Bundesministerium
für Arbeit und Soziales

Sign language coding, 3D behavior data ... and ANVIL

Michael Kipp

DFKI
Embodied Agents Research Group
Cluster of Excellence
Multimodal Computing and Interaction
Saarland University

OTIM / ILIKS Workshop
24 May 2011
LPL, Aix-en-Provence

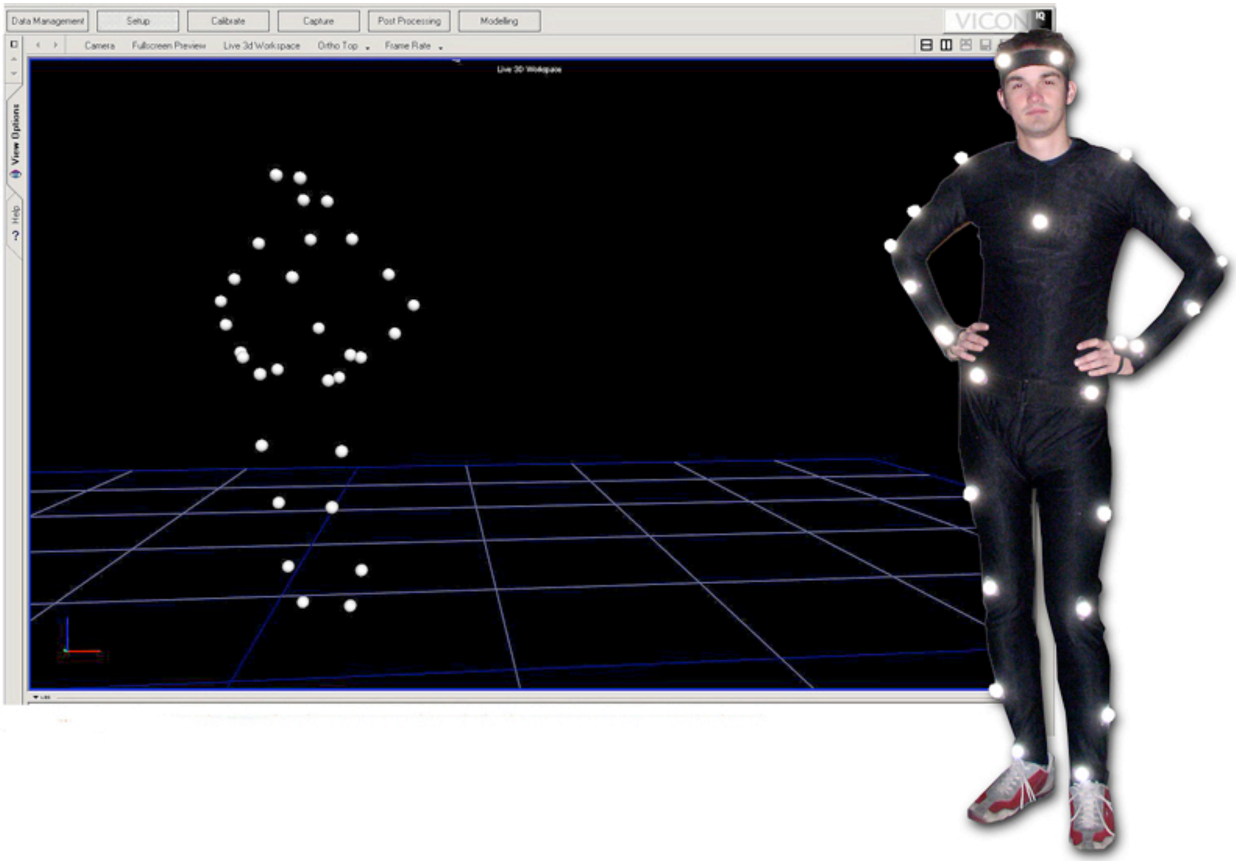
Joint work with:
Alexis Heloir, DFKI
Quan Nguyen, DFKI
Michael Neff, UC Davis

Overview

- Multimodal corpora for animation
- Sign language avatars
- ANVIL

Announcements:
Workshop on Multimocal Corpora: Taking Stock and Roadmapping the Future
held in conjunction with ICMI-2011 (Heylen, Paggio, Kipp), 18 November
Watch www.multimodal-corpora.org

Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT),
with ACM ASSET 2011, Dundee, UK.
Watch http://embots.dfki.de/SLTAT



Corpora for Animation

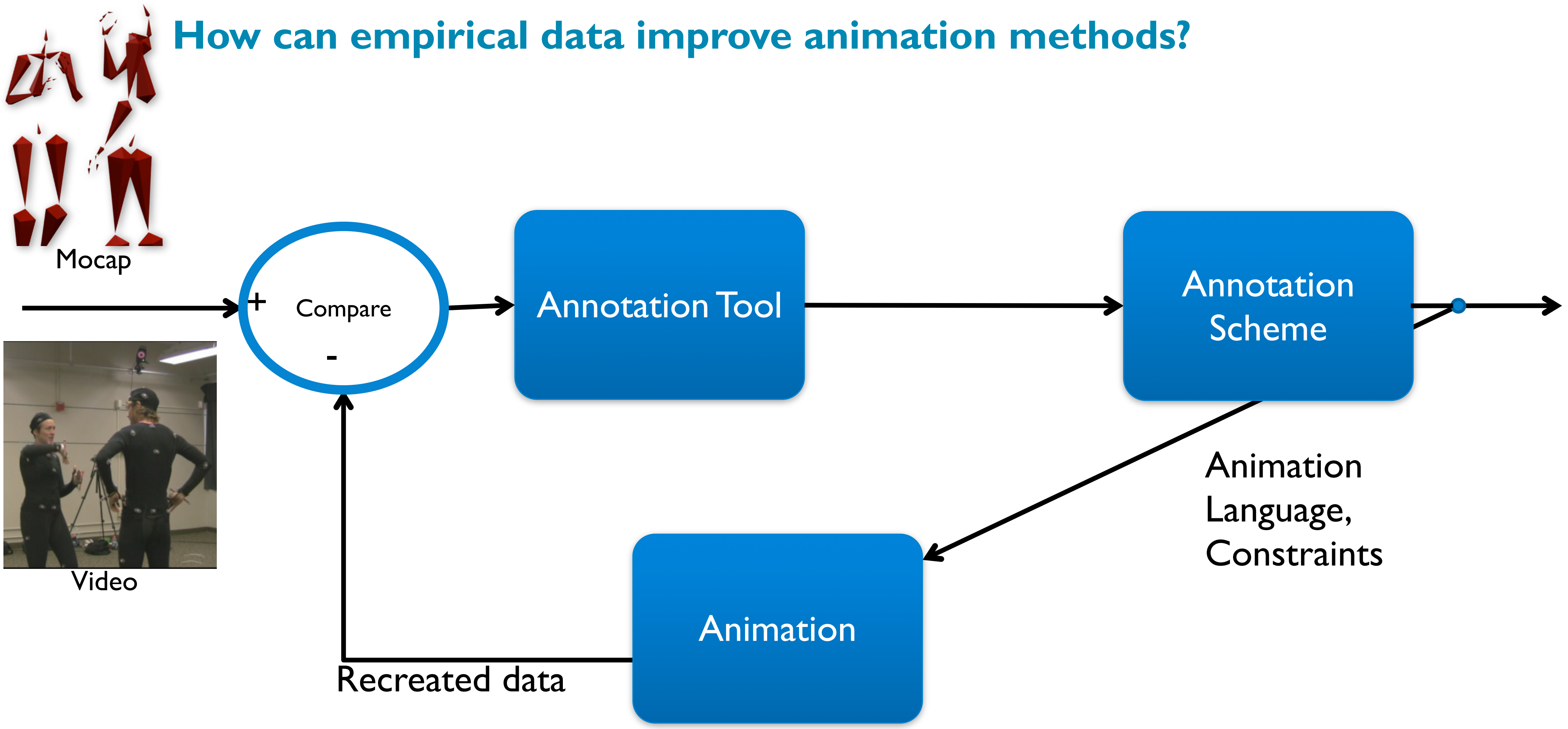
What Can be Learned from Motion Data ?

- Ambient movements (Egges et al. 2005)
- Balance control (Neff et al. 2009)
- Motion graphs (Kovar et al. 2002)
- Recreation of gesture from annotations (Kipp et al. 2008)

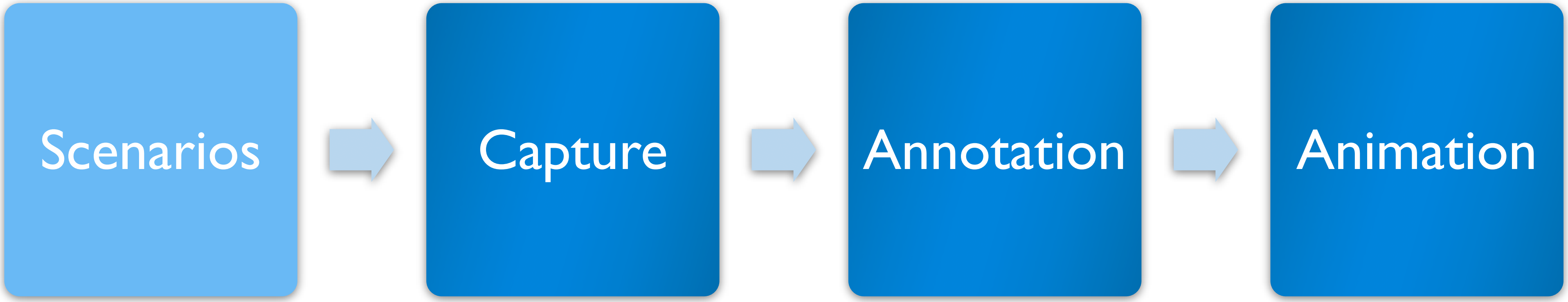
- We are interested in building generative models of communicative gestures (in dyadic conversations)

The analysis and synthesis loop

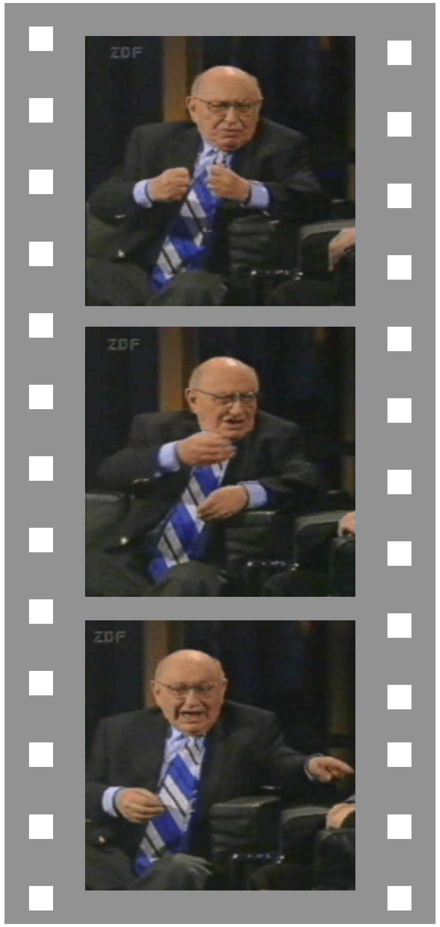
How can empirical data improve animation methods?



Data Acquisition and Processing Pipeline



Video Corpus
(Neff et al. 2008)



Mocap/Video Corpus
(Heloir et al. 2010)

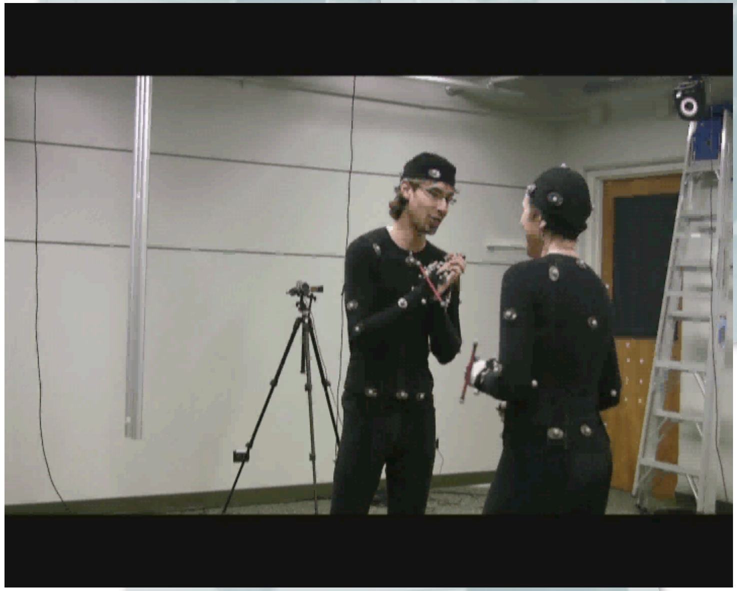
Recent Capture Session (UC Davis)

- Improvised acting
- 19 dyadic scenarios (two friends meet …)
- Status high/low + agree/disagree
- Proxemic behavior + NVB synchronization in dyads

