User Authentication for Natural User Interfaces (NUIs)

Janusz Konrad

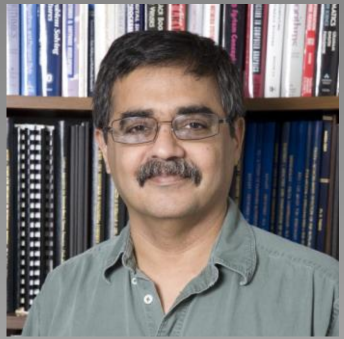
BOSTON UNIVERSITY

Boston University Electrical & Computer Engineering

IEEE Signal Processing Society

Acknowledgments

Prof. Prakash Ishwar



Prof. Nasir Memon (NYU)

Dr. Jonathan Wu (Amazon)

Dr. Jiawei Chen (Oppo)

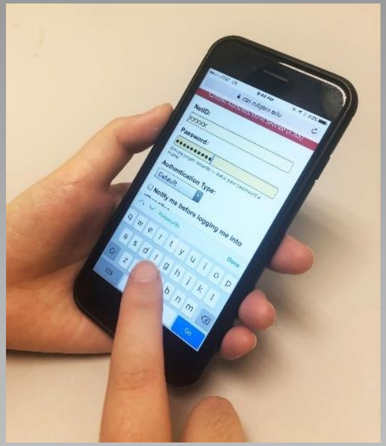
Dr. Napa Sae-Bae (RMUTSB, Thailand)

This work was supported in part by the National Science Foundation, and by Boston University. This lecture has been made possible by the IEEE Signal Processing Society.

What is user authentication ?

The process of verifying someone's identity, for example to access a restricted device, fetch restricted data, sign a document, vote, etc.

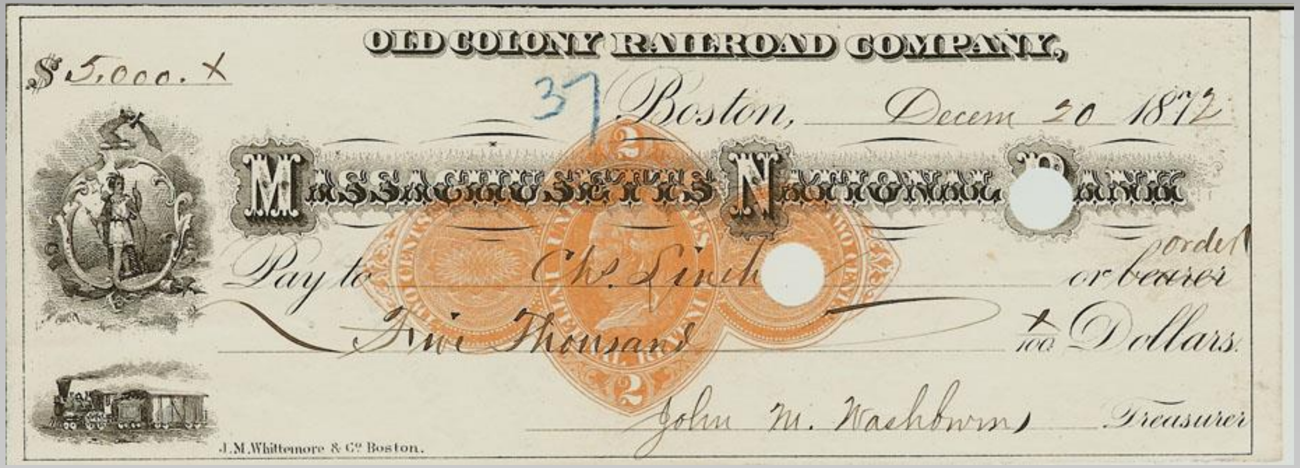
AUTHENTICATION



Authentication exploits something: **you have,**
you know,
you are

Authentication yesterday

Something you have



Something you know and you are

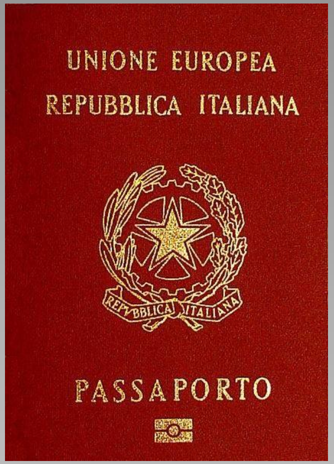
Authentication today (frequent)

Magnetic swipe card

Proximity card (e.g., RFID)

Something you have

Authentication today (occasional)



Claimed identity

Proof of identity

Something you have and you are

Authentication today (very frequent)

Claimed identity

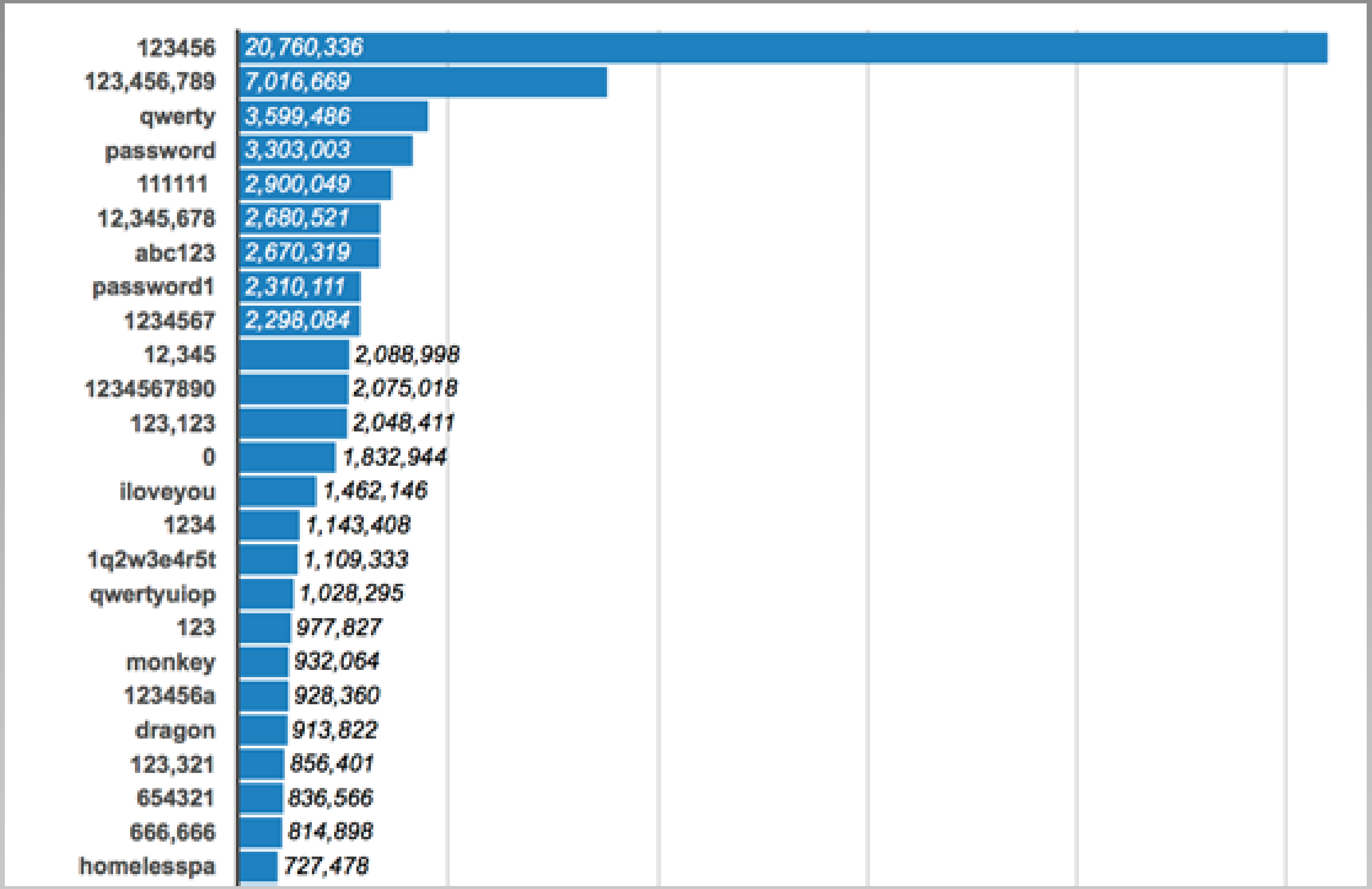
Proof of identity

Something you know

Passwords: should be unique and obey hygiene

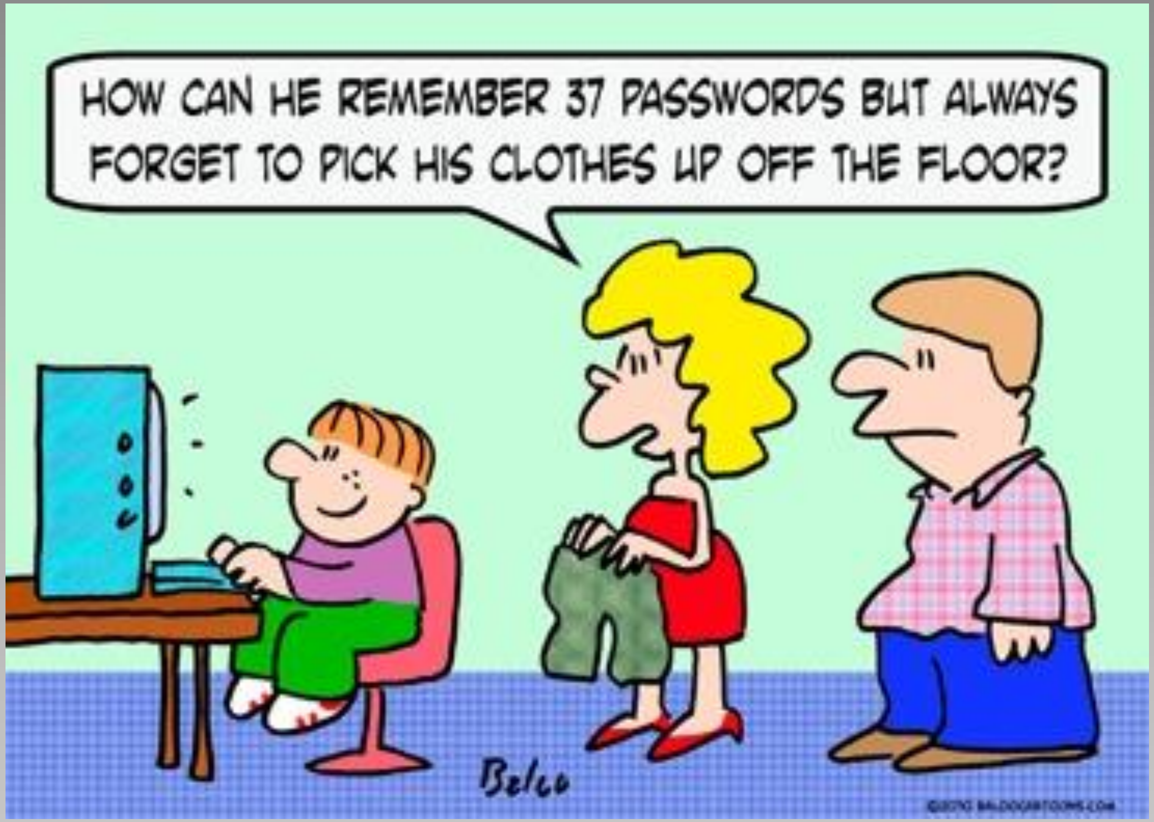
But passwords require mental effort, so ….

Top 100 most vulnerable passwords



NY Post, June 12, 2018
(https://nyp.st/2y8DoI3)

We need to remember more and more of them …



but as we age, it gets more and more difficult

Is there any hope ?

Authentication wish list:

- Simple
- Effortless (easy to remember)
- Secure

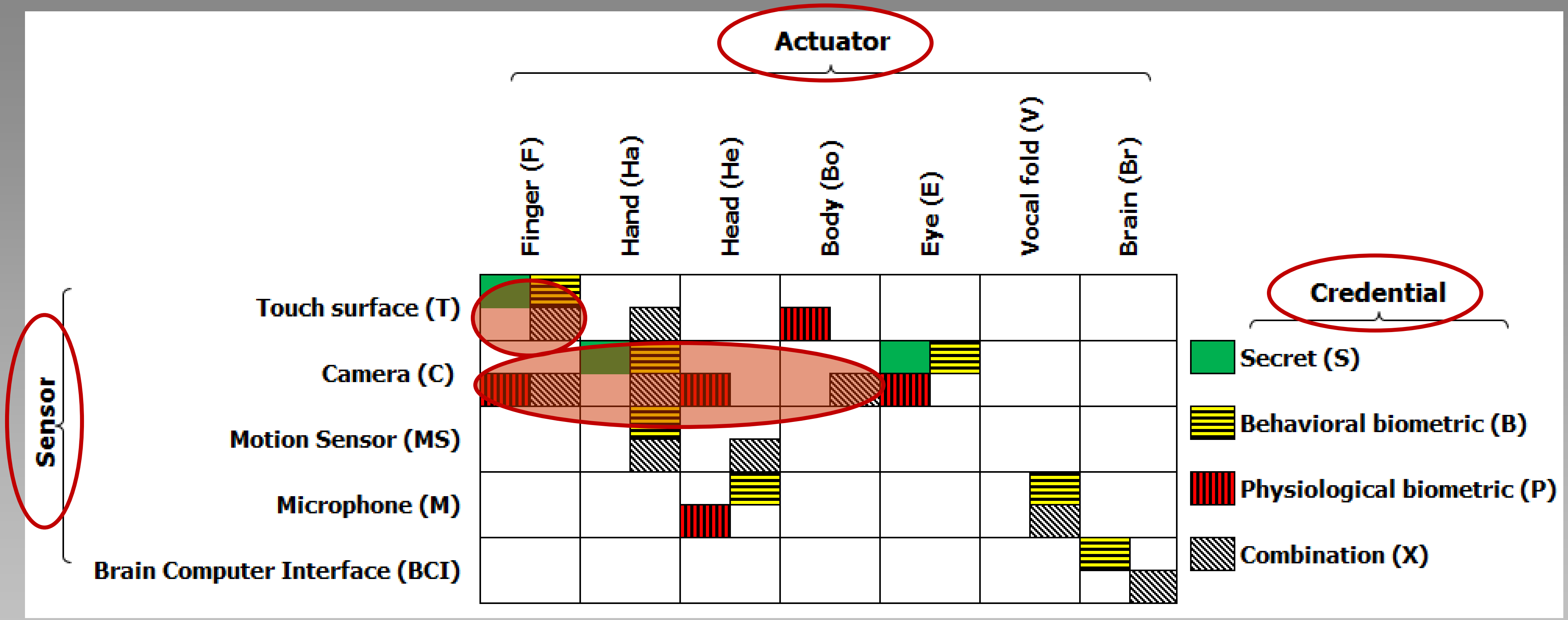
Natural User Interfaces (NUIs)

Touch surface | 3-D camera | Multiple cameras | AR headset / Smartwatch

- Emerging modes of user interaction with devices
- Natural user behavior
- Can NUIs be leveraged for user authentication ?

