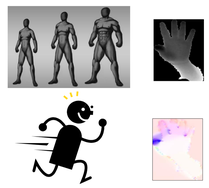
Two-Stream CNNs for Gesture-Based User Recognition: Learning User Style

Jonathan Wu, Prakash Ishwar, and Janusz Konrad
{jonwu, pi, jkonrad}@bu.edu

Motivation

- Use **body** or **hand** gestures to recognize users

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



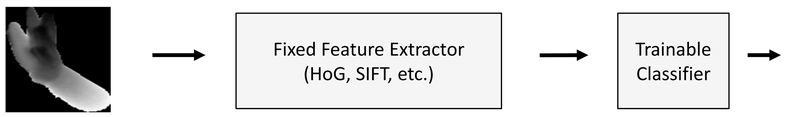
- Can obtain gesture depth-maps with **time-of-flight** cameras (Kinect V1 and V2)

- **Previous works:** compared gesture representations, assessed the value of multiple views, studied spoof attacks
- **Focus of this study:**

→ Use two-stream convolutional networks to recognize users
→ Evaluate gesture generalization performance: **learn** user style
→ Visualize deep gesture features with t-SNE

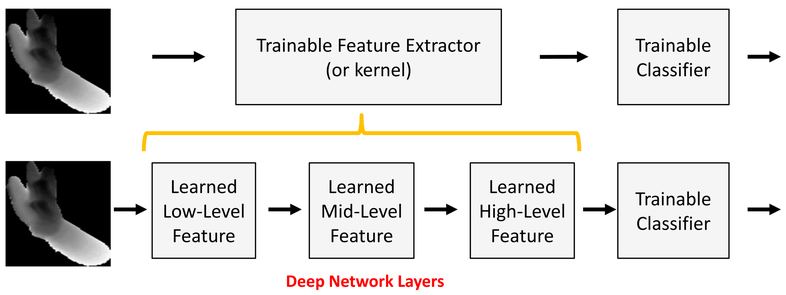
Deep Learning Recap

Traditional Learning Pipeline



Deep Learning Pipeline

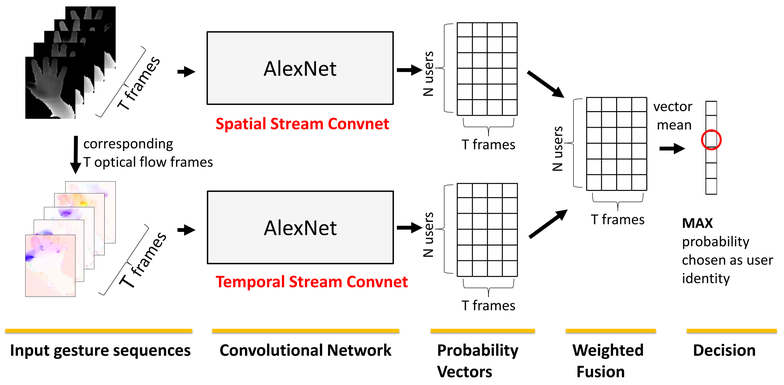
- Learn feature representation directly from image (end-to-end learning)
- Hidden "weight" layers are a composition of non-linear transformations



Two-Stream Convolutional Networks

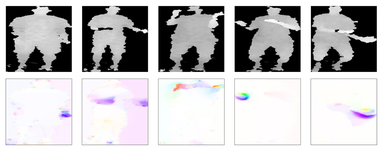
Method: Adapt a two-stream convolutional network [1] for user **identification**

- Leverages static and dynamic information of a gesture
- Learns two separate image-based convolutional networks
- AlexNet [2] used as network of choice (5 conv layers, 3 fc layers)
- Pre-trained from ImageNet [3], then fine-tuned

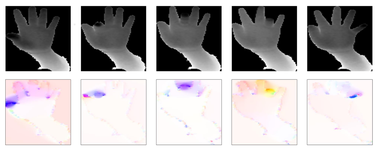


Public Gesture Datasets

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

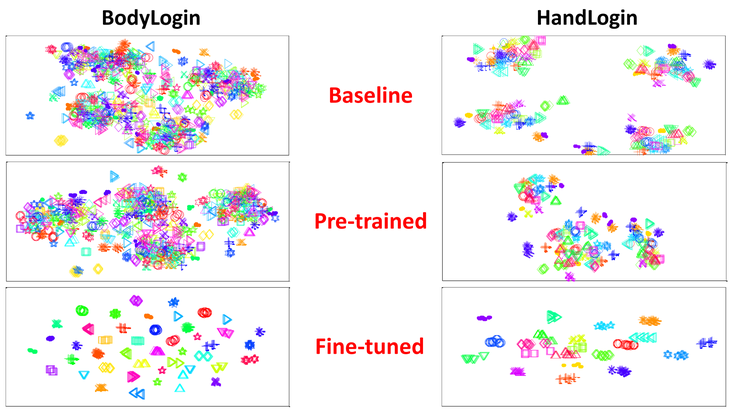


Hand Gesture Dataset (HandLogin):
21 users, 4 gestures



Feature Visualization

- Marker Point ⇔ Gesture, Marker Color **and** Shape ⇔ user identity
- t-SNE [4] shows strong user separation after fine-tuning



User Identification Experiments

Evaluate **Correct Classification Error (CCE = 100% - CCR)** for various scenarios:

1. Training and testing with all gestures

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%

2. Testing with gestures unseen in training (left-out), evaluates generalization performance

	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%

3. Suppression of dynamics (use first few frames only in training and testing)

Dataset	Scenario	Spatial Stream	Baseline
HandLogin	All frames	0.24%	6.43%
	No dynamics	1.90% ↑	9.29% ↑
BodyLogin	All frames	0.05%	1.15%
	No dynamics	1.00% ↑	32.60% ↑

*Baseline: temporal hierarchy of depth-aware silhouette tunnels [5]

Conclusions

- Quite feasible to learn a user's gesture style from a bank of gesture types
- Possible to generalize user style to similar gestures with only slight degradation in performance
- Convolutional networks offer drastic improvements over state-of-the-art
- Temporal/dynamic information always valuable

- **Check out paper:** additional experiments and analysis
- Data and trained models **available** online **[see QR-code]** (http://vip.bu.edu/projects/hcis/deep-login)

CVPR 2016

This work is supported in part by the National Science Foundation under award CNS-1228869.

[1] Simonyan et. al. *Two-stream convolutional networks for action recognition in videos*. NIPS 2014.
[2] Krizhevsky et. al. *Imagenet classification with deep convolutional neural networks*. NIPS 2012.
[3] Russakovsky et. al. *Imagenet large scale visual recognition challenge*. IJCV 2015.
[4] Van der Maaten et. al. *Visualizing data using t-SNE*. JMLR 2008.
[5] Wu et. al. *Leveraging Shape and Depth in User Authentication from In-Air Hand Gestures*. ICIP 2015.