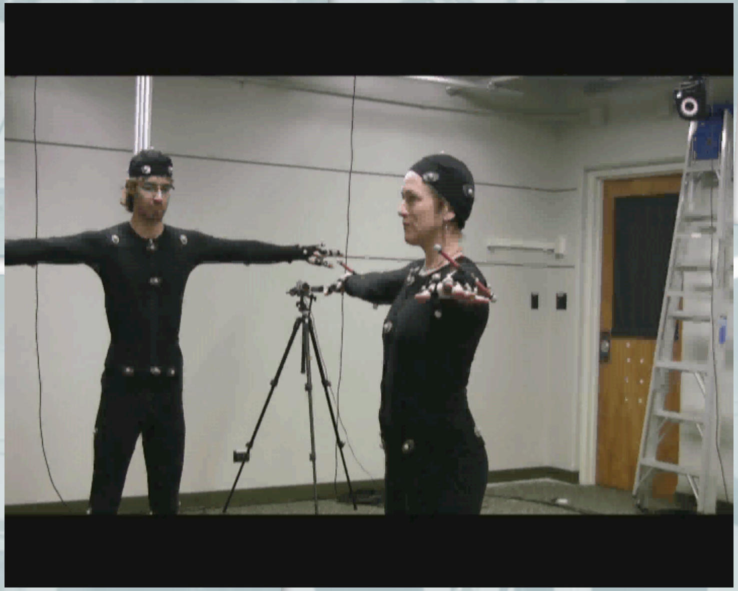
Status

Liking

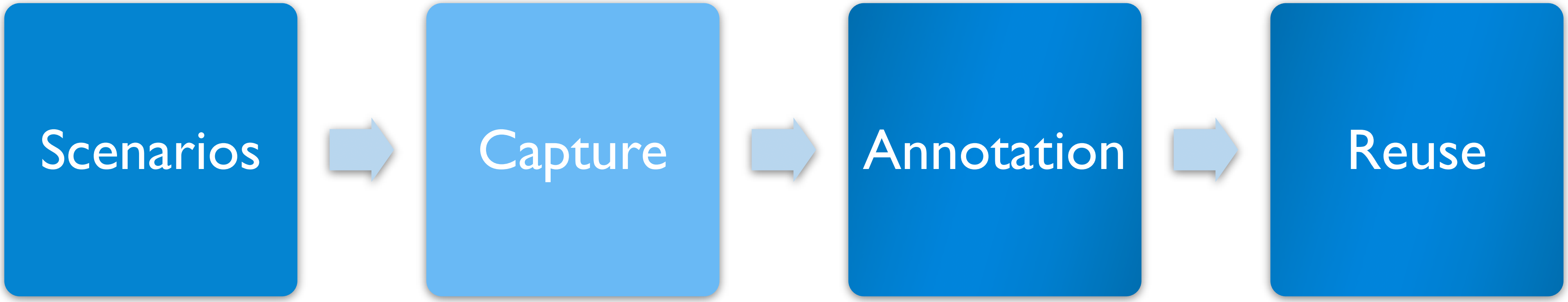
Example: Two People Meet



They like each other



They dislike each other



Scenarios

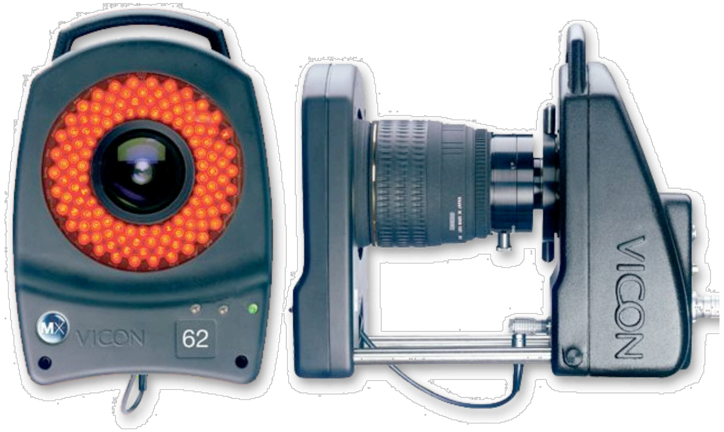
Capture

Annotation

Reuse

Technical setup

- Optical Motion Capture
  - Vicon MX 40
  - 12 Cameras

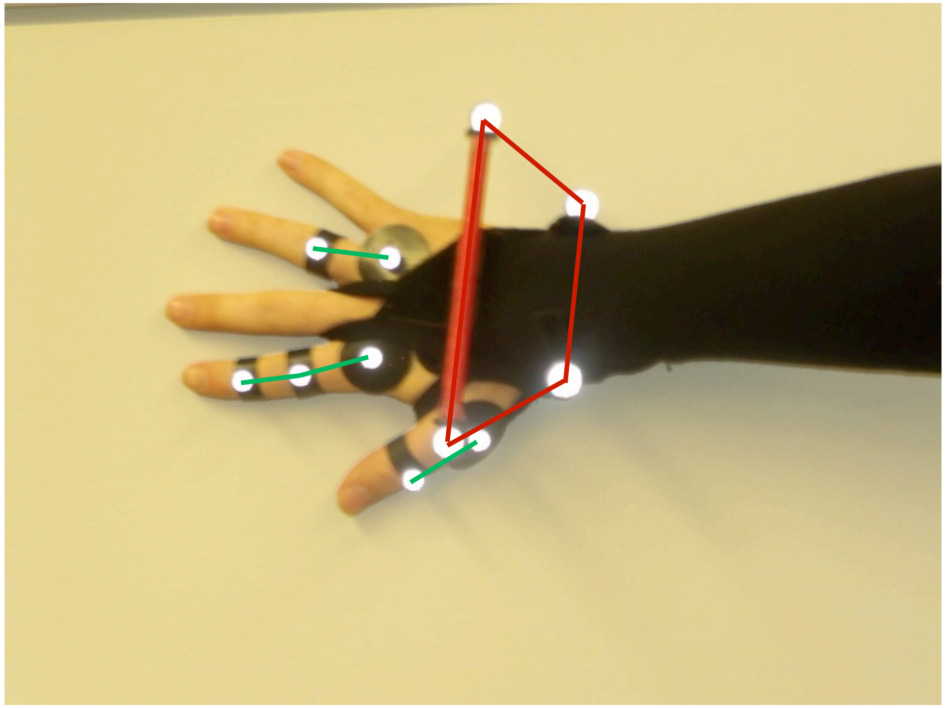


- Video recorder (x2)
  - HD

- Camera mounted microphones (not recommended)

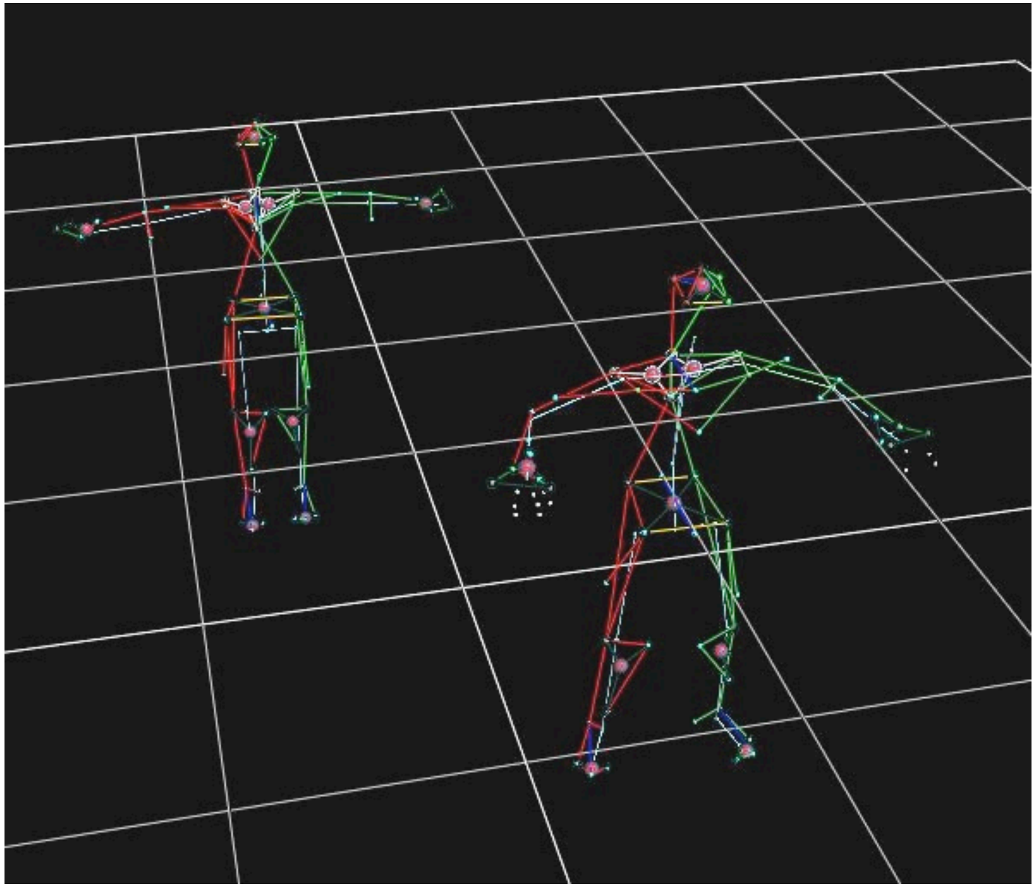
Capturing Handshape

- Occlusions are frequent between fingers
- Impossible to record motion for all fingers
- We used a reduced set of markers
  - index finger
  - thumb
  - „rest“
- Similar to (Chang et al., 2007)



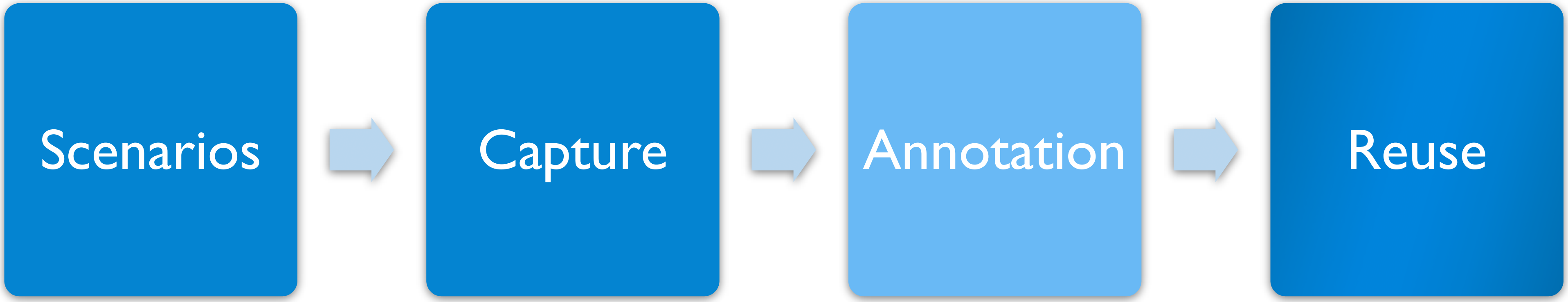
Reconstruction

- From marker clouds to skeleton
- Semi automated process
- Significant manual processing required
  - labeling correction (occlusion, confusion, mainly hands)
- Postprocessing work: 1 : 40



Result

for instance: BVH files



Scenarios

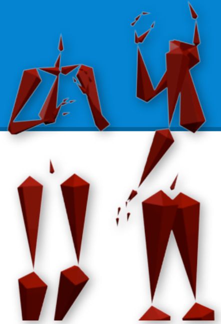
Capture

Annotation

Reuse

Annotation: From Raw Data to Symbolic Representation

Raw Data



Mocap

Video

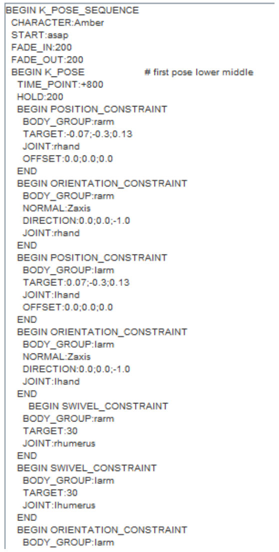
- fine-grained
- no „meaning“
- difficult to manipulate
- highly realistic

Annotation scheme
Annotation tool

- **Gesture Phases**
- **Gesture Phrases**
- **Gesture Units**
- **Handedness**
- **Wrist Trajectories**
- **Gaze direction**
- **Head/Body orientation**

- **Visualization tools**
- **Automated feature detection**

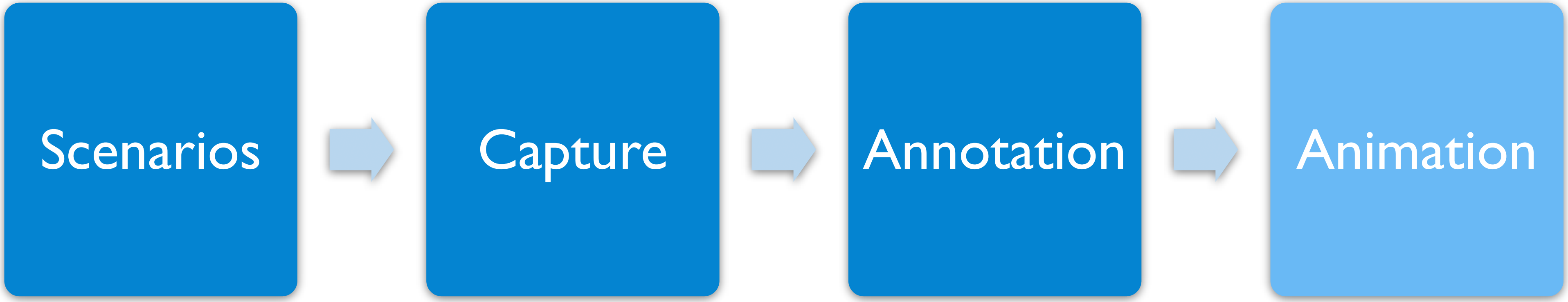
Symbolic Representation



```
BEGIN K_POSE_SEQUENCE
 CHARACTER:Amber
 START:asap
 FADE_IN:200
 FADE_OUT:200
 BEGIN K_POSE                    # first pose lower middle
  TIME_POINT:+800
  HOLD:200
  BEGIN POSITION_CONSTRAINT
   BODY_GROUP:rarm
   TARGET:-0.07;-0.3;0.13
   JOINT:rhand
   OFFSET:0.0;0.0;0.0
  END
  BEGIN ORIENTATION_CONSTRAINT
   BODY_GROUP:rarm
   NORMAL:Zaxis
   DIRECTION:0.0;0.0;-1.0
   JOINT:rhand
  END
  BEGIN POSITION_CONSTRAINT
   BODY_GROUP:larm
   TARGET:0.07;-0.3;0.13
   JOINT:lhand
   OFFSET:0.0;0.0;0.0
  END
  BEGIN ORIENTATION_CONSTRAINT
   BODY_GROUP:larm
   NORMAL:Zaxis
   DIRECTION:0.0;0.0;-1.0
   JOINT:lhand
  END
    BEGIN SWIVEL_CONSTRAINT
   BODY_GROUP:rarm
   TARGET:30
   JOINT:rhumerus
  END
  BEGIN SWIVEL_CONSTRAINT
   BODY_GROUP:larm
   TARGET:30
   JOINT:lhumerus
  END
  BEGIN ORIENTATION_CONSTRAINT
   BODY_GROUP:larm
```

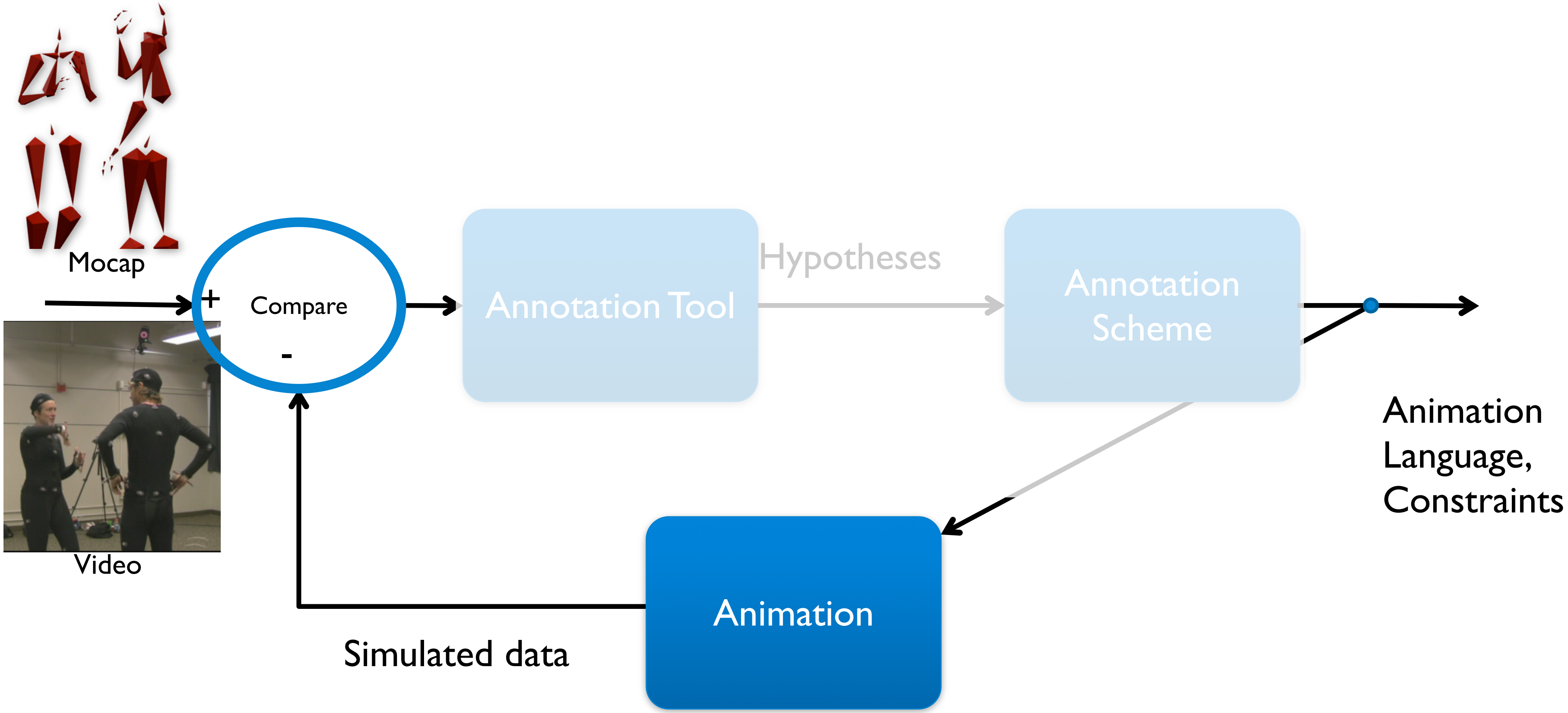
Animation Language, set of constraints

- compact
- meaningful
- easy to manipulate
- realistic ???