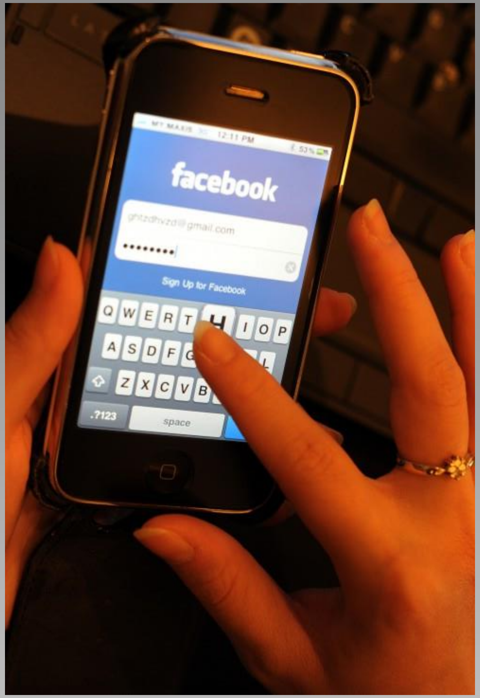
NSF (CISE-SATC) collaborative project between BU and NYU

NUI taxonomy

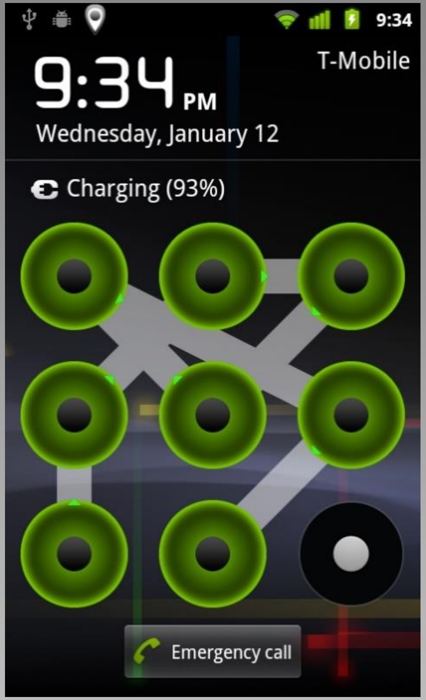


N. Sae-Bae, J. Wu, N. Memon, J. Konrad, and P. Ishwar
"Emerging NUI-based methods for user authentication: A new taxonomy and survey,"
IEEE Trans. Biometrics, Behavior, and Identity Science, vol. 1, pp. 5-31, Jan. 2019.

Touch surface: Early attempts



Password entry on touch keyboard:
- significant effort, slow
- subject to shoulder-surfing attack



Android pattern lock:
- easier to memorize
- also subject to smudge attack

Something you know

Further attempts

Microsoft Picture Password

Graphical passwords

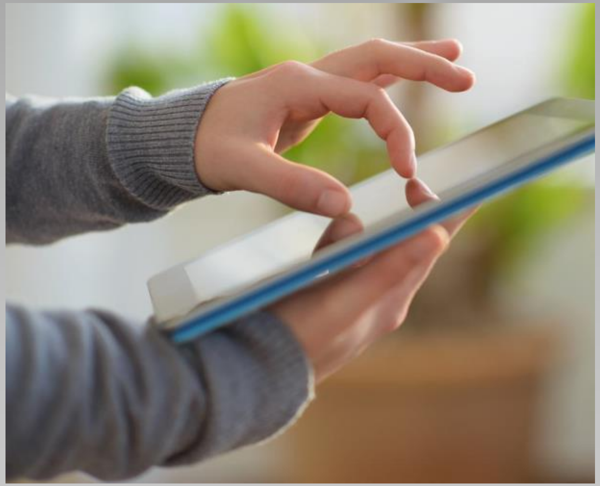
Something you know

How to exploit ``something you are'' (biometric features) on a touch surface ?

Common gestures performed on a touchscreen

Swiping

Pressing



Pinching

Multi-touch swiping

Multi-touch gestures

- Biometrically rich (much richer than single-touch gestures)
- More resilient against shoulder surfing than typing a password
- Natural action (easy to memorize)
- Can be renewed if compromised (unlike fingerprint, retinal scan, face image)

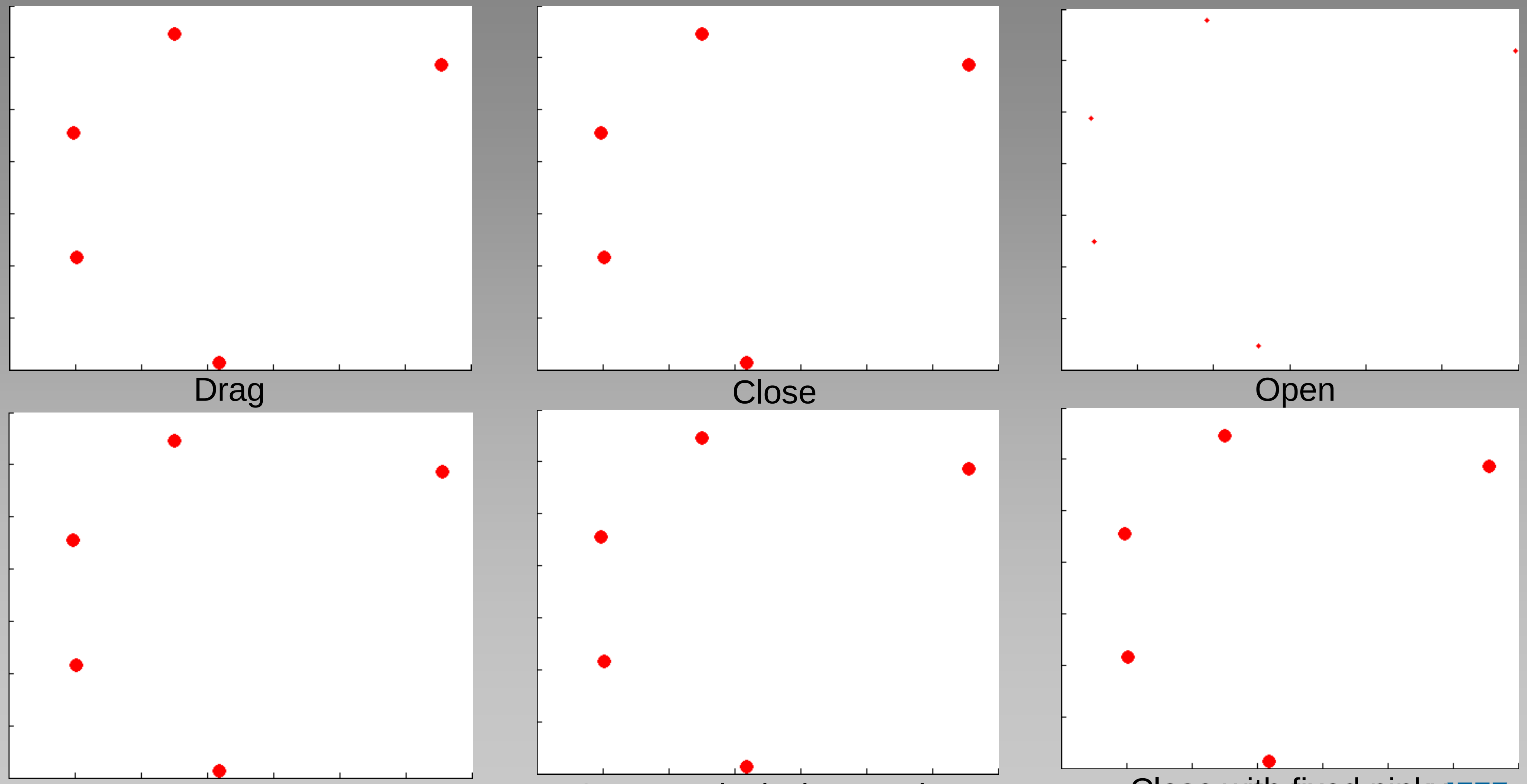
Multi-touch gesture test set (NYU study)

Annotation	Palm movement	Fingertip movement	Dynamic fingertips
'CCR'	Static	Circular(CCW)	All
'CR'	Static	Circular(CW)	All
'Closed'	Static	Close	All
'Drag'	Dynamic(↓)	Parallel	All
'DDC'	Dynamic(↘)	Close	All
'DUO'	Dynamic(↖)	Open	All
'FBD'	Static	Parallel(↓)	Fixed thumb and pinky
'FBSB'	Static	Parallel(⟨ shape)	Fixed thumb and pinky
'FBSA'	Static	Parallel(⟩ shape)	Fixed thumb and pinky
'FPCCR'	Static	Circular(CCW)	Fixed pinky
'FPC'	Static	Close	Fixed pinky
'FPO'	Static	open	Fixed pinky
'FPP'	Static	Parallel(↓)	Fixed pinky
'FTCCR'	Static	Circular(CCW)	Fixed thumb
'FTCR'	Static	Circular(CW)	Fixed thumb
'FTC'	Static	Close	Fixed thumb
'FTO'	Static	Open	Fixed thumb
'FTP'	Static	Parallel(↓)	Fixed thumb
'Flick'	Dynamic(↘)	Parallel	All(Quick)
'Opened'	Static	Open	All
'Scrawl'	Dynamic(Customized)	Parallel	All
'Swipe'	Dynamic(→)	Parallel	All

[Sae-Bae et al., TIFS, 2015]

Examples

Data courtesy of Memon

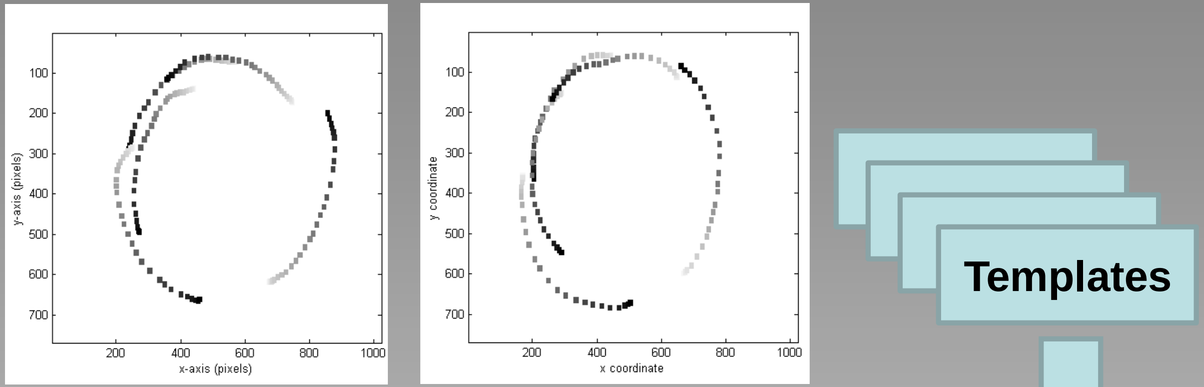


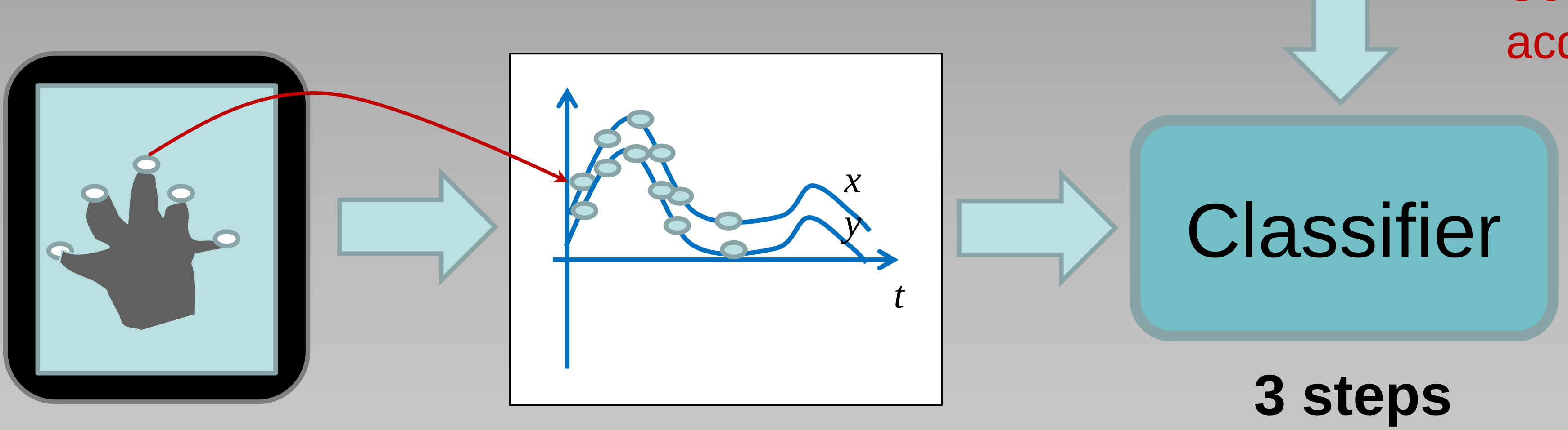
Drag, Close, Open

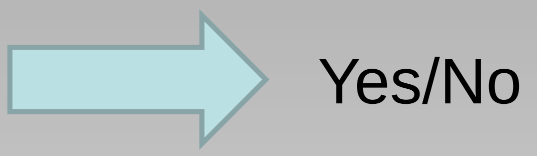
Clockwise rotation, Counter-clockwise rotation, Close with fixed pinky

