$

- Danach ganz normale DTW-Kostenmatrix und Warping-Pfad bestimmen
- Existiert ein Pfad, dessen Gesamtkosten unterhalb eines definierten Werts liegen, wird das Muster als erkannt gewertet