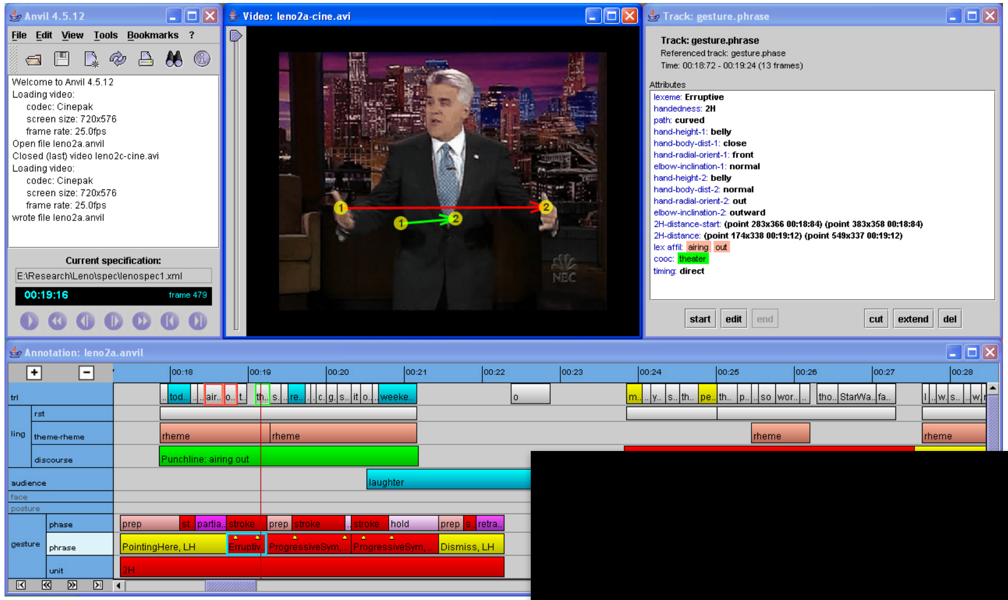


Scenarios → Capture → Annotation → Animation

Validation by Recreation



Just „recreated“ (some call it „reanimated“)



Welcome to Anvil 4.5.12
Loading video:
    codec: Cinepak
    screen size: 720x576
    frame rate: 25.0fps
Open file leno2a.anvil
Closed (last) video leno2c-cine.avi
Loading video:
    codec: Cinepak
    screen size: 720x576
    frame rate: 25.0fps
wrote file leno2a.anvil

Track: gesture.phrase
Referenced track: gesture.phase
Time: 00:18:72 - 00:19:24 (13 frames)

Attributes

lexeme: Erruptive
handedness: 2H
path: curved
hand-height-1: belly
hand-body-dist-1: close
hand-radial-orient-1: front
elbow-inclination-1: normal
hand-height-2: belly
hand-body-dist-2: normal
hand-radial-orient-2: out
elbow-inclination-2: outward
2H-distance-start: (point 283x366 00:18:84) (point 383x358 00:18:84)
2H-distance: (point 174x338 00:19:12) (point 549x337 00:19:12)
lex affil: airing out
cooc: theater
timing: direct

Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style

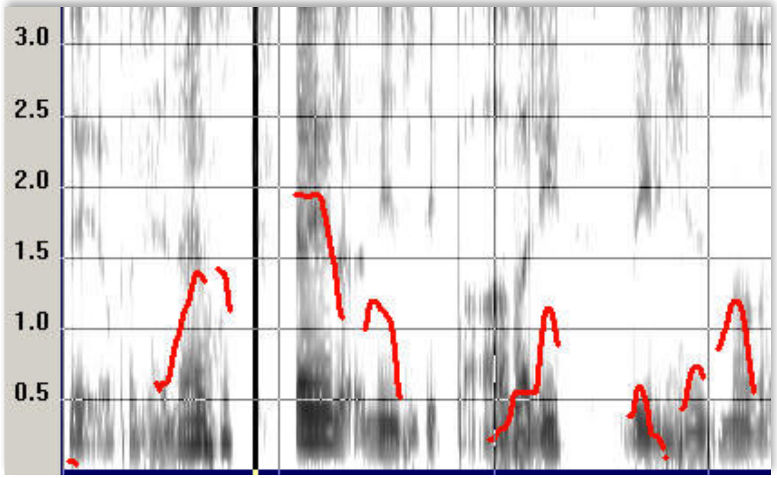
[Neff et al. 2008] ACM Transactions on Graphics

[Kipp et al. 2007] JLRE (coding scheme)

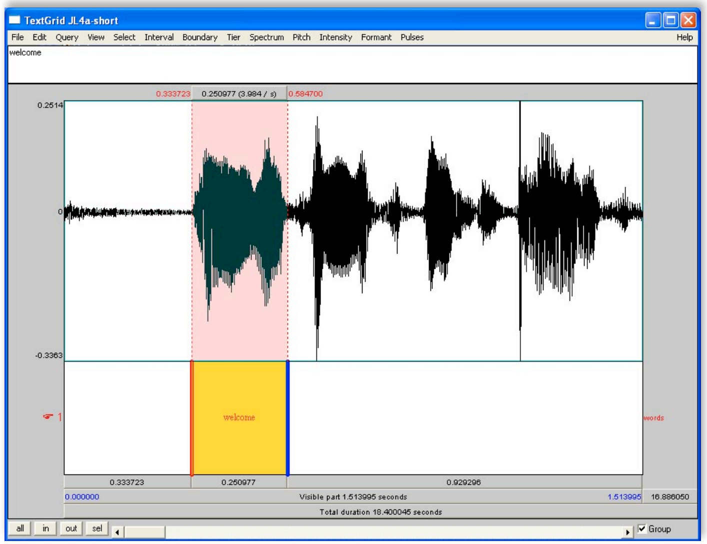
- Question: Longer **G-Units** => more natural?
- Hypothesis: Yes
- Experiment
  - **G version**: synthesized
  - **S version**: manipulation (made singular)
- Results
  - **G version**:
    - **more natural** p < .01
    - **more friendly** p < .001
    - **more trustworthy** p < .001
  - **S version**:
    - **more nervous** p < .001

[Kipp et al. 2007] IVA 07, Best paper

Why Motion Capture?

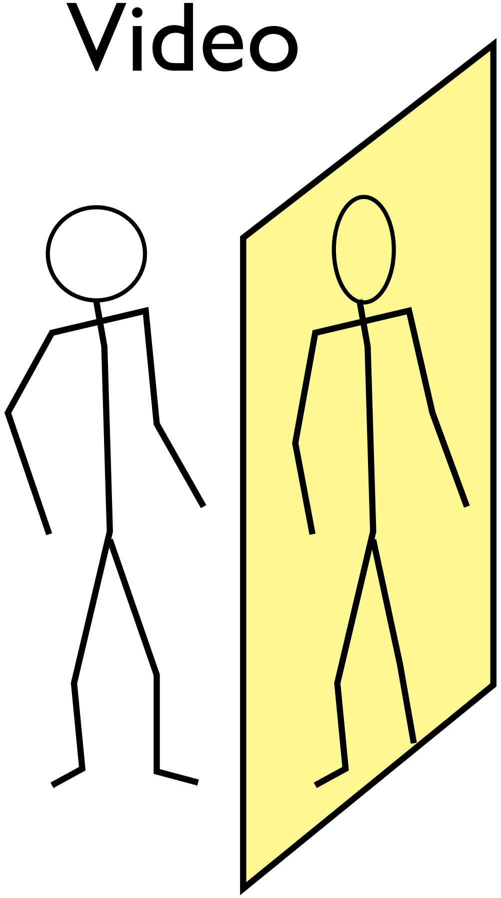


signal
=>
spectrogram
waveform
intensity
pitch
=>
segmentation
categorization


TextGrid JL4a-short
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
welcome
0.333723
0.250977 (3.984 / s)
0.584700
0.2514
0
-0.3363
1
welcome
words
0.333723
0.250977
0.929296
0.000000
Visible part 1.513995 seconds
1.513995
16.886050
Total duration 18.400045 seconds
all
in
out
sel
☑ Group

vs.

Video



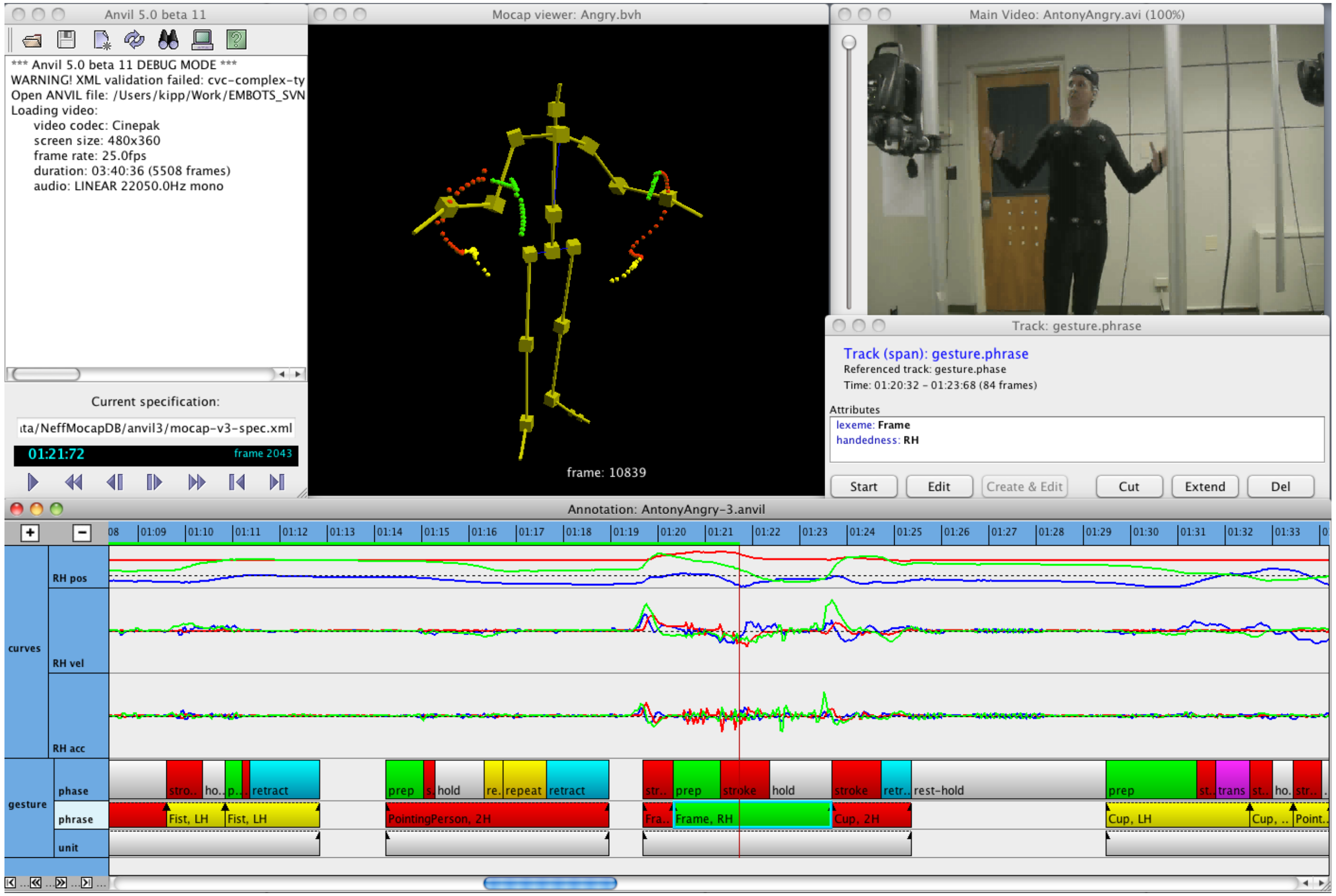
projection onto a 2D screen
merged with background
=> degraded, noisy signal
=> seg. + cat.

Why Motion Capture?