IEEE Signal Processing Society

Multi-touch verification







Biometric data samples are never identical, although are similar

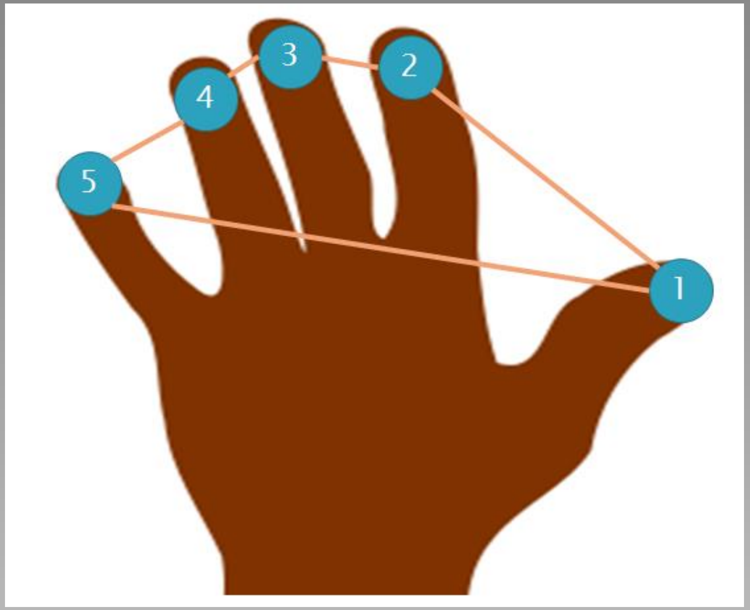
One sample is not good enough to represent a person

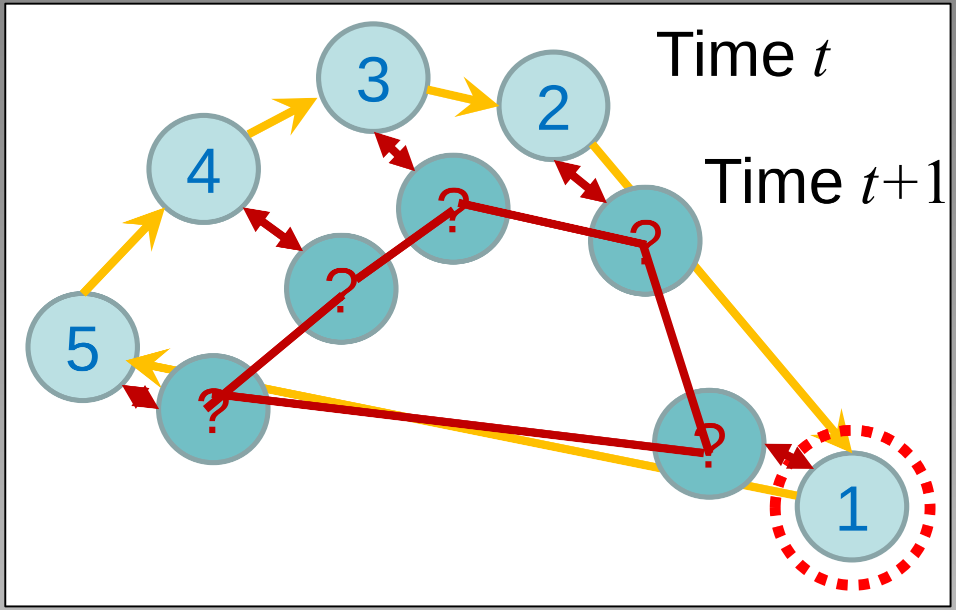
Several samples need to be acquired during enrollment

Diagram courtesy of Memon

Step 1: Data alignment

Graphics courtesy of Memon



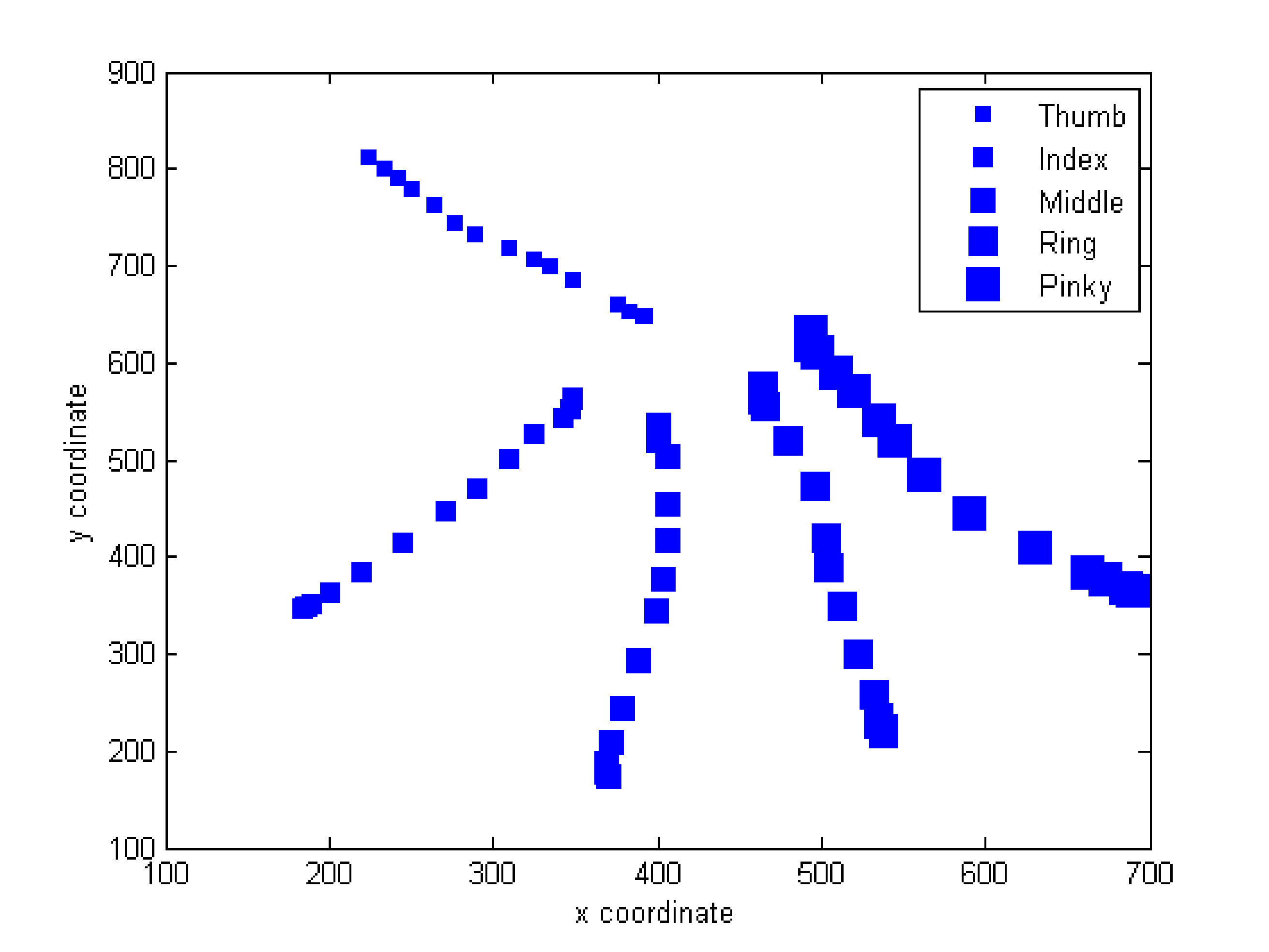


Possible Paths

- 5 – 4 – 3 – 2 – 1
- 5 – 3 – 4 – 2 – 1
- 5 – 3 – 2 – 4 – 1
- 5 – 3 – 2 – 1 – 4

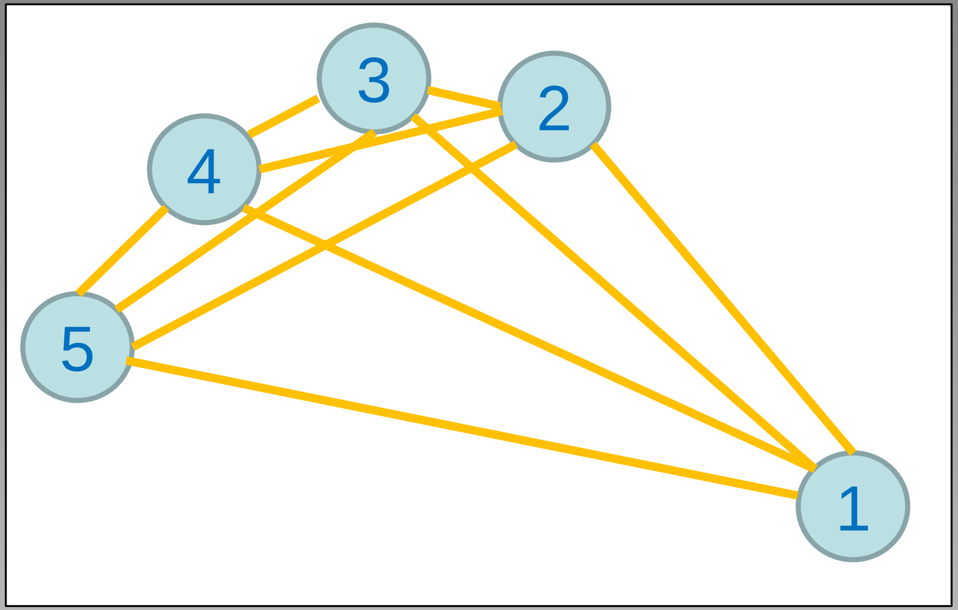
1. Locate the thumb
2. Track individual touch points by minimizing the sum of distances between same-ID touchpoints so that new IDs form a simple polygon

Aligned template



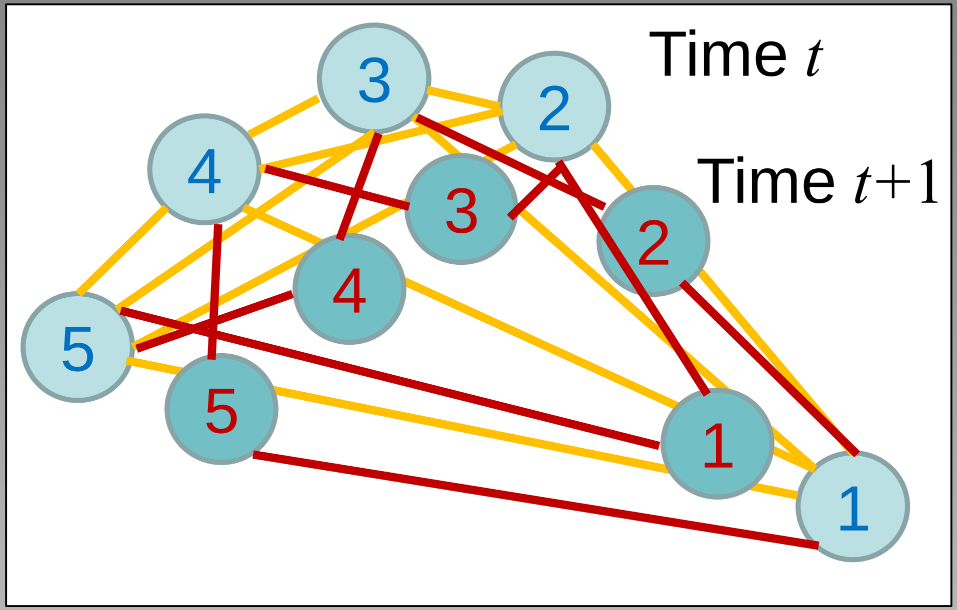
Data courtesy of Memon

Step 2: Feature vectors



10 Euclidean distances between 5 touch points at time t

Feature vector $\boldsymbol{p}_t \in R^{10}$



Additional 10 distances between each touch point k at time $t+1$ and touch points k-1 and k+1 at time t, to account for movement direction and speed

Feature vector $\boldsymbol{p}_t \in R^{20}$

Step 3: Assessing gesture similarity

- Given two feature sequences:

$$\boldsymbol{p} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_n], \qquad \boldsymbol{q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_m]$$

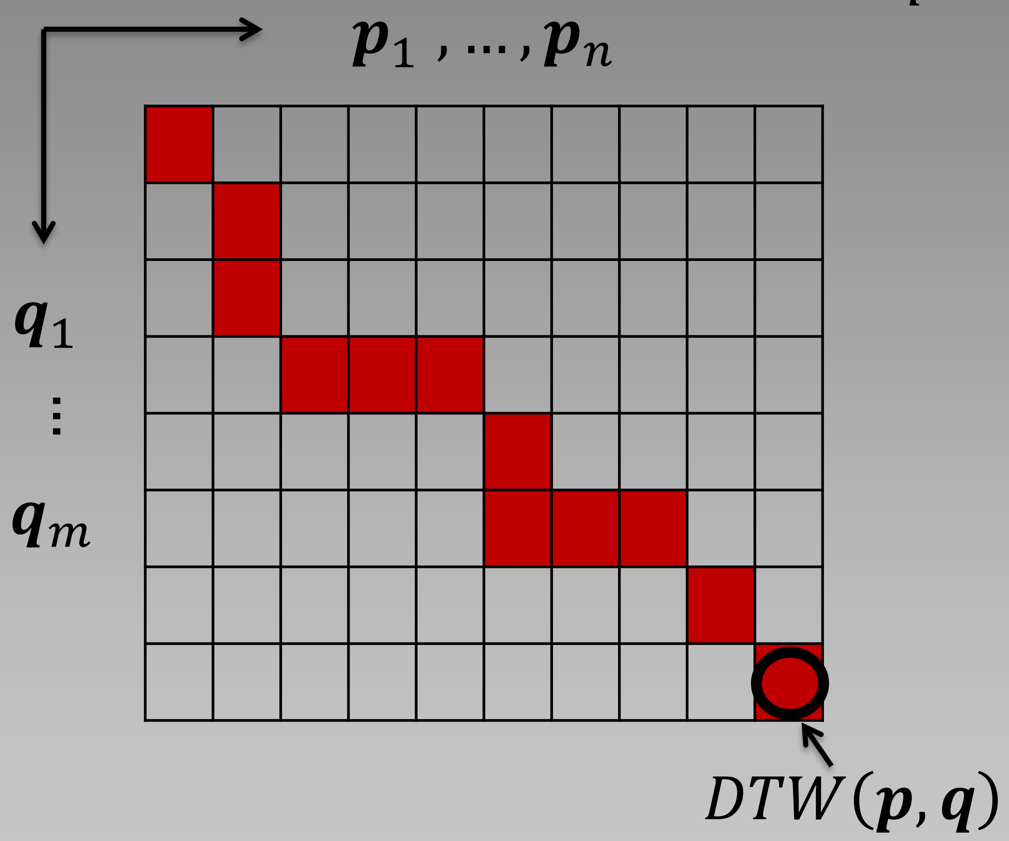
  are they similar ?

- Gestures may be of different time duration ($m \neq n$). How to compare ?

- Apply **Dynamic Time Warping** (DTW):

  constrained, piece-wise linear mapping of the time axes to align the two sequences while minimizing cumulative warping cost.

DTW



$$pathcost(\text{path}, \boldsymbol{p}, \boldsymbol{q}) = \sum_{(i_k, j_k) \in \text{path}} cost(\boldsymbol{p}_{i_k}, \boldsymbol{q}_{j_k})$$

$$cost(\boldsymbol{p}_i, \boldsymbol{q}_j) = \sqrt{\sum_{k=1}^{d} (p_i^k - q_j^k)^2}$$ Euclidean

$$cost(\boldsymbol{p}_i, \boldsymbol{q}_j) = \sum_{k=1}^{d} |p_i^k - q_j^k|$$ Manhattan

$$cost(\boldsymbol{p}_i, \boldsymbol{q}_j) = 1 - \frac{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}{\|\boldsymbol{p}_i\| \|\boldsymbol{q}_j\|}$$ Cosine

Decision rule: The same person or not?

Cost of best alignment (smallest dissimilarity):

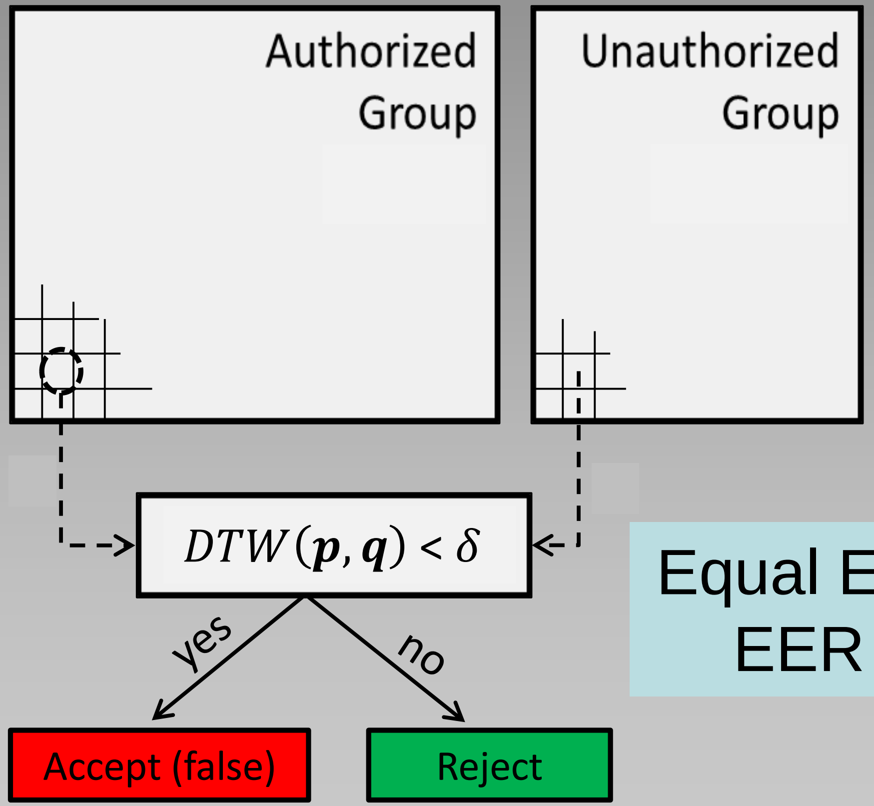
$$DTW(\boldsymbol{p},\boldsymbol{q}) = \min_{\text{path}} pathcost(\text{path},\boldsymbol{p},\boldsymbol{q})$$

$DTW(\boldsymbol{p},\boldsymbol{q}) < \delta \;\rightarrow$ accept as the same person

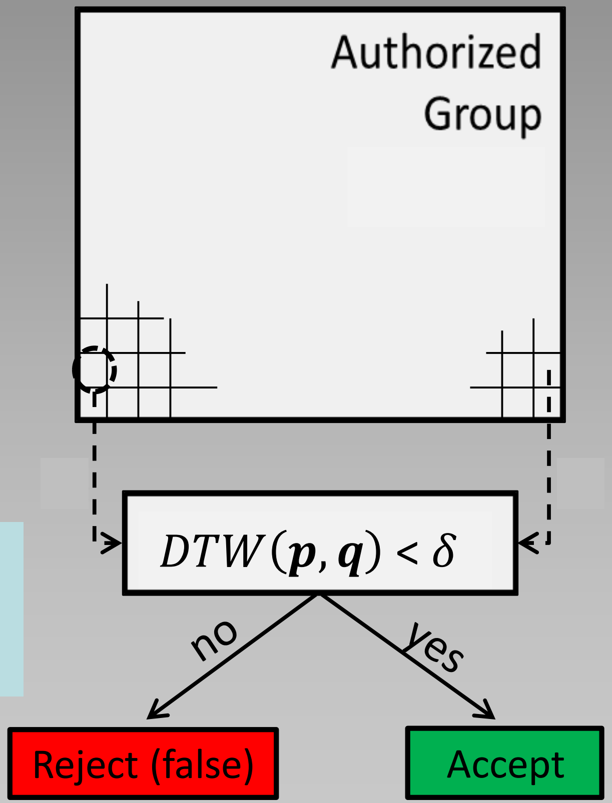
$DTW(\boldsymbol{p},\boldsymbol{q}) > \delta \;\rightarrow$ reject

Evaluation

Unauthorized-user test (False Acceptance Rate – FAR)



Authorized-user test (False Rejection Rate – FRR)



Equal Error Rate (EER):
EER = FAR = FRR

Results: Distance metrics (34 participants)

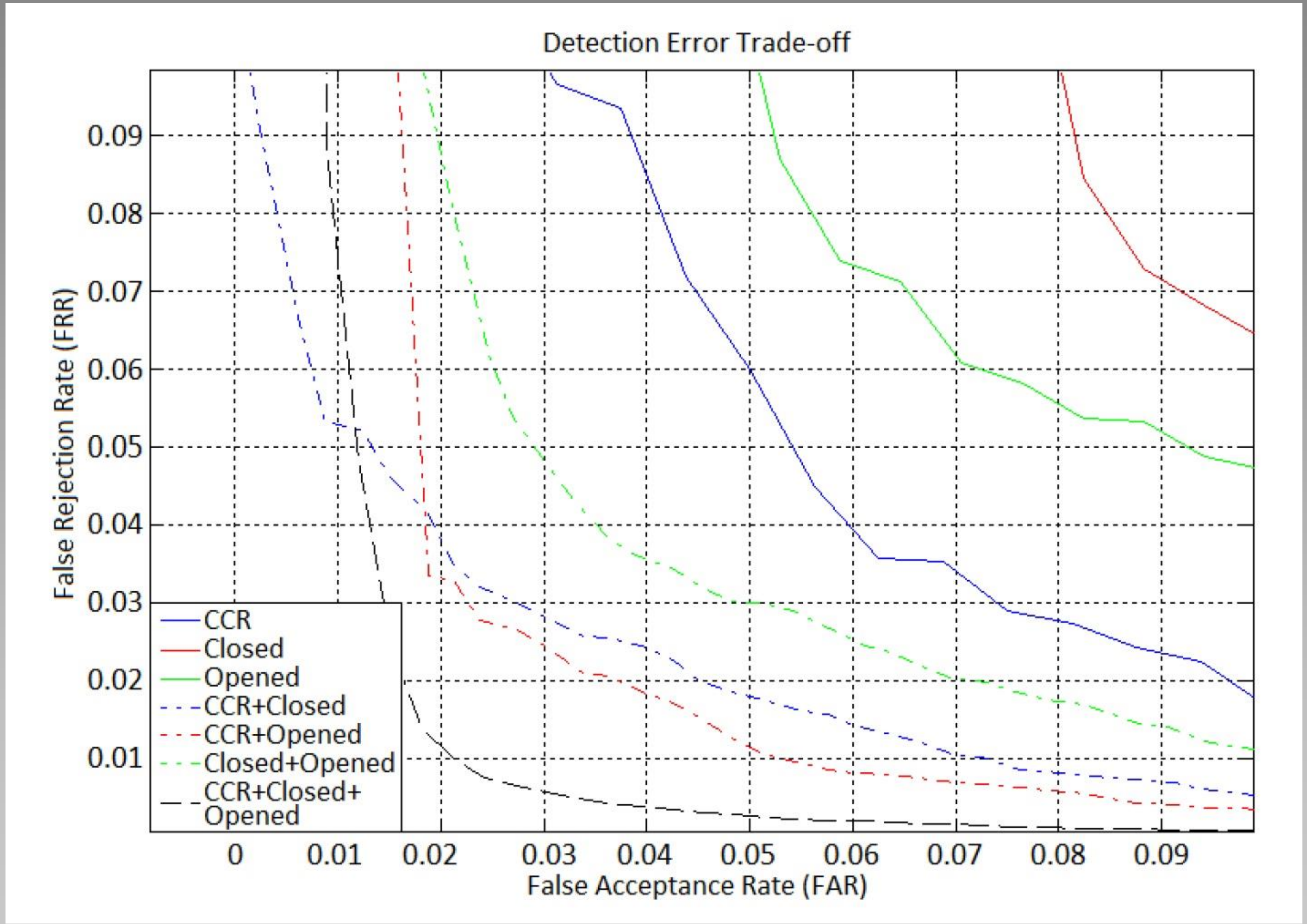
EER FOR DTW DISTANCE FUNCTION OF 20 FEATURES SET WITH THREE DIFFERENT COST FUNCTIONS

Gesture	Manhattan	Euclidean	Cosine
'CCW'	5.50	4.95	8.14
'CW'	7.21	7.26	9.45
'Pinch'	8.34	9.02	9.15
'Drag'	9.50	9.56	8.69
'DDC'	4.46	4.43	8.14
'DUO'	6.80	6.53	8.70
'FBD'	11.53	11.62	13.13
'FBSB'	6.85	7.89	6.61
'FBSA'	9.96	9.84	11.27
'FPCCW'	10.60	10.60	10.63
'FPC'	8.83	8.87	11.46
'FPO'	13.32	14.45	12.42
'FPP'	11.01	10.80	13.85
'FTCCW'	4.48	4.54	5.33
'FTCW'	6.22	6.42	7.98
'FTC'	5.88	5.94	8.88
'FTO'	9.52	9.39	9.98
'FTP'	4.66	4.91	7.36
'Flick'	10.75	10.98	12.85
'Open'	6.80	8.02	9.90
'Swipe'	8.25	9.00	10.14
'User-defined'	2.98	2.85	5.86
Average EER	**7.88**	**8.09**	**9.54**

L1 (Manhattan) norm slightly better than Euclidean norm

[Sae-Bae et al.,TIFS, 2015]

Results: One versus two consecutive gestures



Sequence of
3 gestures better
than 2 gestures
which is better
than 1 gesture

[Sae-Bae et al., TIFS, 2015]