- **Objective measures**
  - ➡ speed / velocity (acceleration)
    - ‣ rhythm analysis, interpersonal synchrony, correlation with intonation
  - ➡ shape of the gesture
    - ‣ trajectory, motion contour
  - ➡ hand location in gesture space (automatic/robust)
  - ➡ direction of a gesture (vector)
  - ➡ distance and orientation of interlocutors (proxemics)
- **Viewing**
  - ➡ Watch from any angle
  - ➡ Zoom in/out without quality loss!
  - ➡ Virtual world visualization support
    (motion trails, coordinate system, boundary planes, vector arrows)
- **Automation** (segmentation, categories)

ANVIL for Gesture Annotation

- supports motion capture
  - synchronization of video, sound and mocap



Anvil 5.0 beta 11

```
*** Anvil 5.0 beta 11 DEBUG MODE ***
WARNING! XML validation failed: cvc-complex-ty
Open ANVIL file: /Users/kipp/Work/EMBOTS_SVN
Loading video:
    video codec: Cinepak
    screen size: 480x360
    frame rate: 25.0fps
    duration: 03:40:36 (5508 frames)
    audio: LINEAR 22050.0Hz mono
```

Current specification: ıta/NeffMocapDB/anvil3/mocap-v3-spec.xml

01:21:72 frame 2043

Mocap viewer: Angry.bvh

frame: 10839

Main Video: AntonyAngry.avi (100%)

Track: gesture.phrase

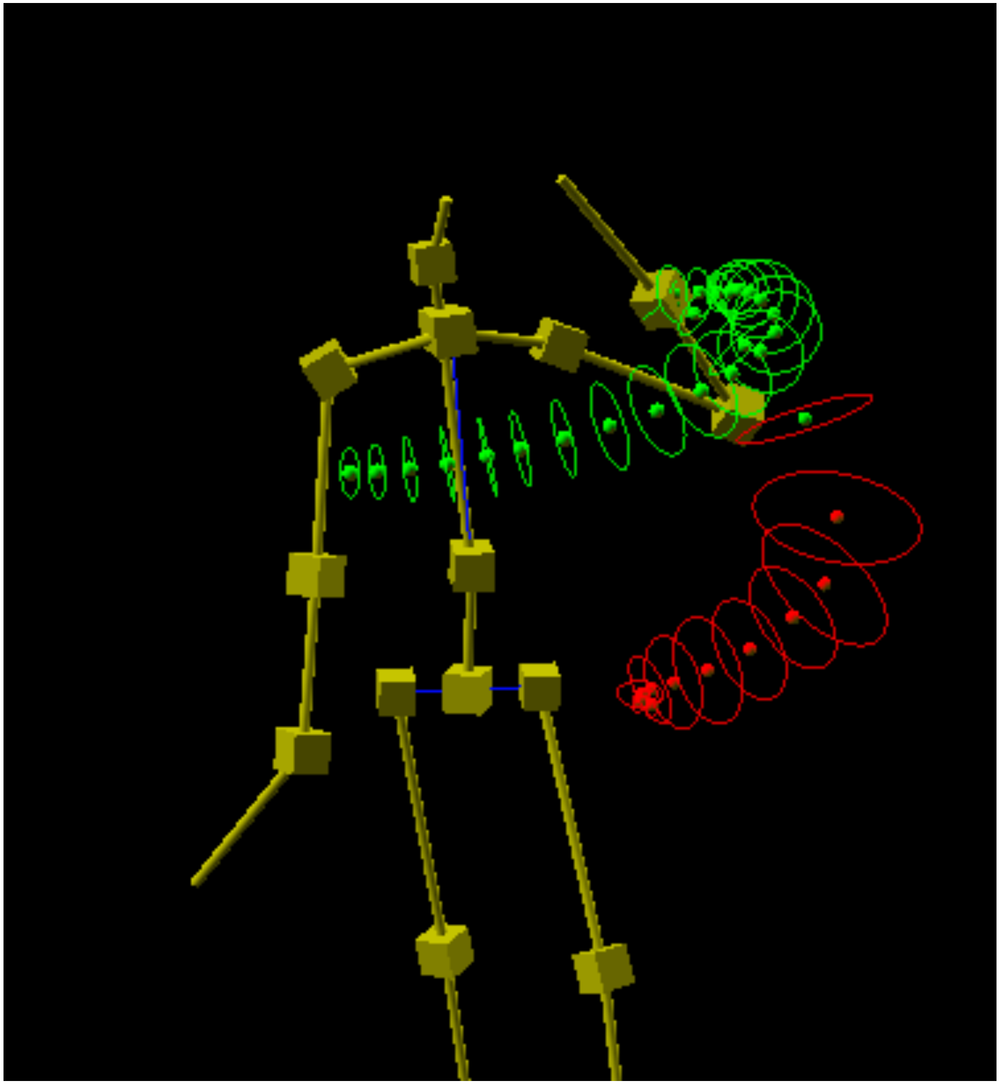
Track (span): gesture.phrase
Referenced track: gesture.phase
Time: 01:20:32 – 01:23:68 (84 frames)

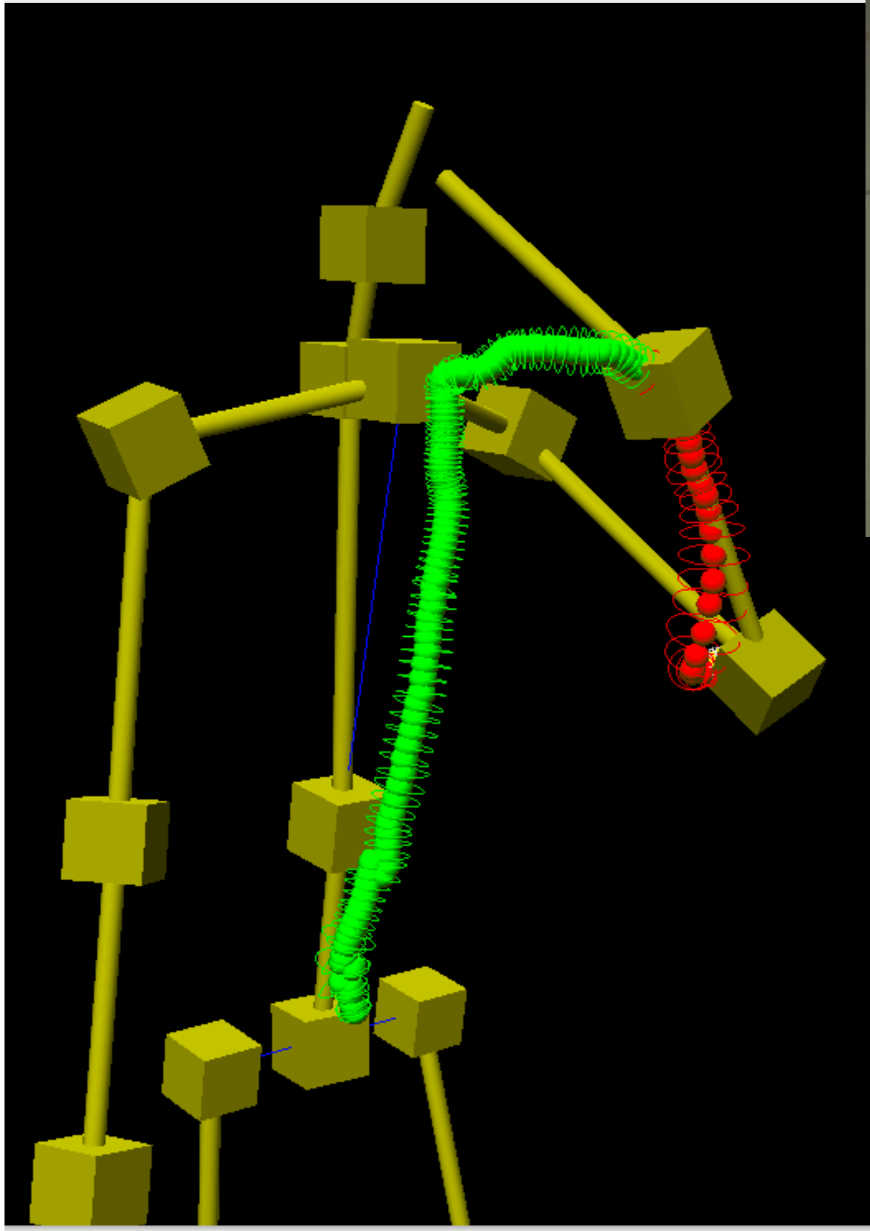
Attributes
lexeme: Frame
handedness: RH

Annotation: AntonyAngry-3.anvil

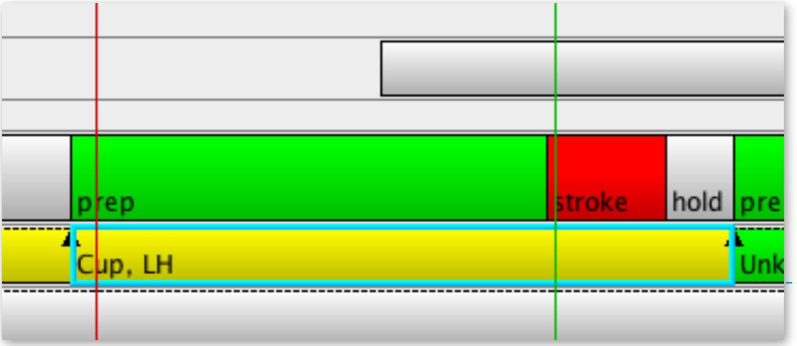
Motion Trails

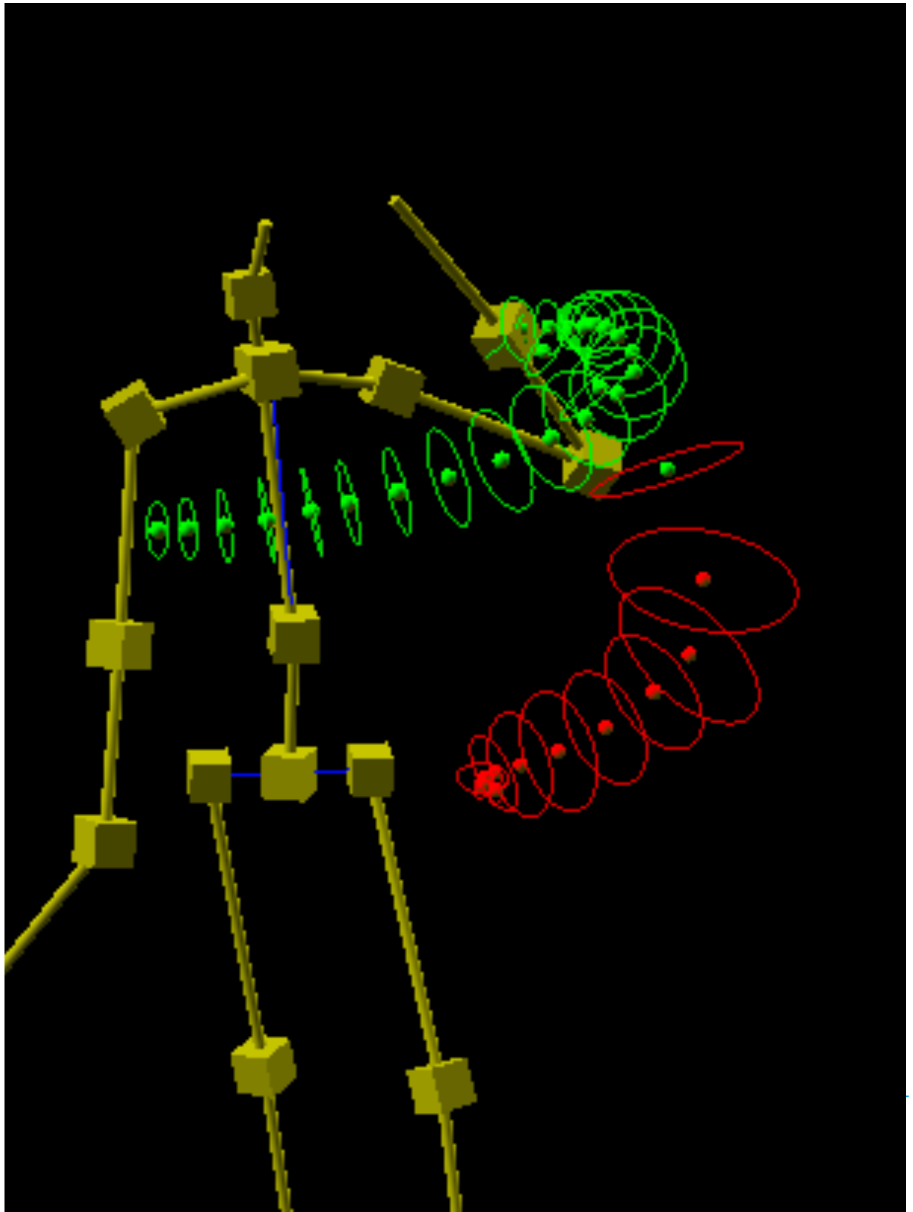
- Continuous representation of motion in 4D
- Shows segmentation by color coding
- Gives an impression of the velocity profile
  - spacing: indicator but too subtle
  - circles: indicate direction vector, can be scaled (gain), do not occlude
- Prevents annotation errors
  - incorrect hand or omission

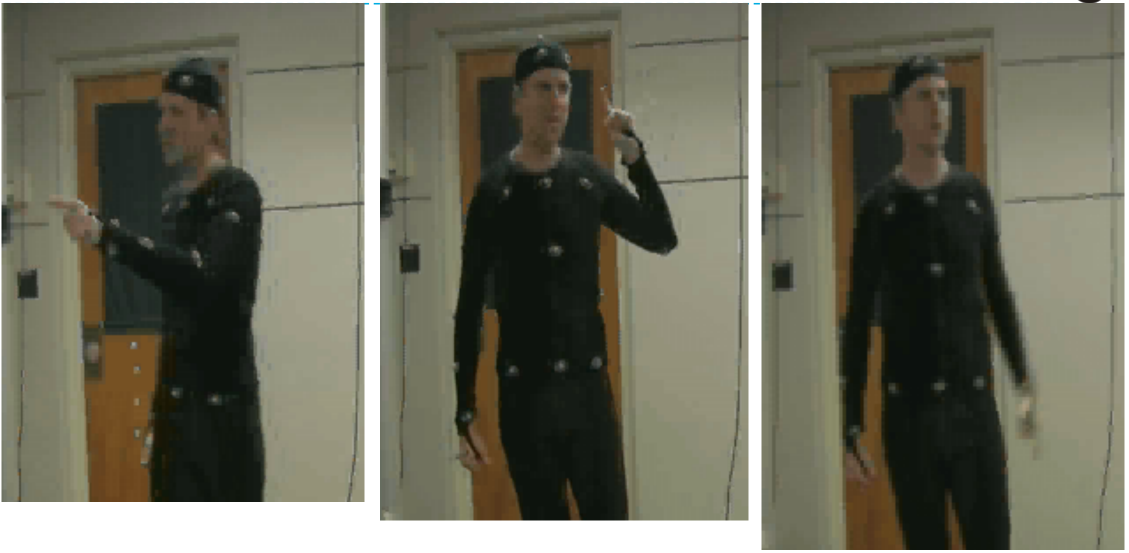




- Slow gesture → small circles, spheres close to each other
- Trail shows sideways motion of hand

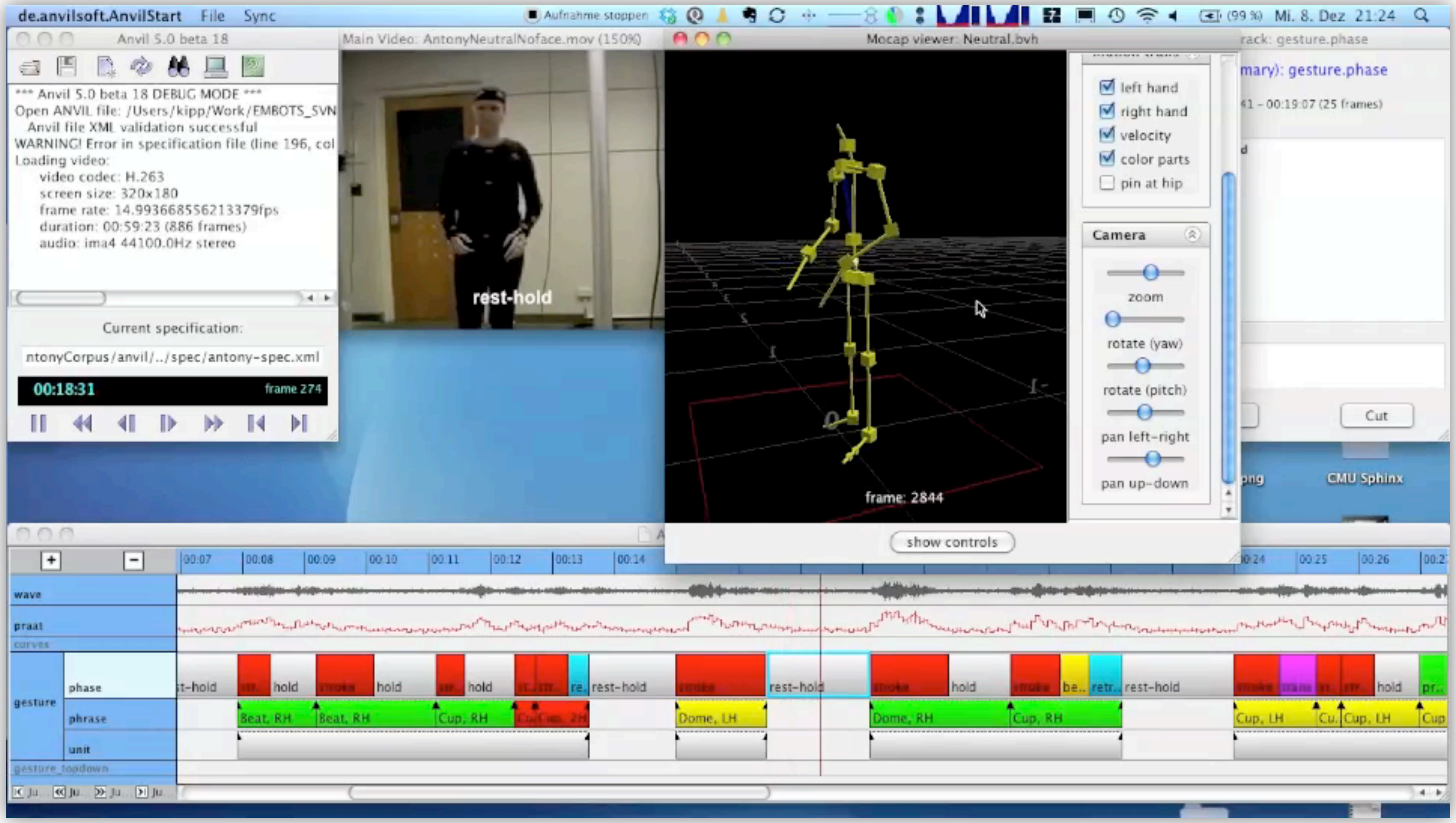






- High velocity → bigger circles,
  spheres more spaced
- Shows nicely that **stroke** has even higher velocity at ist peak