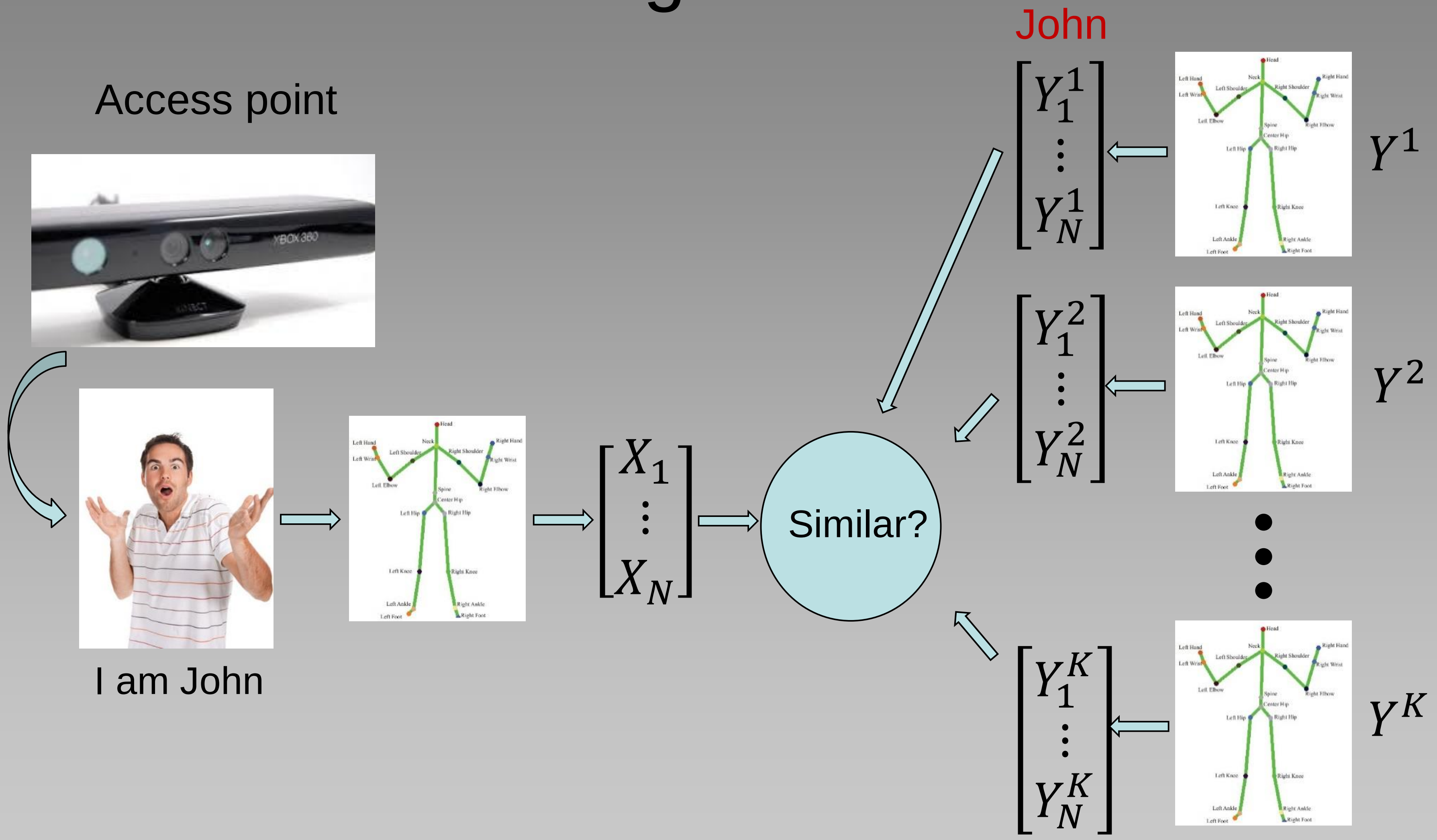
3-D gestures ?

Free-space gestures performed by hands, all limbs or even the whole body:

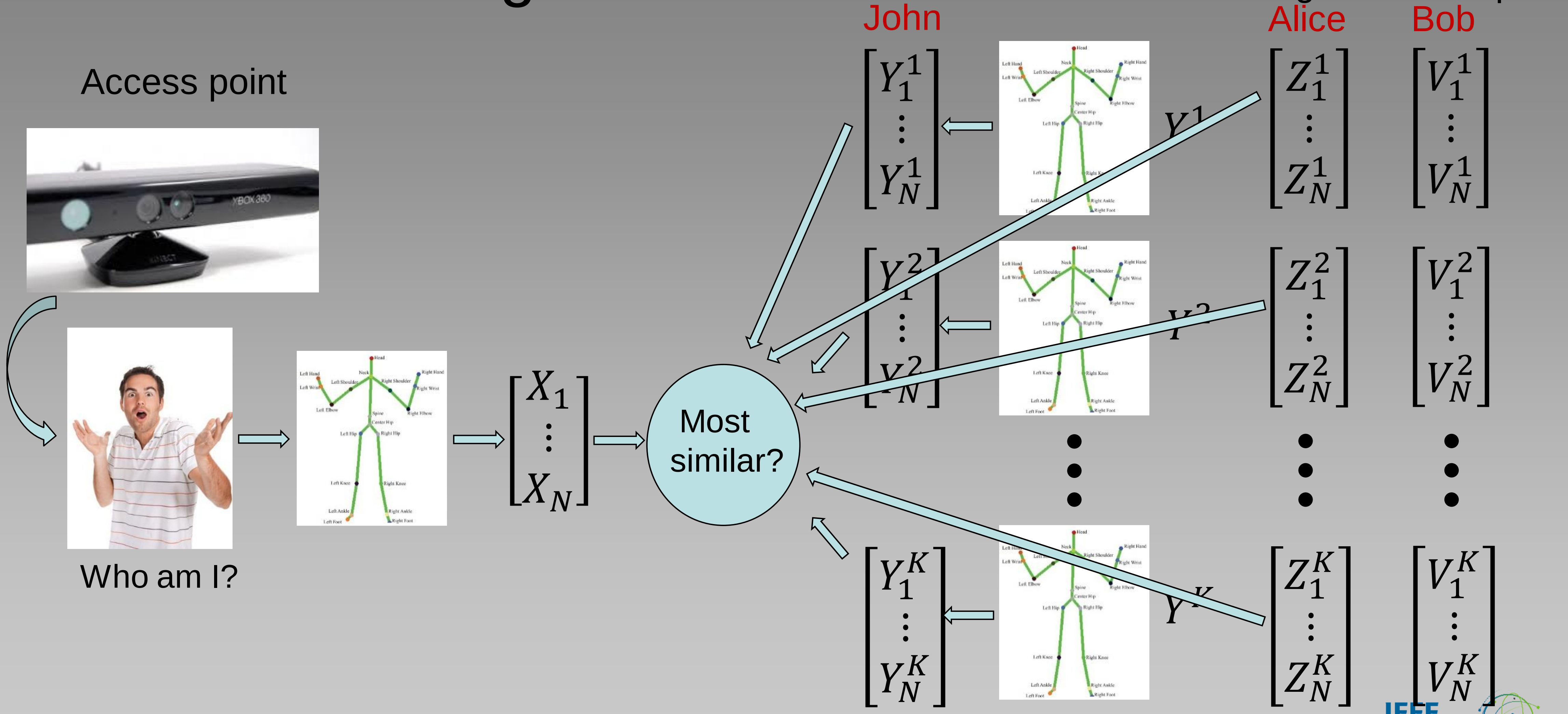
- natural
- can be meaningful, e.g., a hand-wave (easy to memorize)
- biometrically rich

Authentication: Big Picture

Database of enrolled gesture samples

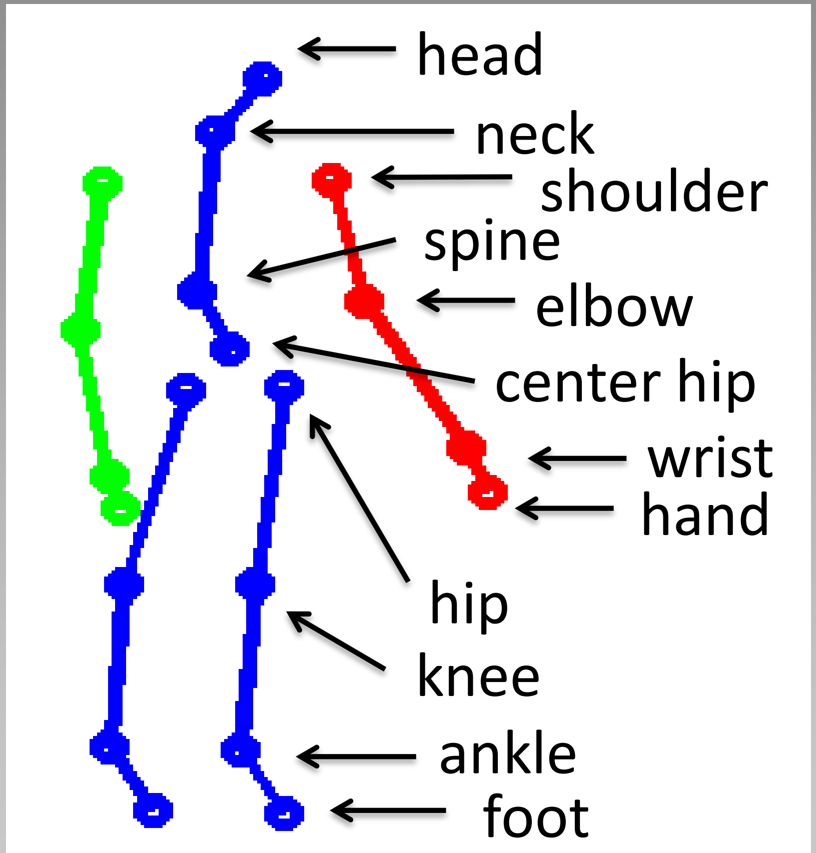


Identification: Big Picture



Method #1: Skeletons

Kinect v1: 20 body joints



Gesture sequence (joint coordinate evolution in time):

$$\boldsymbol{X}_t^g = \begin{pmatrix} \boldsymbol{x}_{1,t}^g \\ \dots \\ \boldsymbol{x}_{j,t}^g \end{pmatrix}$$

g – gesture
t – time
j – joint number (1,…,20)
$\boldsymbol{x}$ – point coordinates

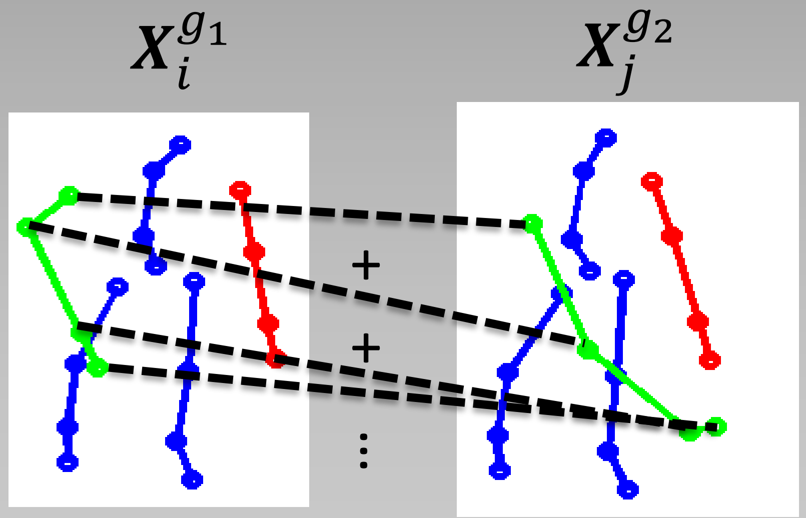
[Wu, Ishwar, Konrad, ICASSP, 2013]

How to find similarity?

Gesture-sequence 1: $\boldsymbol{X}^{g_1} = (\boldsymbol{X}_1^{g_1}, \boldsymbol{X}_2^{g_1}, \ldots \boldsymbol{X}_n^{g_1})$

Gesture-sequence 2: $\boldsymbol{X}^{g_2} = (\boldsymbol{X}_1^{g_2}, \boldsymbol{X}_2^{g_2}, \ldots \boldsymbol{X}_n^{g_2})$

Align them to account for variation in execution speed, and then measure the distance between aligned sequences: **Dynamic Time Warping** (DTW)



DTW

$$pathcost(\text{path}, \boldsymbol{X}^{g_1}, \boldsymbol{X}^{g_2}) = \sum_{(i_k, j_k) \in \text{path}} cost\left(\boldsymbol{X}_{i_k}^{g_1}, \boldsymbol{X}_{j_k}^{g_2}\right)$$

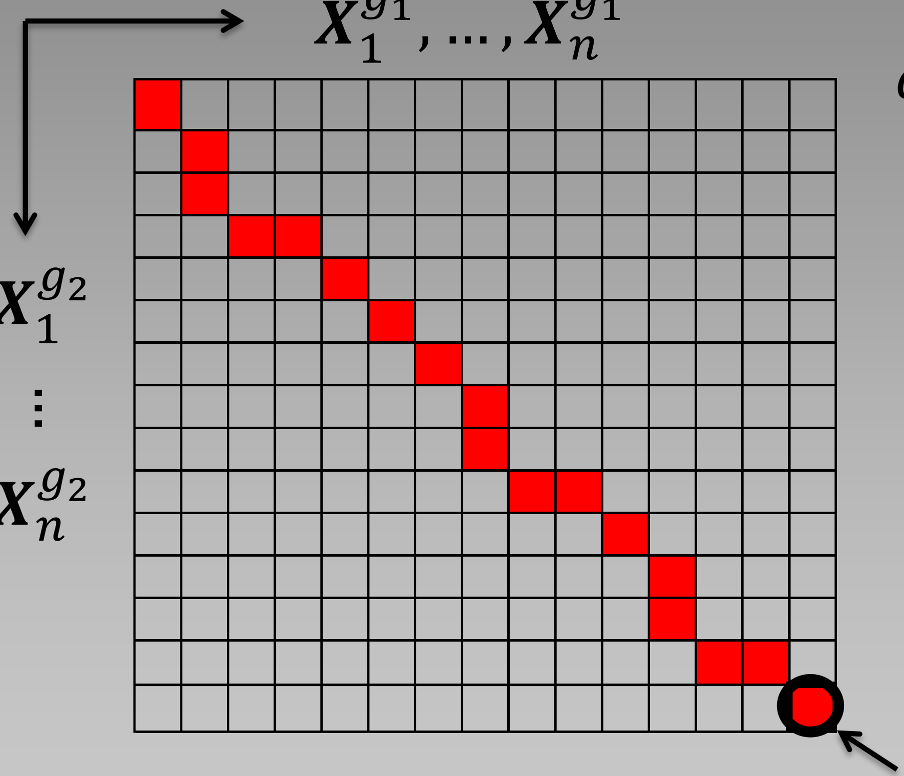
$$cost\left(\boldsymbol{X}_i^{g_1}, \boldsymbol{X}_j^{g_2}\right) = \sum_{p=1}^{d} \left\| \boldsymbol{x}_{p,i}^{g_1} - \boldsymbol{x}_{p,j}^{g_2} \right\|$$