Video + Mocap + 4D Trails



```
*** Anvil 5.0 beta 18 DEBUG MODE ***
Open ANVIL file: /Users/kipp/Work/EMBOTS_SVN
  Anvil file XML validation successful
WARNING! Error in specification file (line 196, col
Loading video:
    video codec: H.263
    screen size: 320x180
    frame rate: 14.99366855621337 9fps
    duration: 00:59:23 (886 frames)
    audio: ima4 44100.0Hz stereo
```

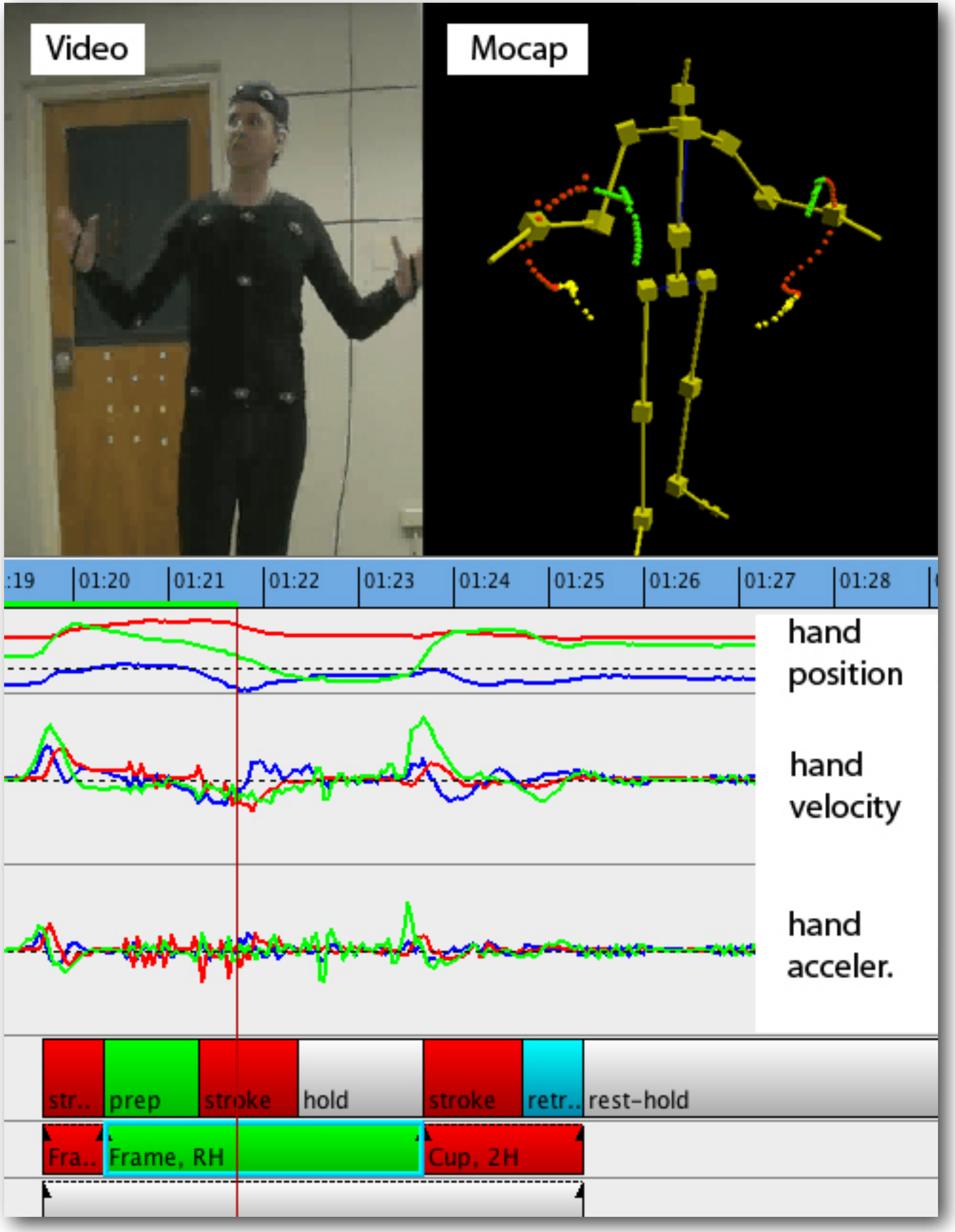
Current specification: ntonyCorpus/anvil/../spec/antony-spec.xml

00:18:31 frame 274

☑ left hand
☑ right hand
☑ velocity
☑ color parts
☐ pin at hip

41 – 00:19:07 (25 frames)

[Kipp 2011, Heloir, Neff, Kipp, 2010]



Video
Mocap
:19
01:20
01:21
01:22
01:23
01:24
01:25
01:26
01:27
01:28
hand position
hand velocity
hand acceler.
str..
prep
stroke
hold
stroke
retr..
rest-hold
Fra..
Frame, RH
Cup, 2H

Automated Handedness Detection

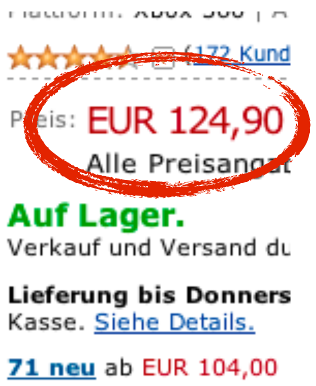
- Handedness automatically detected gesture-wise (= single gesture)

- Compare the normalized path length of right hand L_{RH} and left hand L_{LH} over a gesture of length d (time)
  → if $\frac{|L_{RH} - L_{LH}|}{d} < 0.12 \frac{m}{s}$ then gesture labeled 2H

- First test on 269 gestures: 83% correct

So you don‘t have 100,000$ for your personal mocap lab?

Poor Man‘s Mocap



(172 Kund

Preis: EUR 124,90

Alle Preisang

Auf Lager.

Verkauf und Versand du

Lieferung bis Donners

Kasse. Siehe Details.

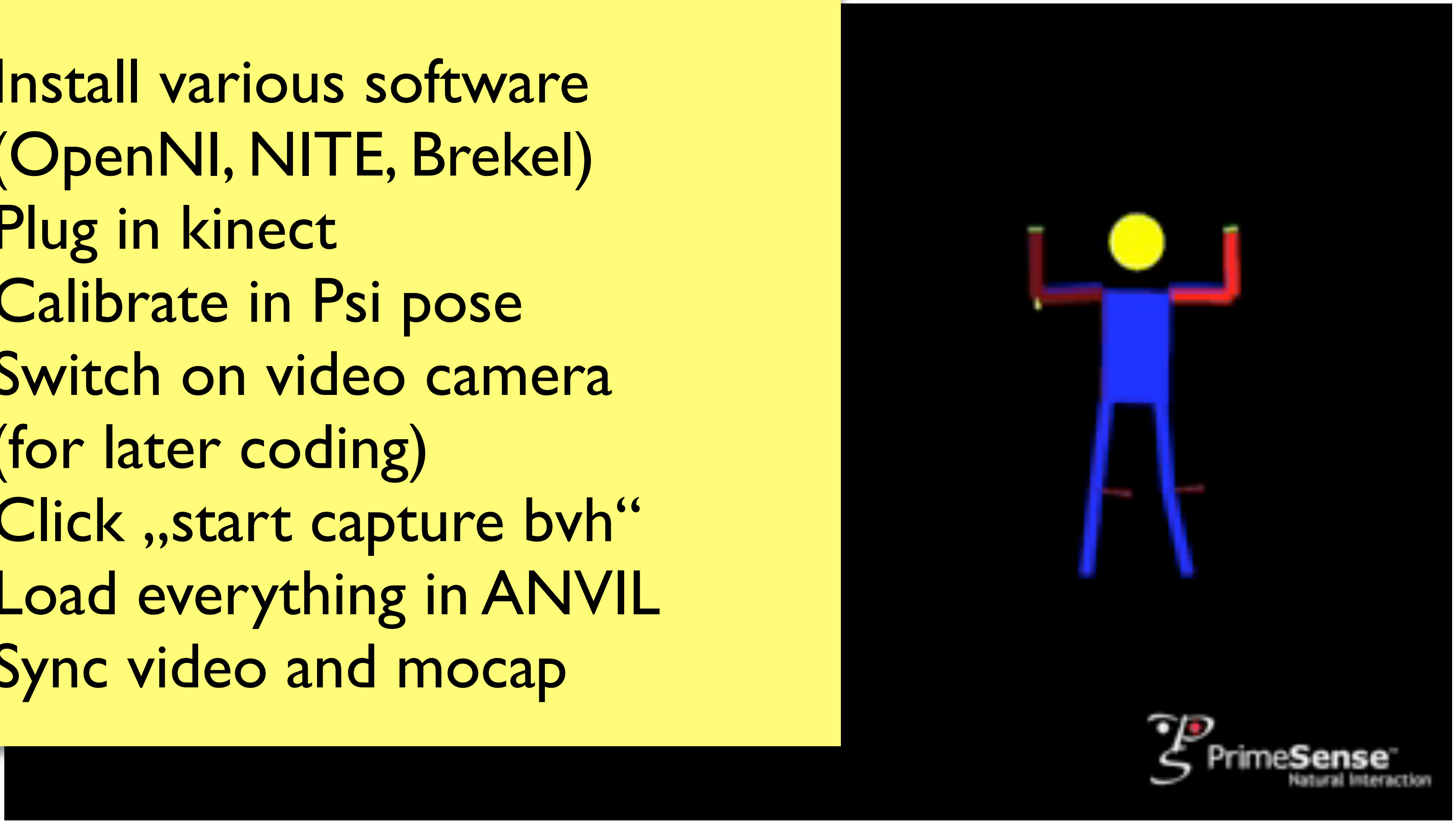
71 neu ab EUR 104,00

- Microsoft Kinect
- Hacked one hour after release
- Free software for skeleton tracking
- No excuses :)

Recipe „Kinect for Anvil“

1. Install various software (OpenNI, NITE, Brekel)
2. Plug in kinect
3. Calibrate in Psi pose
4. Switch on video camera (for later coding)
5. Click „start capture bvh“
6. Load everything in ANVIL
7. Sync video and mocap