$\boldsymbol{X}_1^{g_1}, \dots, \boldsymbol{X}_n^{g_1}$

$\boldsymbol{X}_1^{g_2}$

$\vdots$

$\boldsymbol{X}_n^{g_2}$



$DTW(\boldsymbol{X}^{g_1}, \boldsymbol{X}^{g_2})$

Cost of best alignment:

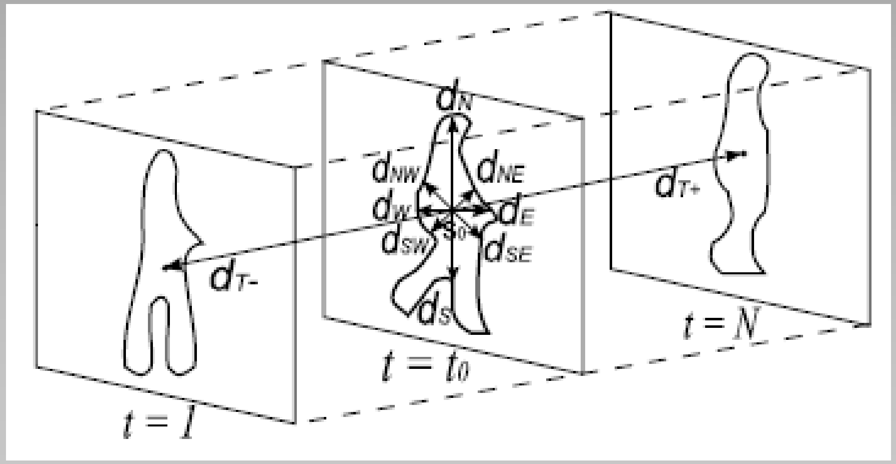
$$DTW(\boldsymbol{X}^{g_1}, \boldsymbol{X}^{g_2}) = \min_{\text{path}} pathcost(\text{path}, \boldsymbol{X}^{g_1}, \boldsymbol{X}^{g_2})$$

If $DTW(\boldsymbol{X}^{g_1}, \boldsymbol{X}^{g_2}) < \delta$, **accept as the same person**

If $DTW(\boldsymbol{X}^{g_1}, \boldsymbol{X}^{g_2}) > \delta$, **reject**

Method #1: Silhouettes

[Lai, Konrad, Ishwar, AVSS, 2012]



13-dimensional feature vector at each silhouette pixel:

$$f_n = \{x, y, t, d_E, d_S, d_W, d_N, d_{NE}, d_{SE}, d_{SW}, d_{NW}, d_{T+}, d_{T-}\}$$

… but this expands dimensionality

Silhouette-based method

Dimensionality reduction *via* 13x13 **covariance matrix**:

$$C = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{f}_n - \boldsymbol{\mu})(\boldsymbol{f}_n - \boldsymbol{\mu})^T, \quad \boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{f}_n$$

Distance metric: log-covariance [Arsigny et al., MRIM, 2006]

$$D(C_1, C_2) = \left\| \log(C_1) - \log(C_2) \right\|_2$$

If $D(C_1, C_2) < \delta$, **accept as the same person**

If $D(C_1, C_2) > \delta$, **reject**

Authentication performance: EER for simple gestures

20 participants

	Skeleton-based			Silhouette-based		
Group Split	19/1	15/5	10/10	19/1	15/5	10/10
Right Swing	3.98%	3.98%	3.98%	4.04%	4.04%	4.01%
Right Push	2.03%	2.03%	1.98%	3.74%	3.73%	3.73%
Right Back	1.01%	1.00%	1.03%	0.00%	0.00%	0.00%
Left Swing	1.12%	1.11%	1.11%	2.01%	2.01%	2.01%
Left Push	2.02%	2.01%	1.96%	2.01%	2.01%	2.01%
Left Back	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Zoom-in	1.02%	1.02%	0.97%	2.45%	2.45%	2.45%
Zoom-out	2.59%	2.59%	2.59%	7.97%	8.02%	7.83%
All gestures	1.89%	1.89%	1.89%	2.79%	2.73%	2.73%

Skeleton-based method performs slightly better than silhouette-based method

Identification performance: EER for simple gestures

Skeleton-based Silhouette-based 20 participants

Group Split	19/1	15/5	10/10	19/1	15/5	10/10
	Skeleton-based			Silhouette-based		
Right Swing	6.02%	6.02%	5.28%	7.07%	6.98%	5.74%
Right Push	3.99%	3.22%	2.91%	8.11%	8.31%	8.70%
Right Back	1.01%	1.01%	1.00%	0.00%	0.00%	0.00%
Left Swing	4.08%	4.02%	2.99%	4.03%	4.03%	3.99%
Left Push	9.05%	8.58%	7.61%	5.04%	4.99%	4.04%
Left Back	1.01%	0.99%	1.01%	0.00%	0.00%	0.00%
Zoom-in	5.02%	4.94%	4.10%	9.57%	9.05%	7.99%
Zoom-out	7.97%	6.31%	5.71%	10.95%	8.95%	7.65%
All gestures	4.14%	4.12%	3.51%	6.92%	6.49%	6.16%

Skeleton-based method performs slightly better than silhouette-based method

Identification performance worse than authentication performance (as expected)

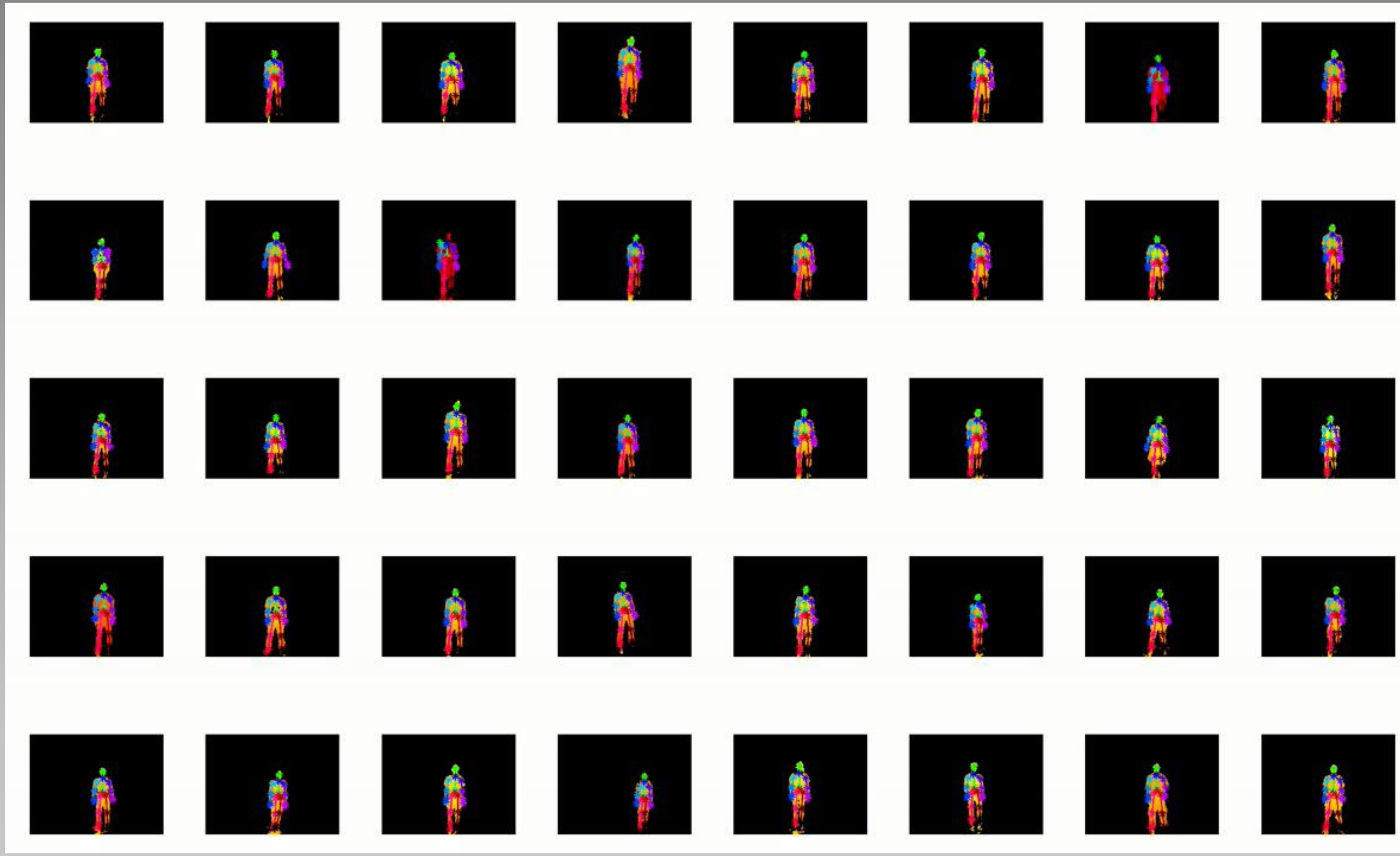
Degradation study: More complex gestures

- 40 participants (27 males/13 females), 2 different gestures:
  - S-shaped movement of both arms
  - User-defined

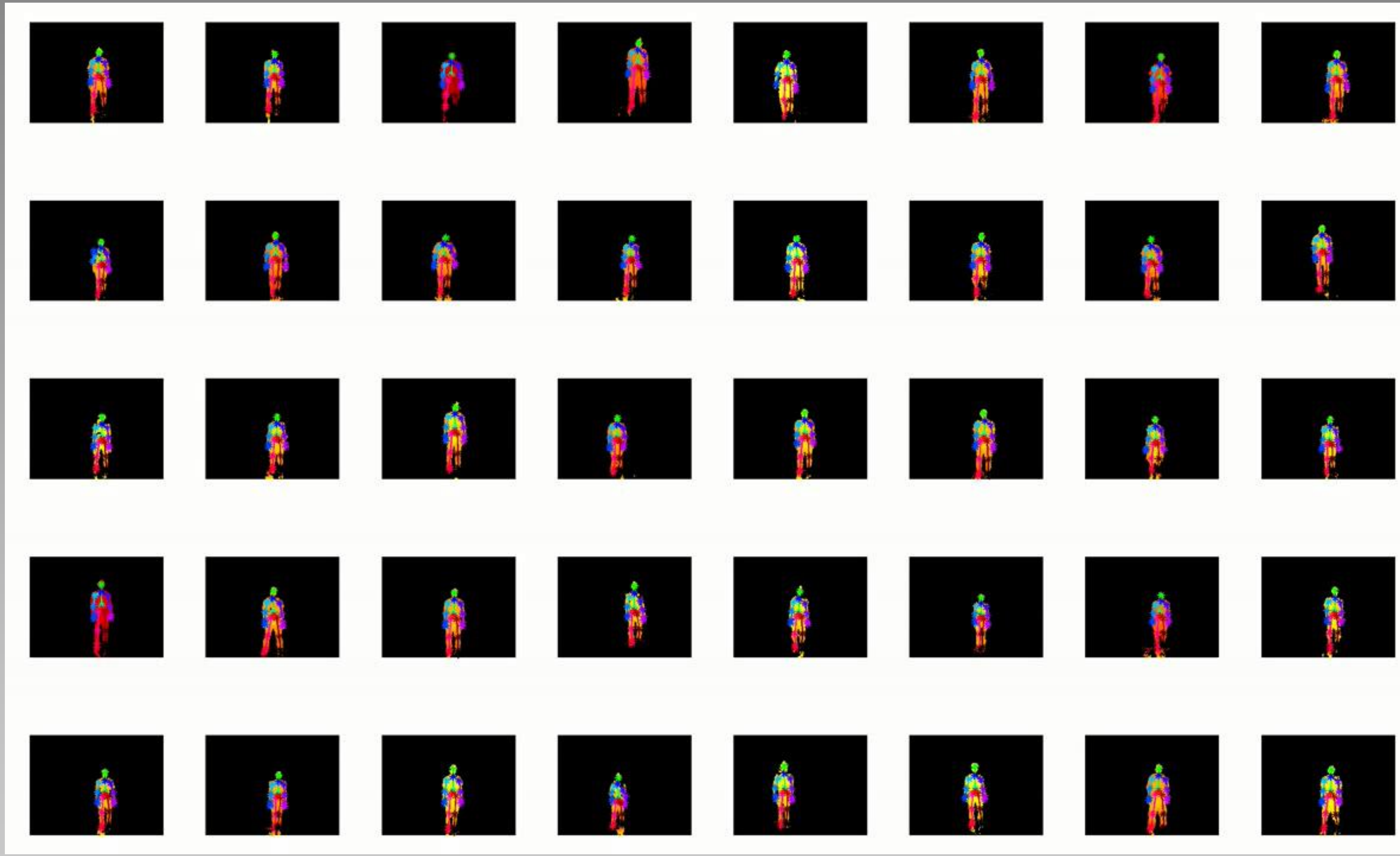
- 20 repetitions of each gesture in 2 sessions:
  - Session 1: test of changing appearance:
    - 5 "clean" gestures (no coats, bags),
    - 5 gestures with either a coat or bag,
  - Session 2: to test time and memory (after 1 week):
    - 5 gestures performed from memory

[Wu, Konrad, Ishwar, AVSS, 2014]

S-gesture



User-defined gesture



Authentication performance

	Train with…	Test with…	EER Silhouette Log-Cov.	EER Skeleton DTW
S-gesture	No degradations	No degradations	3.46%	5.26%
		Personal-effects	11.13%	6.56%
		User memory	17.62%	13.42%
User-defined gesture	No degradations	No degradations	1.12%	0.30%
		Personal-effects	2.51%	0.68%
		User memory	12.14%	2.97%

Identification performance

100% - CCR

	Train with…	Test with…	Silhouette Log-Cov.	Skeleton DTW
S-gesture	No degradations	No degradations	2.50%	1.00%
		Personal-effects	16.00%	5.50%
		User memory	42.50%	21.00%
User-defined gesture	No degradations	No degradations	1.00%	0.00%
		Personal-effects	3.06%	1.02%
		User memory	19.00%	5.00%

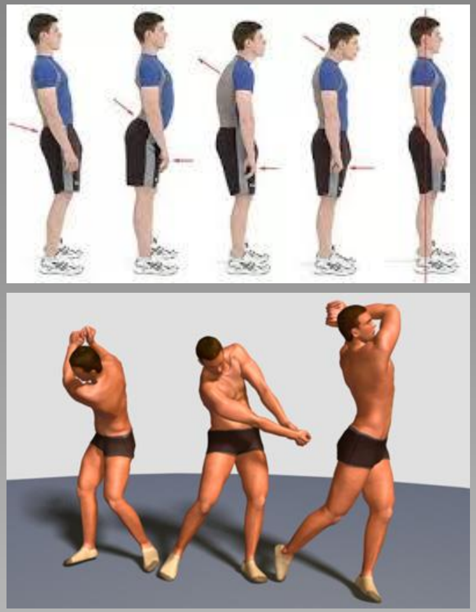
CCR = Correct Classification Rate

Silhouettes or skeletons ?