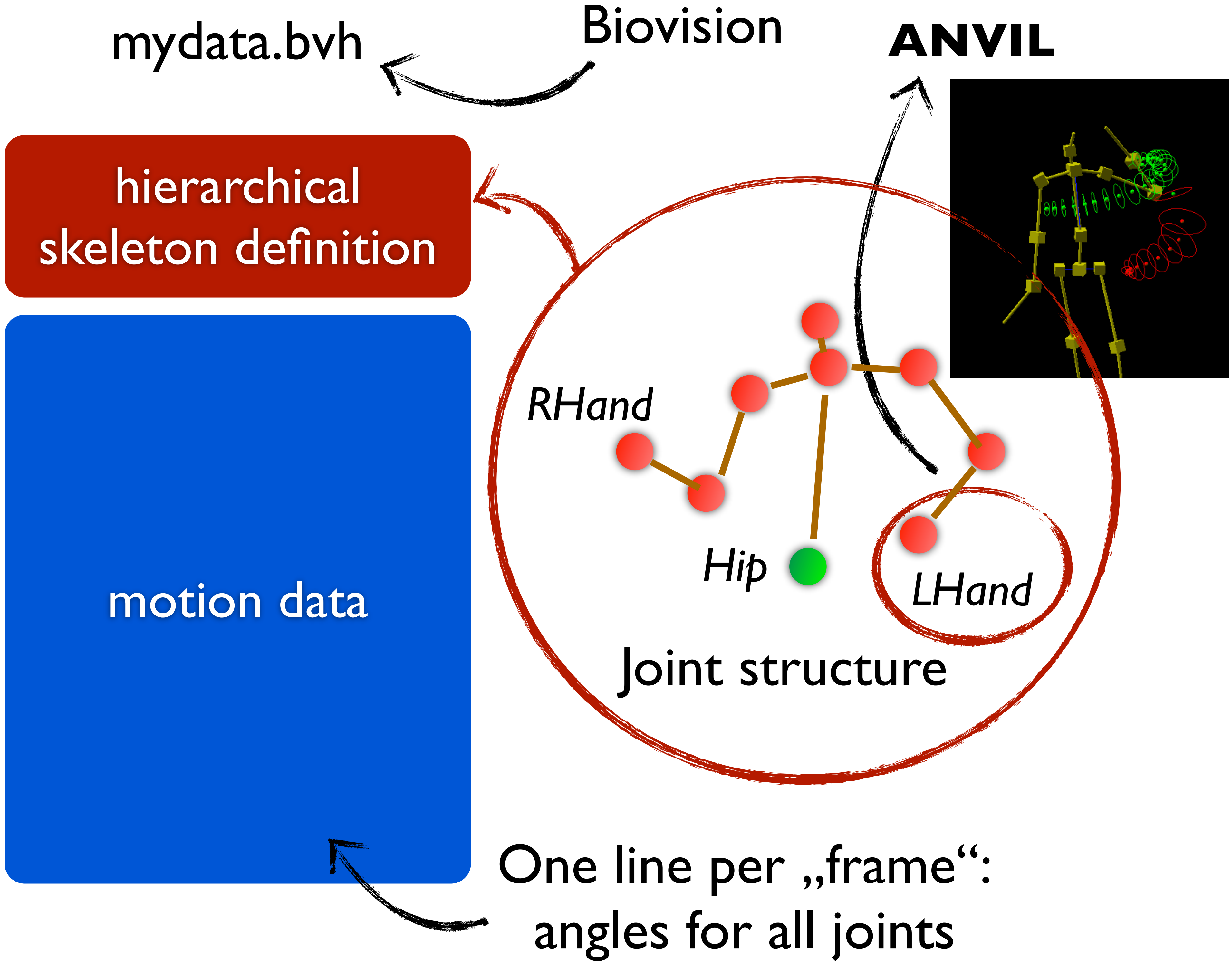
124.90 €



=> can produce a .bvh file (demo)
we will put a „howto“ online soon

=> view & annotate in ANVIL

DEMO



mydata.bvh

Biovision

ANVIL

hierarchical skeleton definition

motion data

Joint structure

One line per „frame“: angles for all joints

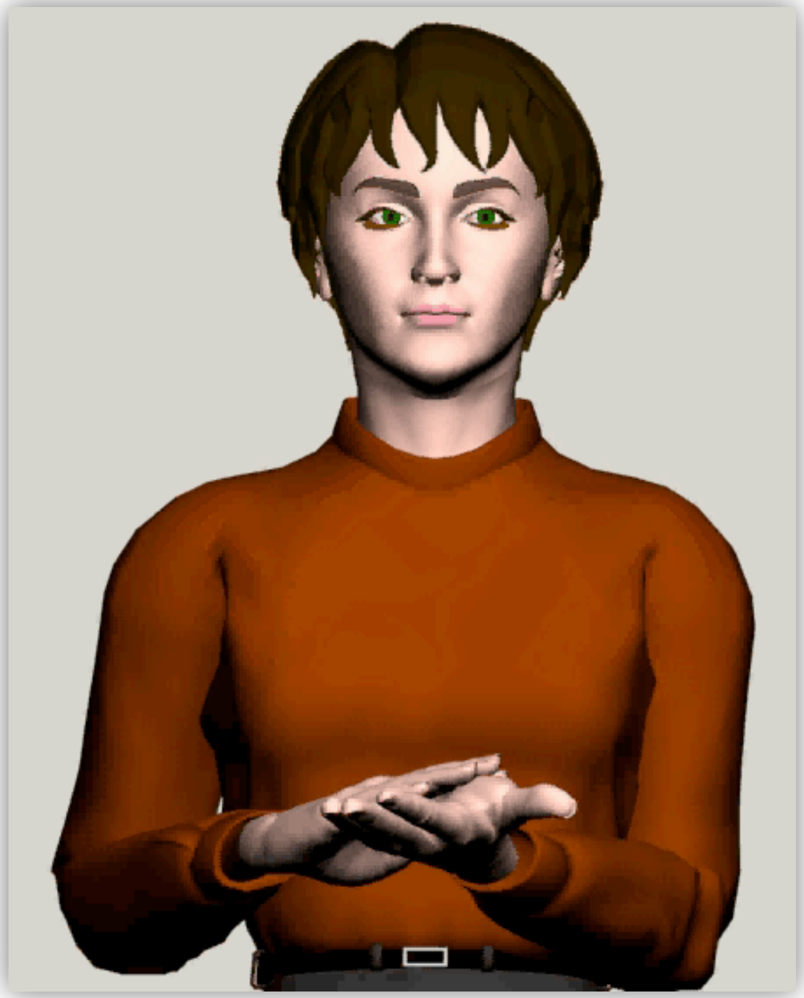
Signing Avatars

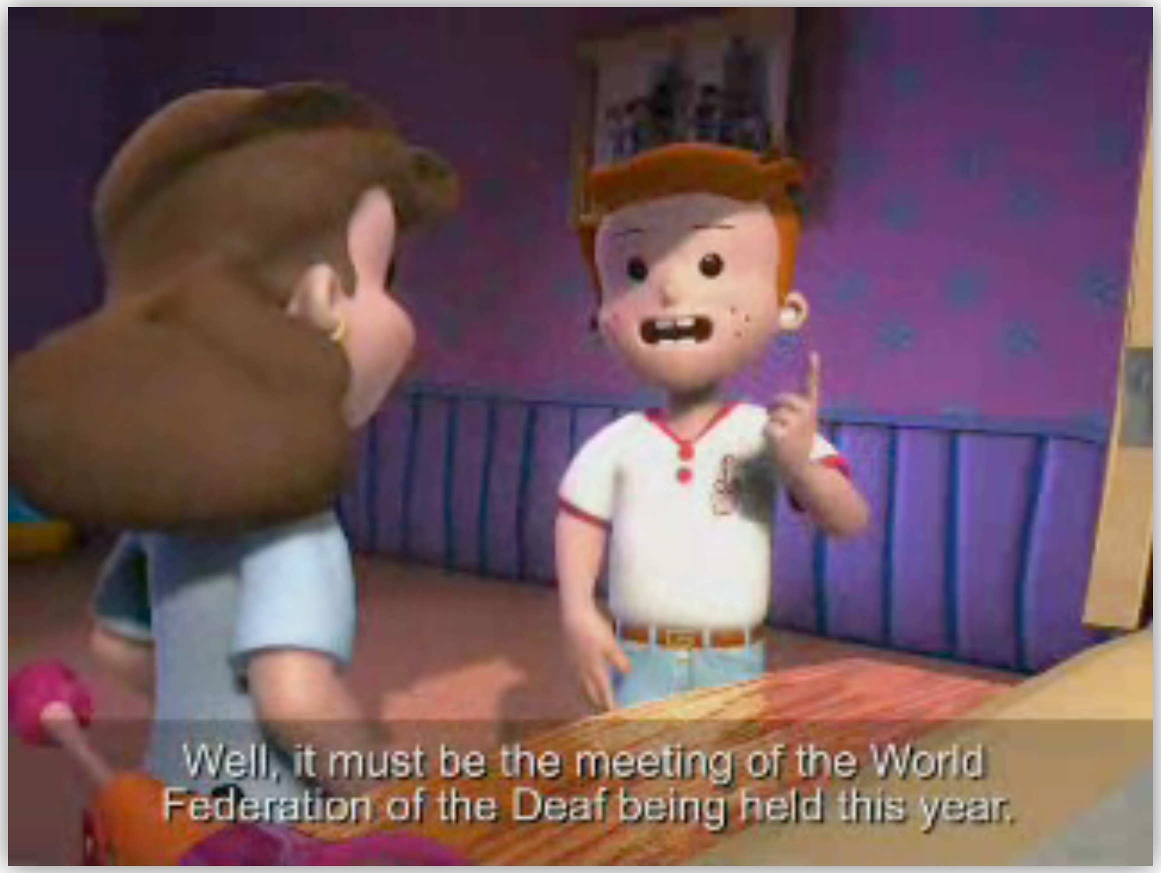
Sign Language Avatars

- 500,000 Deaf in Europe
  - ➡ Mother language / primary language: sign language
  - ➡ Spoken language = second language (hard to learn!)
  - ➡ 80% of the deaf leave school with significant reading/writing problems
- German Federal Ministry for Labour and Social Affairs (BMAS):
  - ➡ Are signing avatars a solution for accessible dynamic web content? (current comprehensibility around 60%)
- Feasibility study
  - ➡ state of the art, research priorities, applications

Focus Group Interviews

Existing Avatars





Well, it must be the meeting of the World Federation of the Deaf being held this year.

Avatar videos: criticism

- Upper body
  - ➡ too little involvement, especially no sideways rotations
  - ➡ important in constructed dialogue
- Face
  - ➡ too little eyebrow movement
  - ➡ hardly any mouthing (important for DGS)
    - ‣ absence of lip movement more striking than bad lip movement
    - ‣ recent CG movie („Lissy“) allowed quite good lip reading!
  - ➡ missing teeth and tongue
(necessary e.g. for letters L and N)

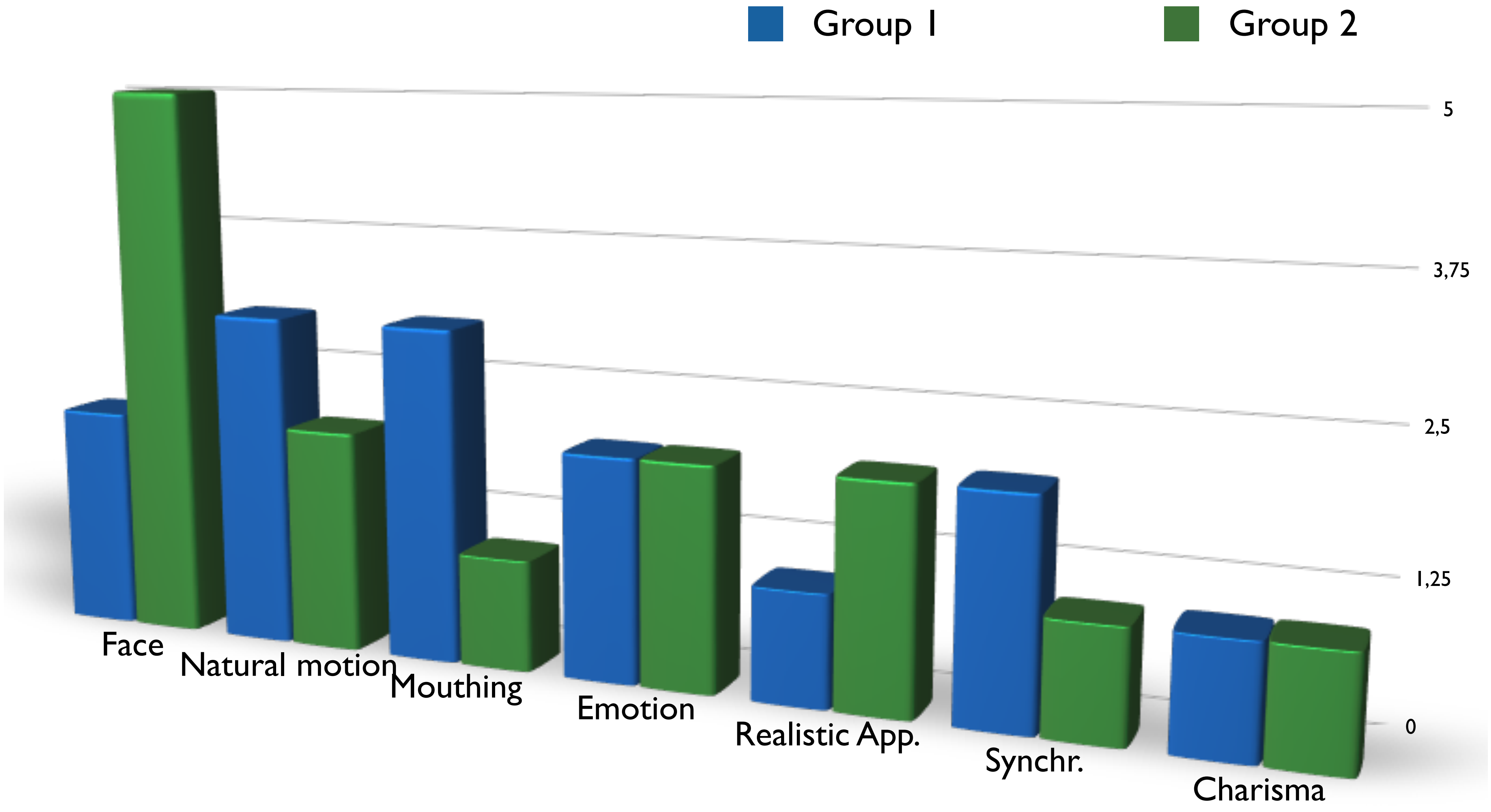
Avatar videos: criticism

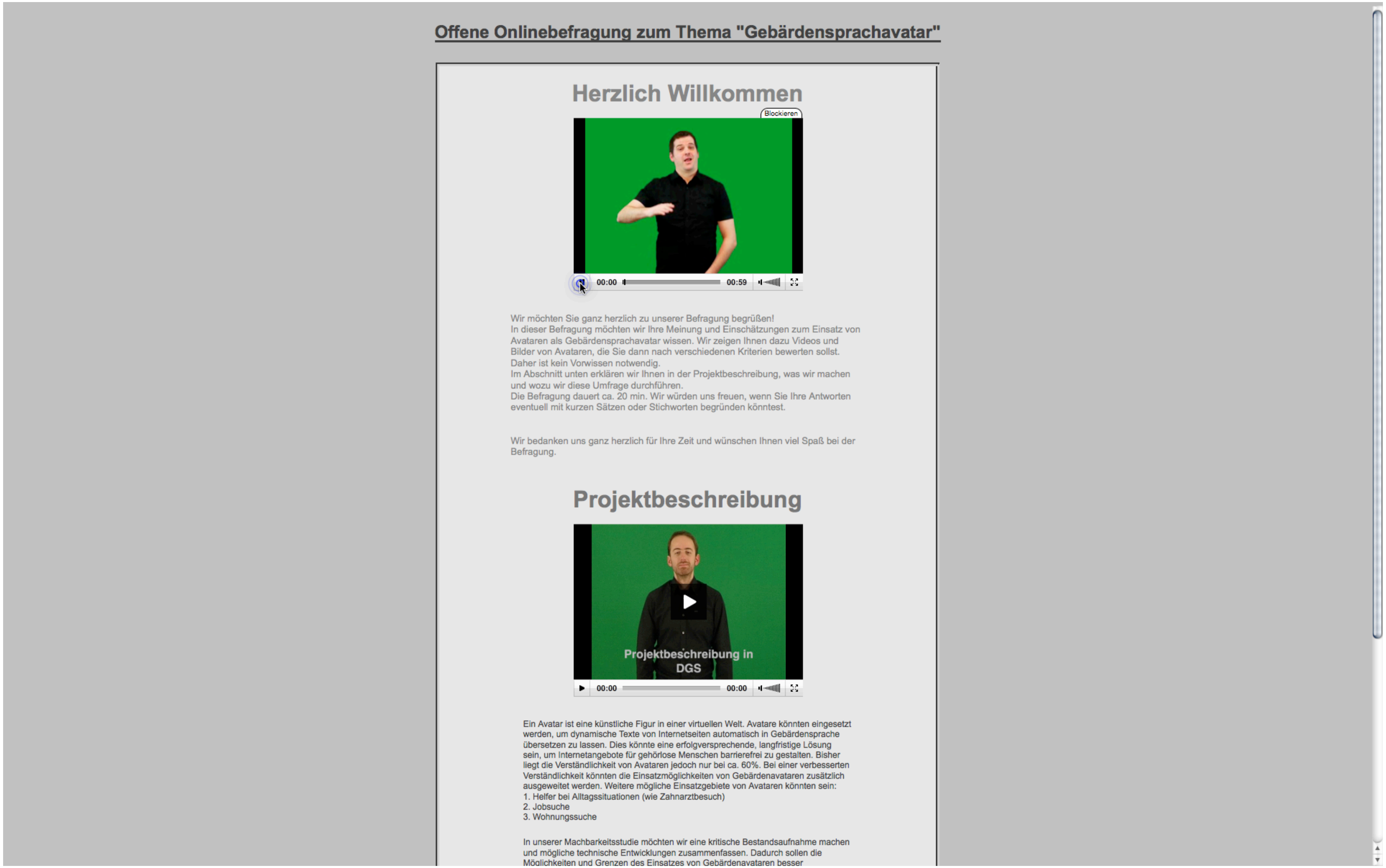
- Style
    - ➡ hardly any emotional expression
    - ➡ stiff / robotic movements
    - ➡ missing personality easily interpreted as cold / unfriendly
- Synchronization
    - ➡ mouthing and signs durations did not match
    - ➡ important for keeping the face as a focus point (!), otherwise focus occillates between hands and face

Avatar videos: criticism

- Technical
  - ➡ good lighting and contrast important (e.g. black clothes are good)
  - ➡ shadows support 3D effect
  - ➡ preferably: speed & perspective under user control
- Avatar Appearance
  - ➡ different avatars for different domains
  - ➡ child avatar & cartoonish: for kids and entertainment
  - ➡ adult & realistic: serious applications (politics, church ...)