- Silhouettes work well for clean data
- Heavy clothing, backpacks degrade performance of both, but skeletons are more robust
- Elapsed time degrades performance, but user-defined gesture performs better especially using skeletons
- Pre-defined gestures work well, but user-defined ones work even better (EER ≈ 1%)

Value of posture, build and dynamics

Biometric information
- static information [**body posture, build**]
- dynamic information [**limb motion**]



Approach:

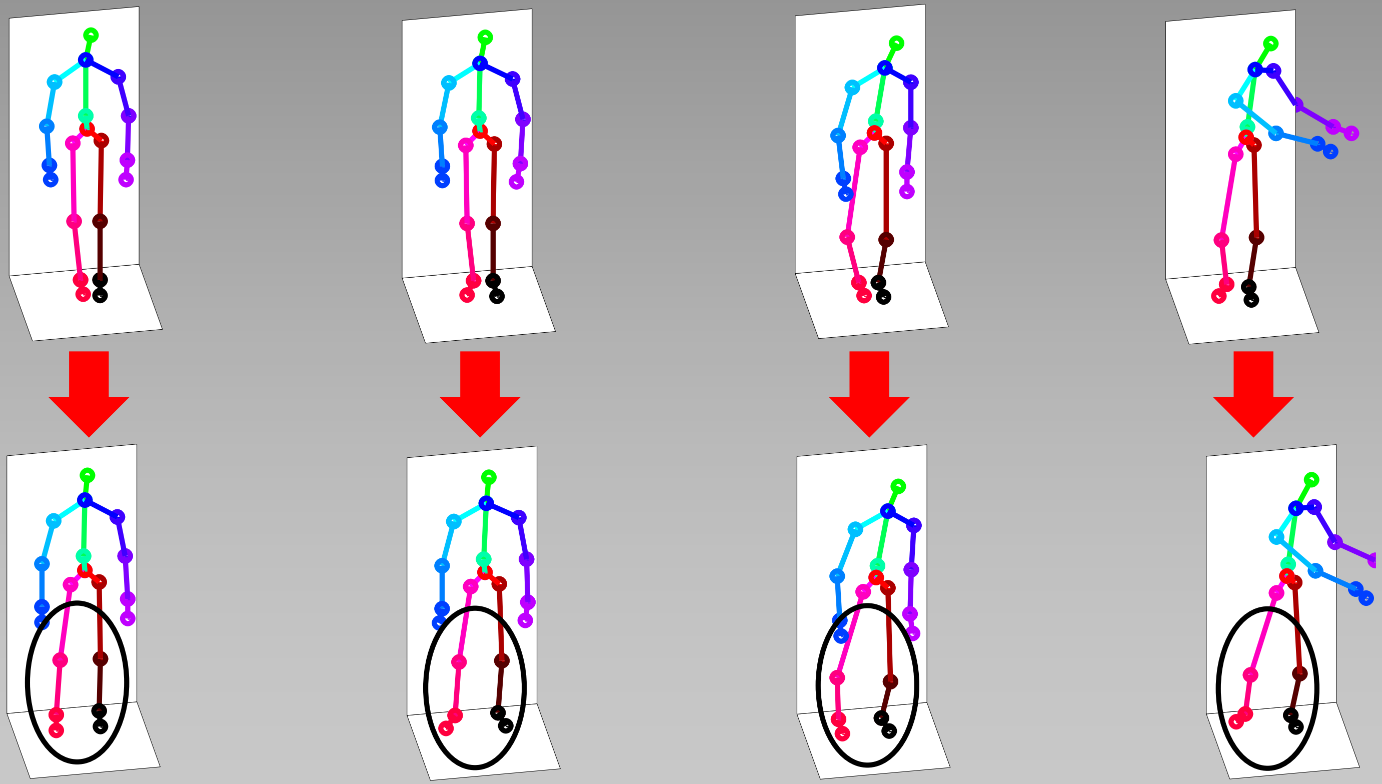
- Suppress various combinations of posture, build and dynamics, and evaluate authentication performance
- Train attackers by showing a gesture video of their easiest "victim" (one with the most similar gesture)

[Wu, Ishwar, Konrad, IJCB, 2014]

Suppressing user posture

Method: User-specific posture → Standard posture

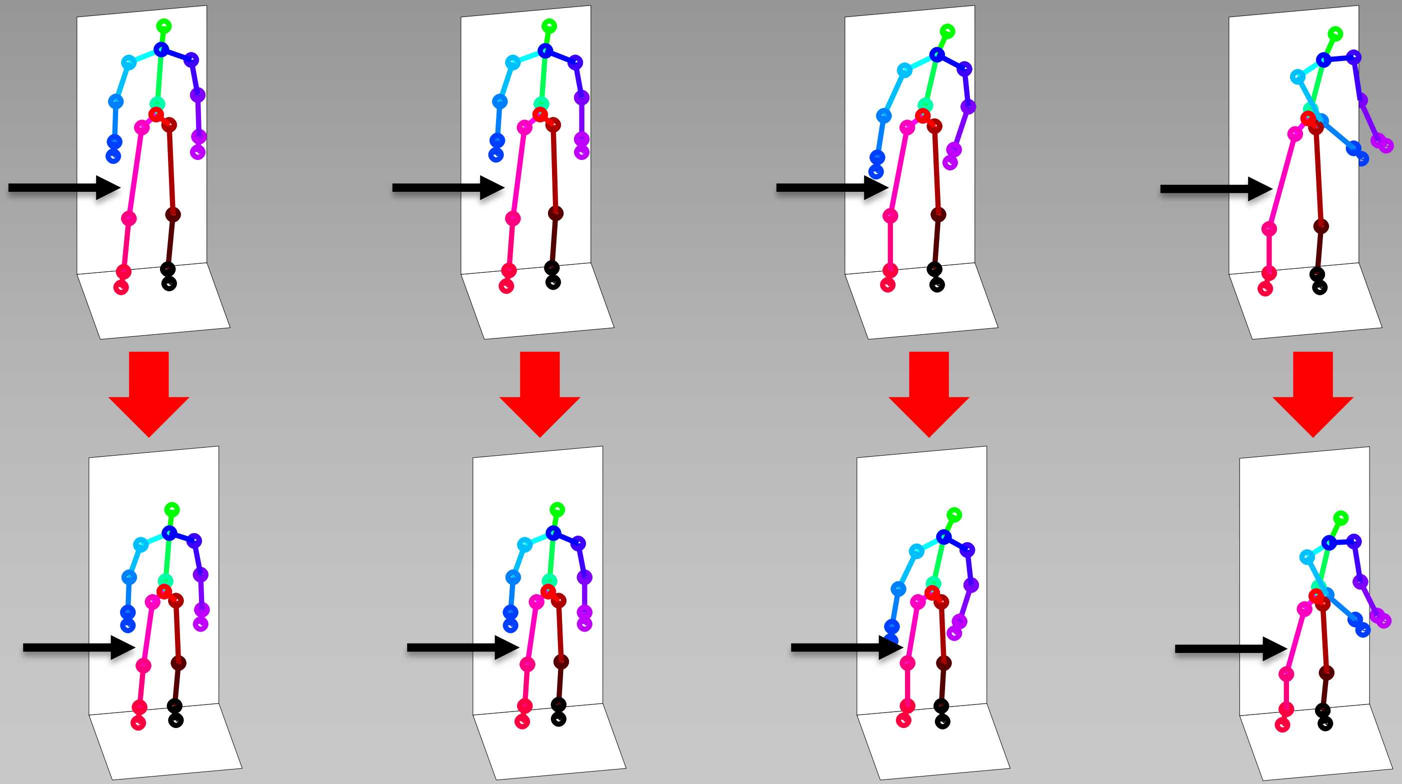
Standard posture = Average of all user initial postures



Suppressing user build

Method: User-specific build → Standard user build

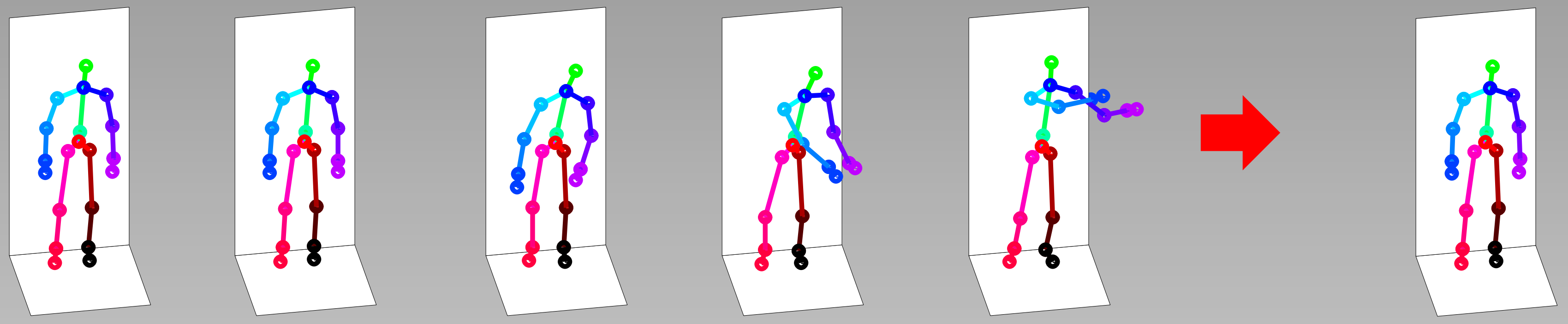
Standard user build = Average of all user limb proportions



Suppressing user dynamics

Method: Suppress limb motion

Simply discard all but the first frame



Results of suppression (36 participants)

	EER for 3 different gestures		
Information Suppressed	**Left-Right**	**Double-handed arch**	**Balancing**
Nothing	1.97%	0.25%	0.68%
Dynamics	3.83%	3.01%	2.12%
Build	2.09%	0.38%	1.20%
Posture	3.75%	0.61%	1.30%
Dynamics + Build	4.29%	4.88%	3.72%
Dynamics + Posture	8.22%	4.76%	4.39%
Posture + Build	6.91%	0.91%	3.22%

Dynamics affect performance more than posture and build

Spoofing study

- Attackers matched to their closest “victims” (similar gesture performance)
- In ``Matched-Spoof’’, the attacker is allowed to study ``victim’s’’ gesture for 1 minute and practice simultaneously seeing ``victim’s’’ and own gesture

Gesture	Matched Zero-Effort EER	Matched Spoof EER
Left-right	2.78%	2.35%
Double-handed arch	1.24%	1.13%
Balancing	2.66%	2.06%

- Surprise: EER improves after spoofing; it suggests that it is difficult to imitate someone’s gesture

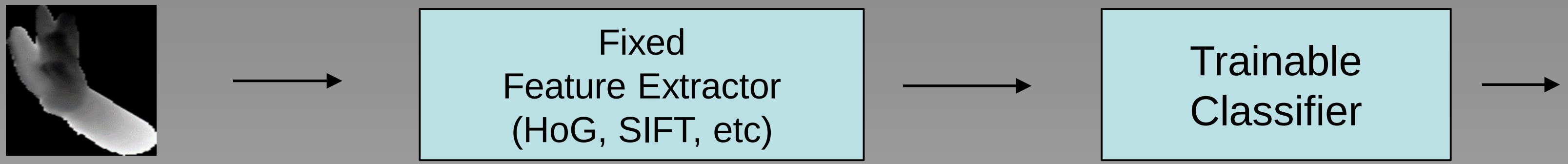
Learning user style

- So far, samples of a user's gesture must be enrolled
- Is it possible to recognize a user regardless of gesture ?
- "Reverse" of gesture recognition
  - Gesture recognition: Learn gesture invariant of user
  - User recognition: Learn user invariant of gesture
- Method: **Deep Convolutional Neural Networks**

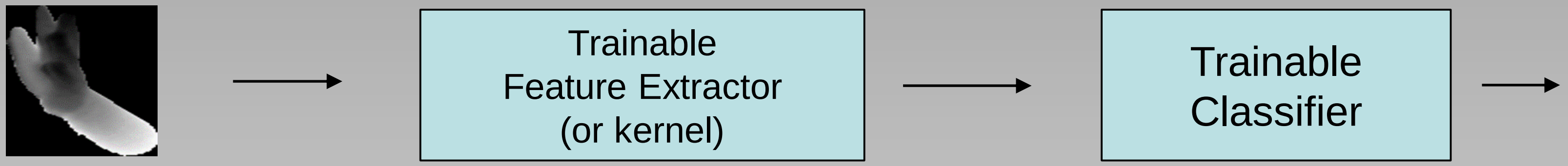
[Wu, Ishwar, Konrad CVPRW, 2016]

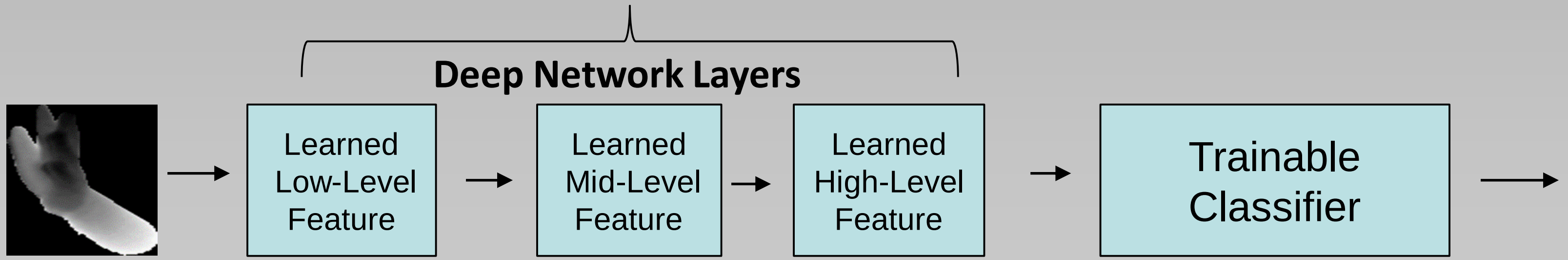
Deep learning

- Traditional Learning

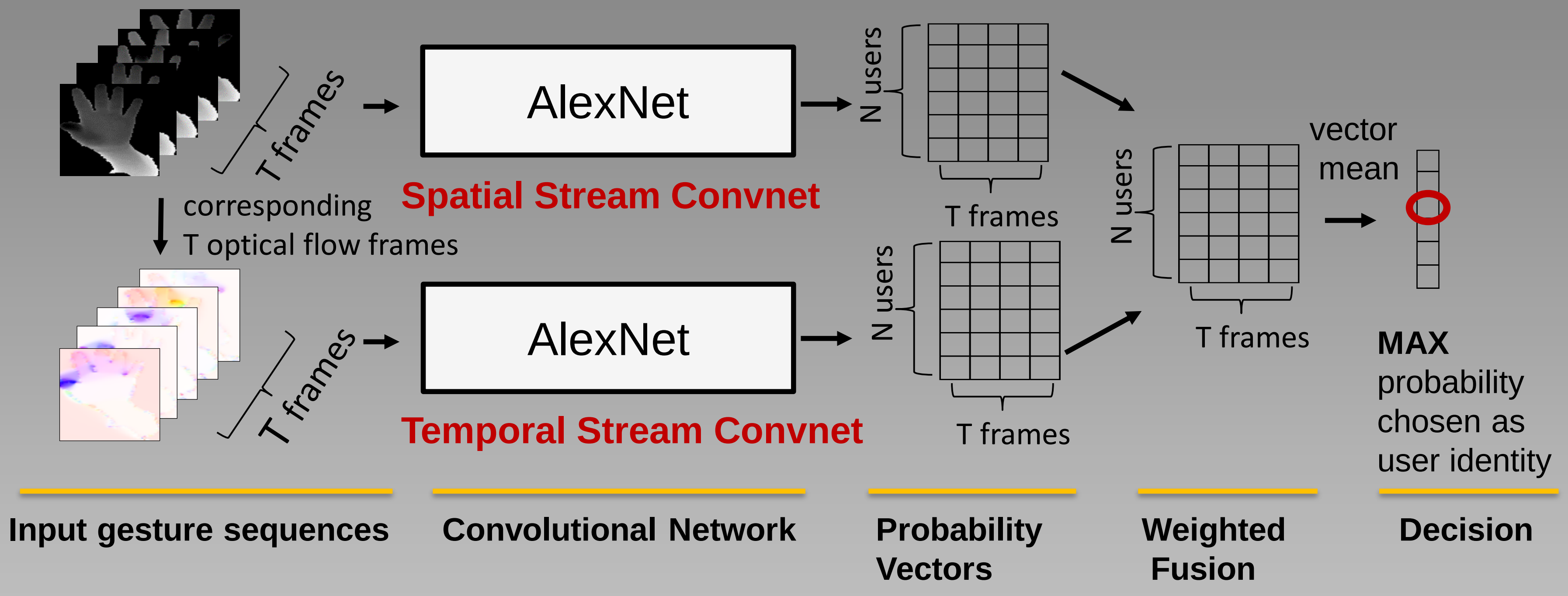


- Deep Learning Pipeline
  - Learn feature representation directly from image
  - Hidden “weight” layers are a composition of non-linear transformations





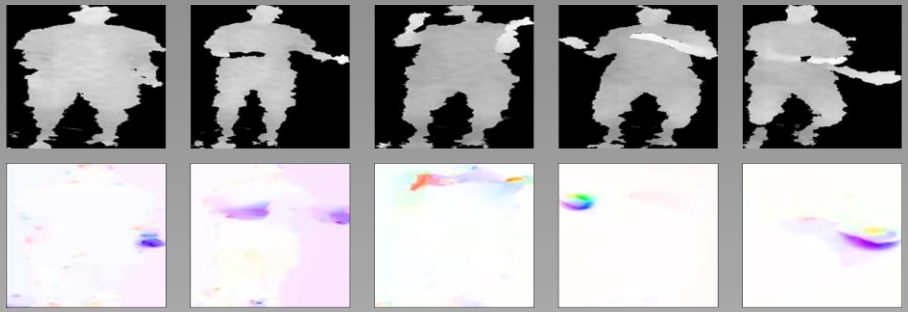
Two-Stream Convolutional Neural Network (CNN)



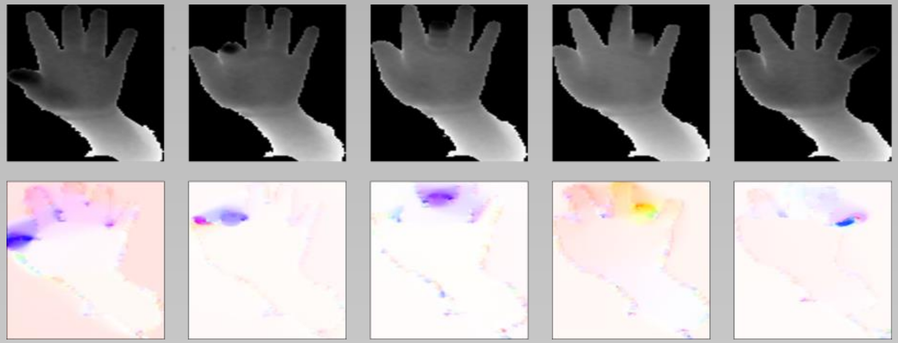
- Adapt a two-stream CNN architecture for identification
- Learn two separate image-based CNNs
- AlexNet used as the CNN of choice; pre-trained from ImageNet, then fine-tuned

Gesture datasets

- Body Gesture Dataset (BodyLogin): 40 users, 5 gestures (1 user-defined)



- Hand Gesture Dataset (HandLogin): 21 users, 4 gestures



Experiments: User identification

- Evaluate Correct Classification Error (CCE = 100% - CCR)

  1. Training and testing with all gestures

  2. Testing with gestures unseen in training (left-out) to evaluate generalization performance

- **Baseline**: silhouette-covariance method over 3 temporal scales (7 covariance matrices concatenated together)

Results: Training and testing with all gestures

CCE

Dataset	Spatial (1,0)	(0.66,0.33)	(0.5,0.5)	(0.33,0.66)	Temporal (0,1)	Baseline
HandLogin	0.24%	0.24%	0.24%	0.71%	4.05%	6.43%
BodyLogin	0.05%	0.05%	0.05%	0.05%	5.01%	1.15%

- Significant improvement over baseline

Results: Testing with gestures unseen in training

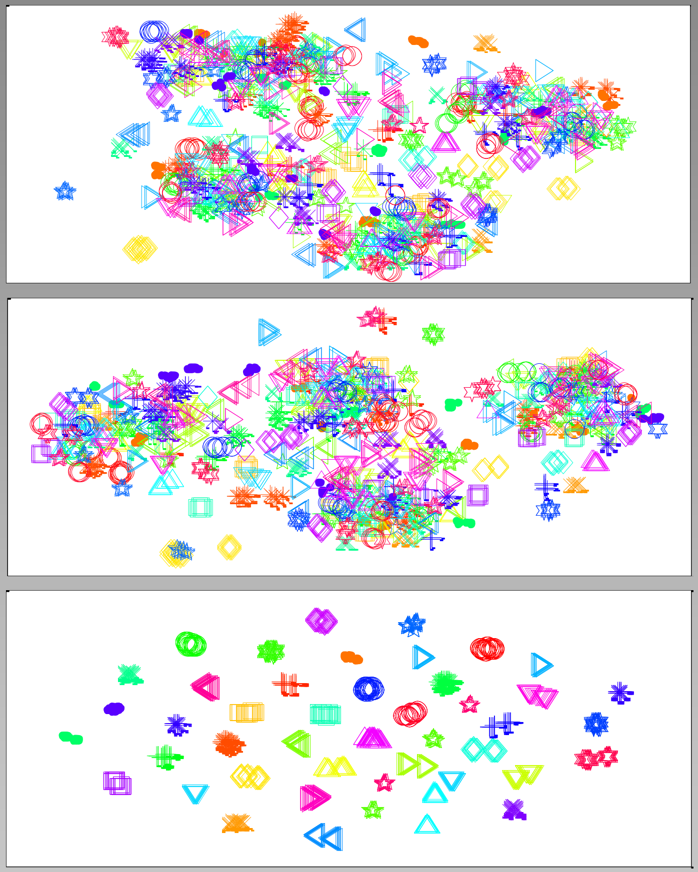
CCE

	Generalizing Gesture	← Spatial (1,0)	(0.66,0.33)	(0.5,0.5)	(0.33,0.66)	Temporal → (0,1)	Baseline
HandLogin	Compass	2.38%	2.86%	4.76%	8.57%	36.19%	82.38%
	Piano	1.91%	0.48%	1.43%	1.91%	12.86%	68.10%
	Push	44.29%	49.05%	54.29%	67.62%	77.14%	79.52%
	Fist	16.67%	15.71%	17.14%	20.00%	31.43%	72.38%
BodyLogin	S motion	0.75%	1.00%	1.25%	1.75%	16.75%	75.75%
	Left-Right	0.88%	1.25%	1.50%	1.88%	11.50%	80.88%
	2-Hand Arch	0.13%	0.13%	0.13%	0.38%	6.25%	74.50%
	Balancing	9.26%	10.01%	13.27%	19.52%	45.06%	77.97%
	User Defined	5.28%	5.53%	6.16%	8.54%	22.49%	71.61%

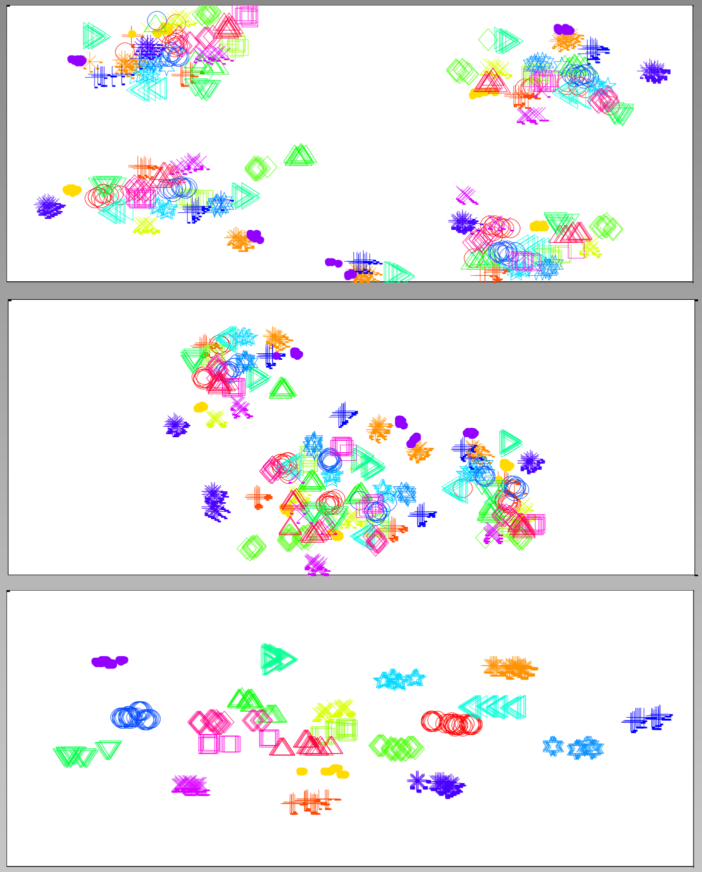
- Strong generalization for similar gestures
- Baseline incapable of generalizing

Feature visualization with t-SNE

BodyLogin



HandLogin



Color coding by user

Baseline

Pre-trained (ImageNet)

Fine-tuned

Strong user separation after fine-tuning

Final thoughts

- Authentication ``anxiety'' will only grow
- Juggling hundreds of passwords is not sustainable
- **Solution:** Leverage renewable biometrics *via* Natural User Interfaces
- **Bonus:** Authentication using NUIs has been shown to increase pleasure and excitement for user-defined gestures
- **Challenge:** How to develop practical authentication systems on NUIs that are robust under a wide range of circumstances ?