Avatar Aspects





Offene Onlinebefragung zum Thema "Gebärdensprachavatar"

Herzlich Willkommen

Wir möchten Sie ganz herzlich zu unserer Befragung begrüßen!
In dieser Befragung möchten wir Ihre Meinung und Einschätzungen zum Einsatz von Avataren als Gebärdensprachavatar wissen. Wir zeigen Ihnen dazu Videos und Bilder von Avataren, die Sie dann nach verschiedenen Kriterien bewerten sollst. Daher ist kein Vorwissen notwendig.
Im Abschnitt unten erklären wir Ihnen in der Projektbeschreibung, was wir machen und wozu wir diese Umfrage durchführen.
Die Befragung dauert ca. 20 min. Wir würden uns freuen, wenn Sie Ihre Antworten eventuell mit kurzen Sätzen oder Stichworten begründen könntest.

Wir bedanken uns ganz herzlich für Ihre Zeit und wünschen Ihnen viel Spaß bei der Befragung.

Projektbeschreibung

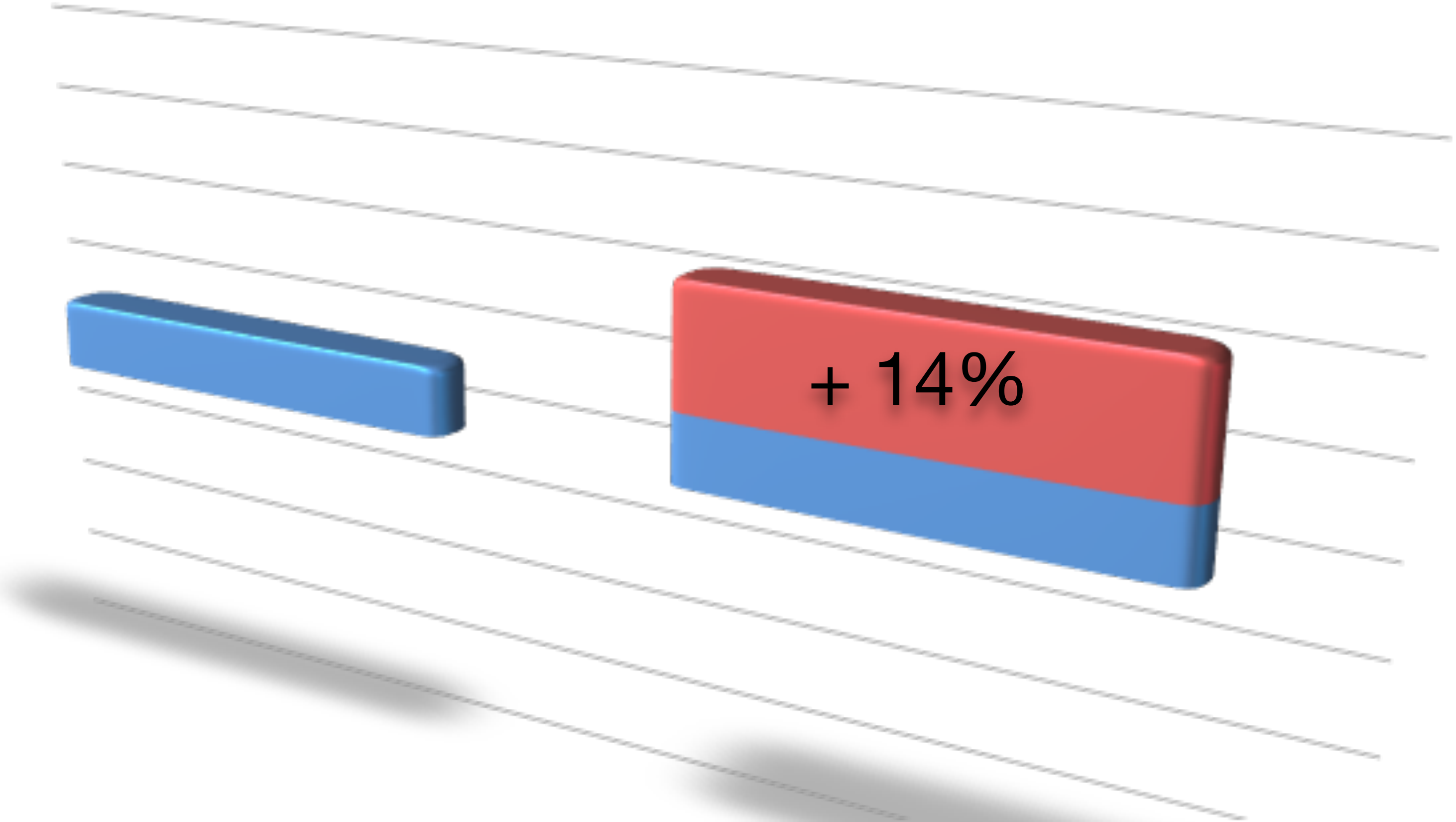
Ein Avatar ist eine künstliche Figur in einer virtuellen Welt. Avatare könnten eingesetzt werden, um dynamische Texte von Internetseiten automatisch in Gebärdensprache übersetzen zu lassen. Dies könnte eine erfolgversprechende, langfristige Lösung sein, um Internetangebote für gehörlose Menschen barrierefrei zu gestalten. Bisher liegt die Verständlichkeit von Avataren jedoch nur bei ca. 60%. Bei einer verbesserten Verständlichkeit könnten die Einsatzmöglichkeiten von Gebärdenavataren zusätzlich ausgeweitet werden. Weitere mögliche Einsatzgebiete von Avataren könnten sein:
1. Helfer bei Alltagssituationen (wie Zahnarztbesuch)
2. Jobsuche
3. Wohnungssuche

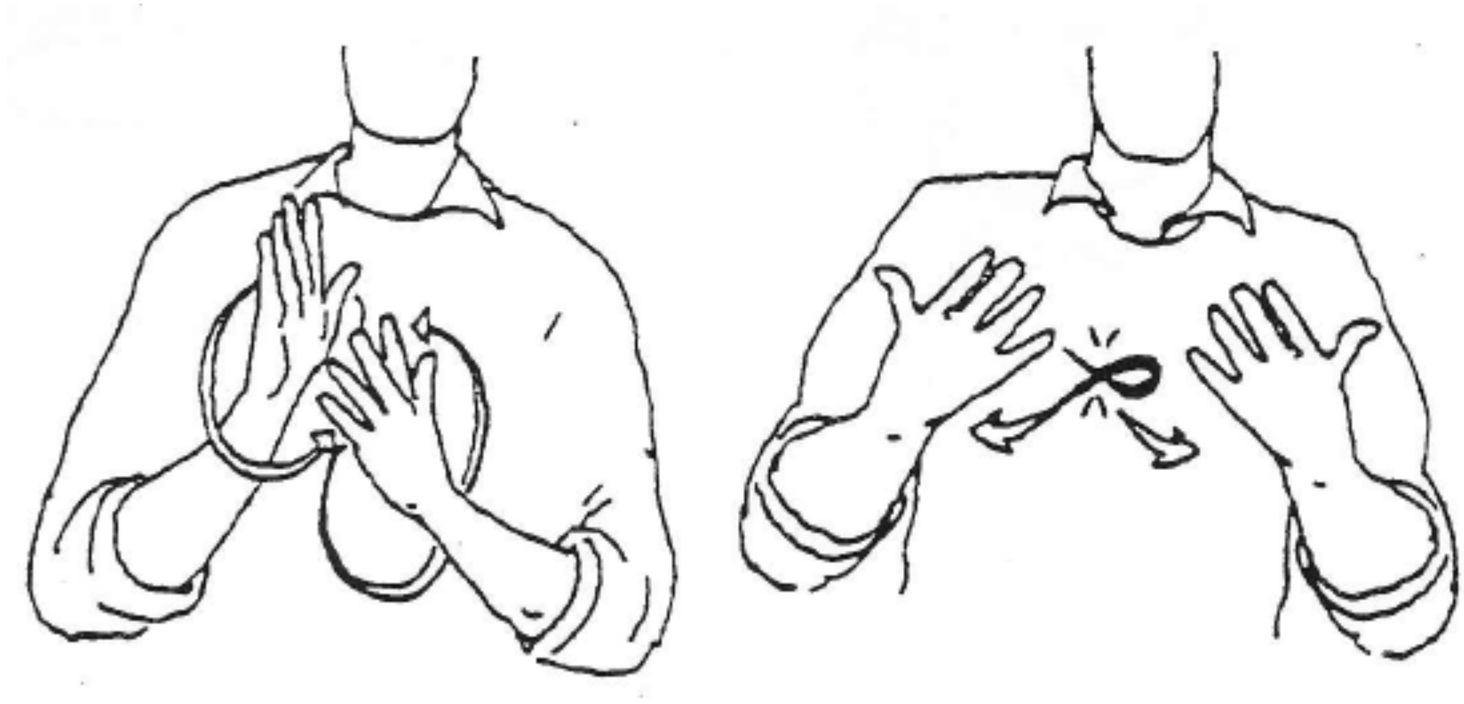
In unserer Machbarkeitsstudie möchten wir eine kritische Bestandsaufnahme machen und mögliche technische Entwicklungen zusammenfassen. Dadurch sollen die Möglichkeiten und Grenzen des Einsatzes von Gebärdenavataren besser

Do you consider avatars useful?

(-2 ... +2)

- Before: +0.4
- After: +0.7





Avatar
Sign Language
Animation

Character Animation

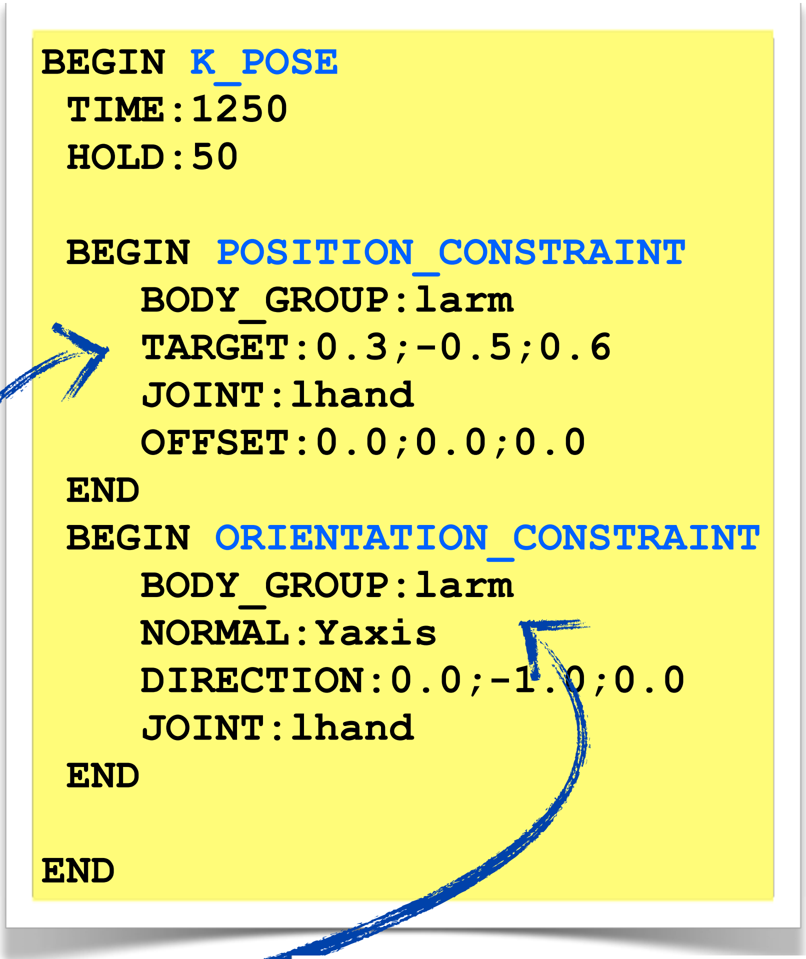
- Initial Motivation
  - ➡ Create a reusable character animation engine
  - ➡ Exploration of coverbal gesture
- Existing systems (e.g. Greta, SmartBody, MAX) proposed high-level control languages
  - ➡ Behavior Markup Language (BML)
  - ➡ e.g. <gesture type=“pointing“ stroke=“x“ />
- Useful layer of abstraction but...
  - ➡ What if you need more control (hand shape, torso involvement, size of gesture...) ?

EMBR:
EMbodied Agent Behavior Realizer

- Our solution: Low-level control language
  - ➡ Wrapper around keyframe animation
  - ➡ Theory-independent (bottom-up approach)
- Used to create gesture lexemes
  - ➡ Add „semantics“ inside the gesture, e.g. this is the stroke, this is a preparatory motion
  - ➡ Also: specify open parameters like hand shape, location, direction (in progress) => exploit knowledge of „stroke“

EMBRScript

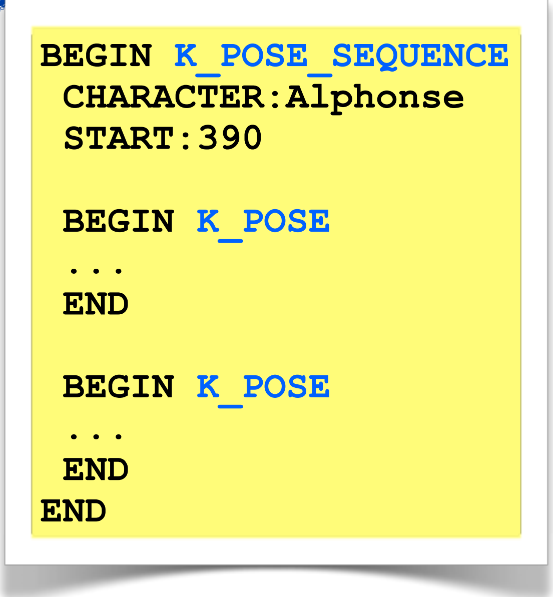
- Pose: body configuration for a single time point (+ hold duration), defined by constraints like
  - ➡ hand at a particular point in space
  - ➡ hand shape, shoulder position
  - ➡ facial expression, level of blushing
- Every constraint applies to part of the body
  - ➡ Channels are inherent (arms, hands, face, shoulders, ...)
- Pose sequence: sequence of poses + start time
  - ➡ a deliberate temporal segmentation
  - ➡ design decision: we use sequences for glosses



```
BEGIN K_POSE
 TIME:1250
 HOLD:50

 BEGIN POSITION_CONSTRAINT
    BODY_GROUP:larm
    TARGET:0.3;-0.5;0.6
    JOINT:lhand
    OFFSET:0.0;0.0;0.0
 END
 BEGIN ORIENTATION_CONSTRAINT
    BODY_GROUP:larm
    NORMAL:Yaxis
    DIRECTION:0.0;-1.0;0.0
    JOINT:lhand
 END

END
```

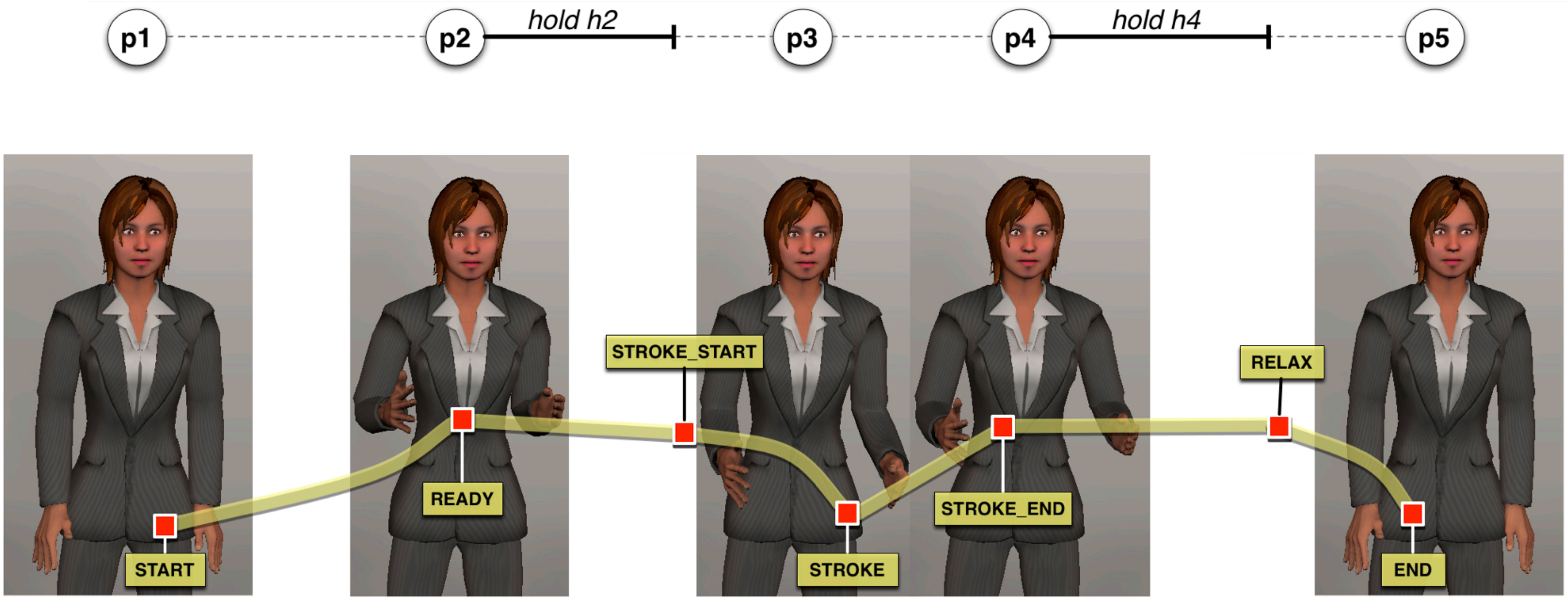


```
BEGIN K_POSE_SEQUENCE
 CHARACTER:Alphonse
 START:390

 BEGIN K_POSE
 ...
 END

 BEGIN K_POSE
 ...
 END
END
```

EMBRScript



Temporal „movement phase“ markers allow synchronization & modification (e.g. drop preparation)

[Heloir, Kipp, 2009, 2010]

BehaviorBuilder tool to create and test EMBRScripts

Sign Language Animation

- Attempt 1:
  - ➡ source video (human)
  - ➡ rotoscope (avatar)

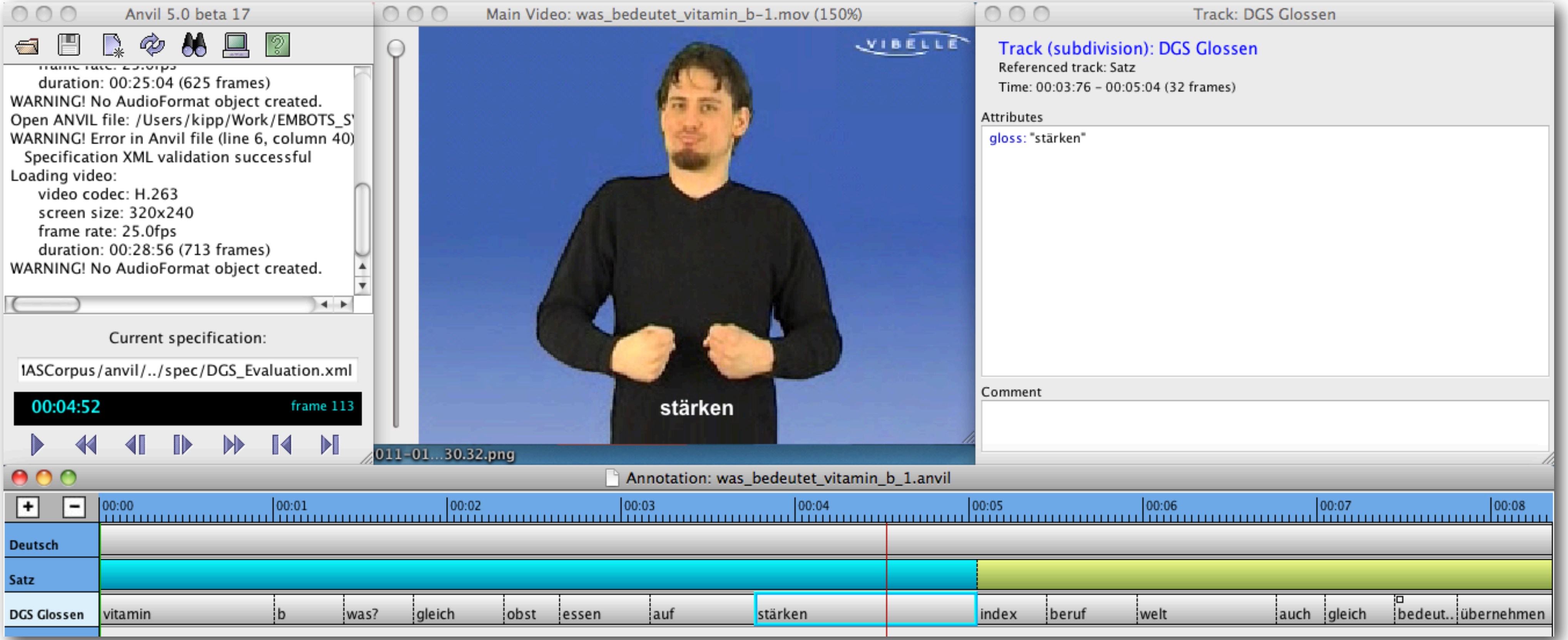
failure

Why?

- Single sign disambiguation
  - ➡ same manual movement, different meaning
  - ➡ mouthing
  - ➡ gaze, facial expression, pose narrow down possible meanings
- identify sentence topic
  - ➡ interrogative facial expression / eyebrow raise
  - ➡ pauses
  - ➡ posture shift
- Face as fixation point
  - ➡ allows parallel observance of face, mouth, hands, torso
  - ➡ hard to do if face is static

Sign Language Animation

- Attempt 1: failure
  - ➡ source video (human)
  - ➡ rotoscope (avatar)
- Attempt 2:
  - ➡ source video (human)
  - ➡ remake (human)
  - ➡ rotoscope (avatar)



```
    duration: 00:25:04 (625 frames)
WARNING! No AudioFormat object created.
Open ANVIL file: /Users/kipp/Work/EMBOTS_S
WARNING! Error in Anvil file (line 6, column 40)
  Specification XML validation successful
Loading video:
    video codec: H.263
    screen size: 320x240
    frame rate: 25.0fps
    duration: 00:28:56 (713 frames)
WARNING! No AudioFormat object created.
```

Current specification:

IASCorpus/anvil/../spec/DGS_Evaluation.xml

00:04:52 frame 113

Track (subdivision): DGS Glossen

Referenced track: Satz

Time: 00:03:76 – 00:05:04 (32 frames)

Attributes

gloss: "stärken"

Comment

Deutsch

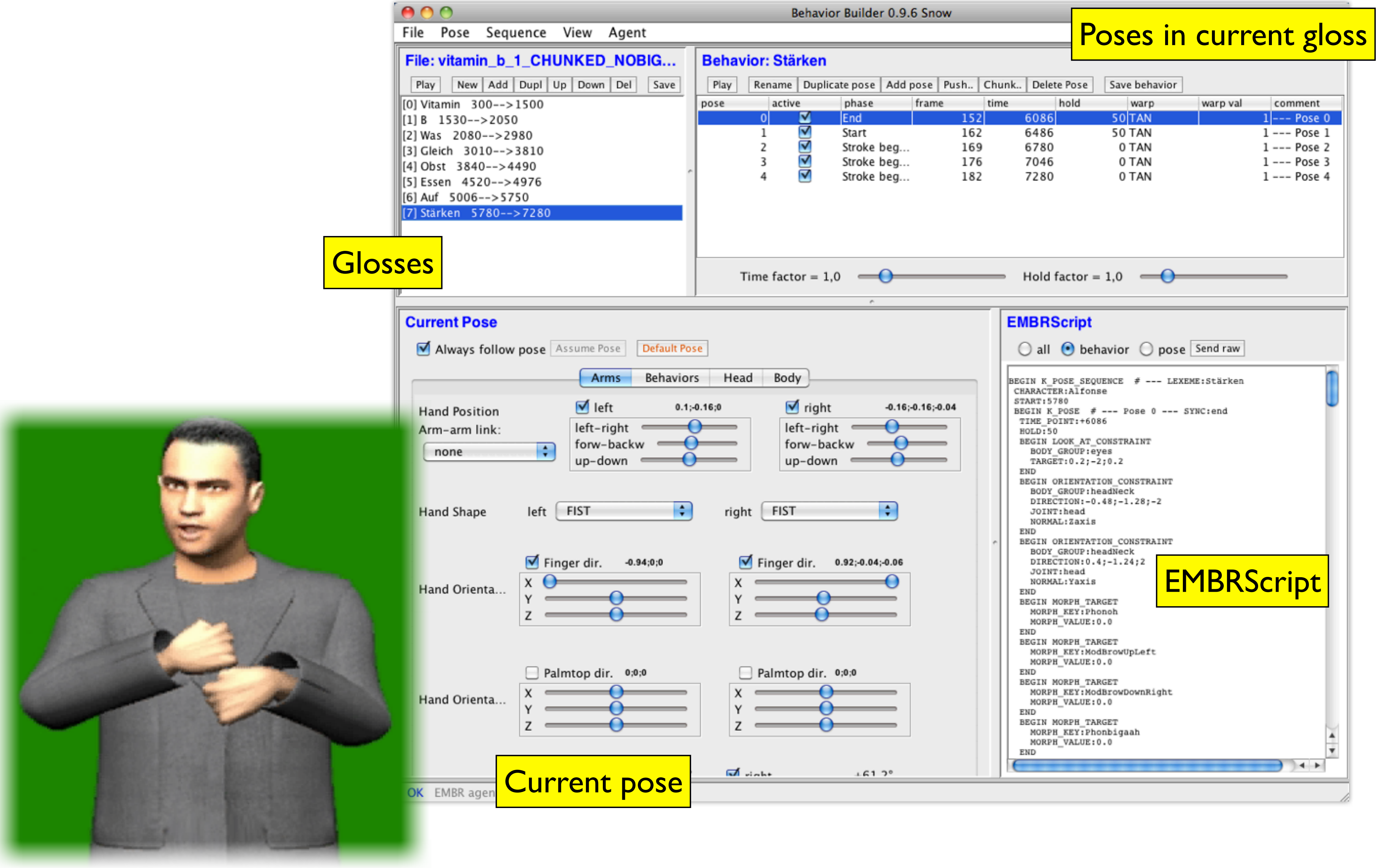
Satz

DGS Glossen: vitamin | b | was? | gleich | obst | essen | auf | stärken | index | beruf | welt | auch | gleich | bedeut.. | übernehmen

Gloss-wise transcription for utterance segmentation.

re-make

BehaviorBuilder Extensions:
Gloss-based Creation of Animation Sequences



File: vitamin_b_1_CHUNKED_NOBIG...

[0] Vitamin 300-->1500
[1] B 1530-->2050
[2] Was 2080-->2980
[3] Gleich 3010-->3810
[4] Obst 3840-->4490
[5] Essen 4520-->4976
[6] Auf 5006-->5750
[7] Stärken 5780-->7280

Behavior: Stärken

pose	active	phase	frame	time	hold	warp	warp val	comment
0	☑	End	152	6086	50	TAN	1	--- Pose 0
1	☑	Start	162	6486	50	TAN	1	--- Pose 1
2	☑	Stroke beg...	169	6780	0	TAN	1	--- Pose 2
3	☑	Stroke beg...	176	7046	0	TAN	1	--- Pose 3
4	☑	Stroke beg...	182	7280	0	TAN	1	--- Pose 4

Time factor = 1,0

Hold factor = 1,0

Current Pose

☑ Always follow pose

Hand Position

Arm-arm link: none

☑ left 0.1;-0.16;0

left-right

forw-backw

up-down

☑ right -0.16;-0.16;-0.04

left-right

forw-backw

up-down

Hand Shape left FIST right FIST

Hand Orienta...

☑ Finger dir. -0.94;0;0 X Y Z

☑ Finger dir. 0.92;-0.04;-0.06 X Y Z

Hand Orienta...

☐ Palmtop dir. 0;0;0 X Y Z

☐ Palmtop dir. 0;0;0 X Y Z

☑ right +61.2°

EMBRScript

○ all ● behavior ○ pose

```
BEGIN K_POSE_SEQUENCE  # --- LEXEME:Stärken
 CHARACTER:Alfonse
 START:5780
 BEGIN K_POSE  # --- Pose 0 --- SYNC:end
  TIME_POINT:+6086
  HOLD:50
  BEGIN LOOK_AT_CONSTRAINT
    BODY_GROUP:eyes
    TARGET:0.2;-2;0.2
  END
  BEGIN ORIENTATION_CONSTRAINT
    BODY_GROUP:headNeck
    DIRECTION:-0.48;-1.28;-2
    JOINT:head
    NORMAL:Zaxis
  END
  BEGIN ORIENTATION_CONSTRAINT
    BODY_GROUP:headNeck
    DIRECTION:0.4;-1.24;2
    JOINT:head
    NORMAL:Yaxis
  END
  BEGIN MORPH_TARGET
    MORPH_KEY:Phonoh
    MORPH_VALUE:0.0
  END
  BEGIN MORPH_TARGET
    MORPH_KEY:ModBrowUpLeft
    MORPH_VALUE:0.0
  END
  BEGIN MORPH_TARGET
    MORPH_KEY:ModBrowDownRight
    MORPH_VALUE:0.0
  END
  BEGIN MORPH_TARGET
    MORPH_KEY:Phonbigaah
    MORPH_VALUE:0.0
  END
```

original

re-make

avatar

Lessons...

- SL is multimodal => change focus from manual gesture to ...
  - ➡ facial expression
  - ➡ mouthing
  - ➡ torso involvement
  - ➡ gaze
- Multimodality means
  - ➡ each modality as important as manual signs
  - ➡ explore synchronization
- Acceptability depends on
  - ➡ presence of style, personality, emotionality
  - ➡ prosody for information structure (topic) and segmentation
  - ➡ visual interest of the face => face as center of attention
- Good reliability test: Is it comprehensible?

Conclusions

- We need motion capture!
- Sign language research needs you!
- Start looking at numbers instead of pixels...

Thanks for listening!