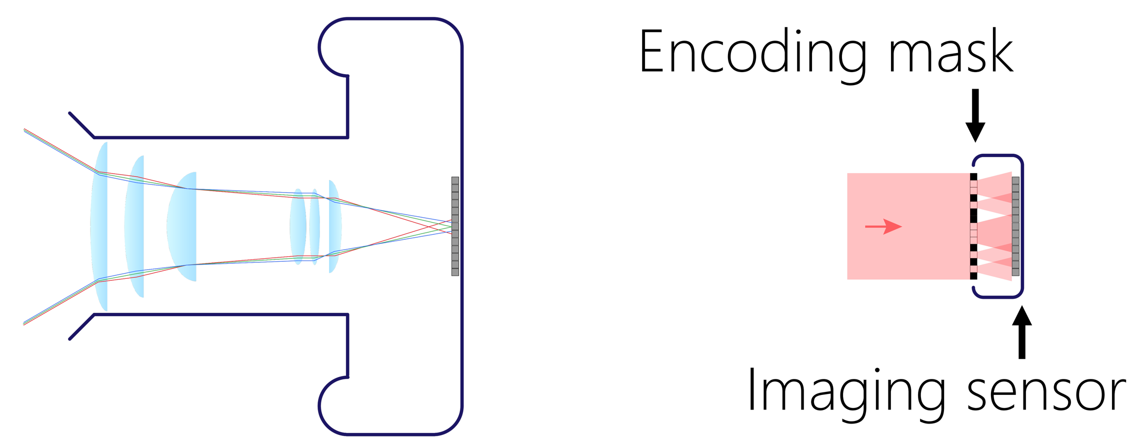
Privacy-Preserving Action Recognition using Coded Aperture Videos

Northwestern University | Microsoft | UF University of Florida | Zillow GROUP

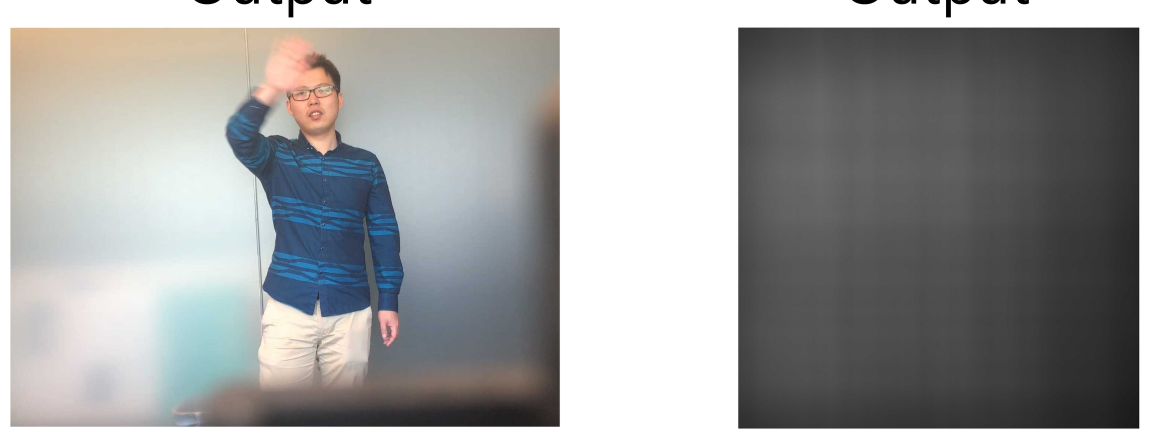
Z. W. Wang[1], V. Vineet[2], F. Pittaluga[3], S. N. Sinha[2], O. Cossairt[1], S. B. Kang[4]

[1]Northwestern University, [2]Microsoft Research, [3]University of Florida, [4]Zillow Group

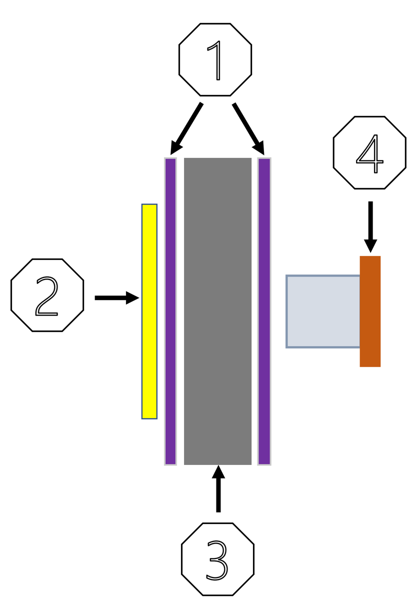
What is a coded aperture camera?



Conventional cam. Output | Lens-free CA cam. Output



Building a lens-free CA camera



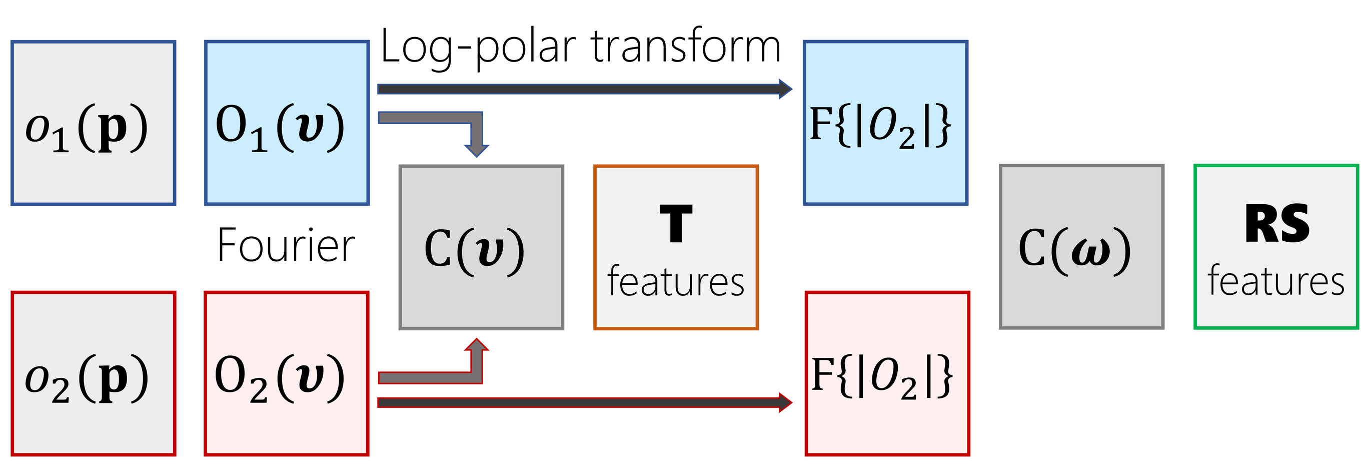
1. Polarizing filters.
2. 550nm long-pass filter
3. Spatial light modulator (SLM)
4. Camera board

Vision from CA images?

5-class image classification	
gray images >95%	CA images ~60%

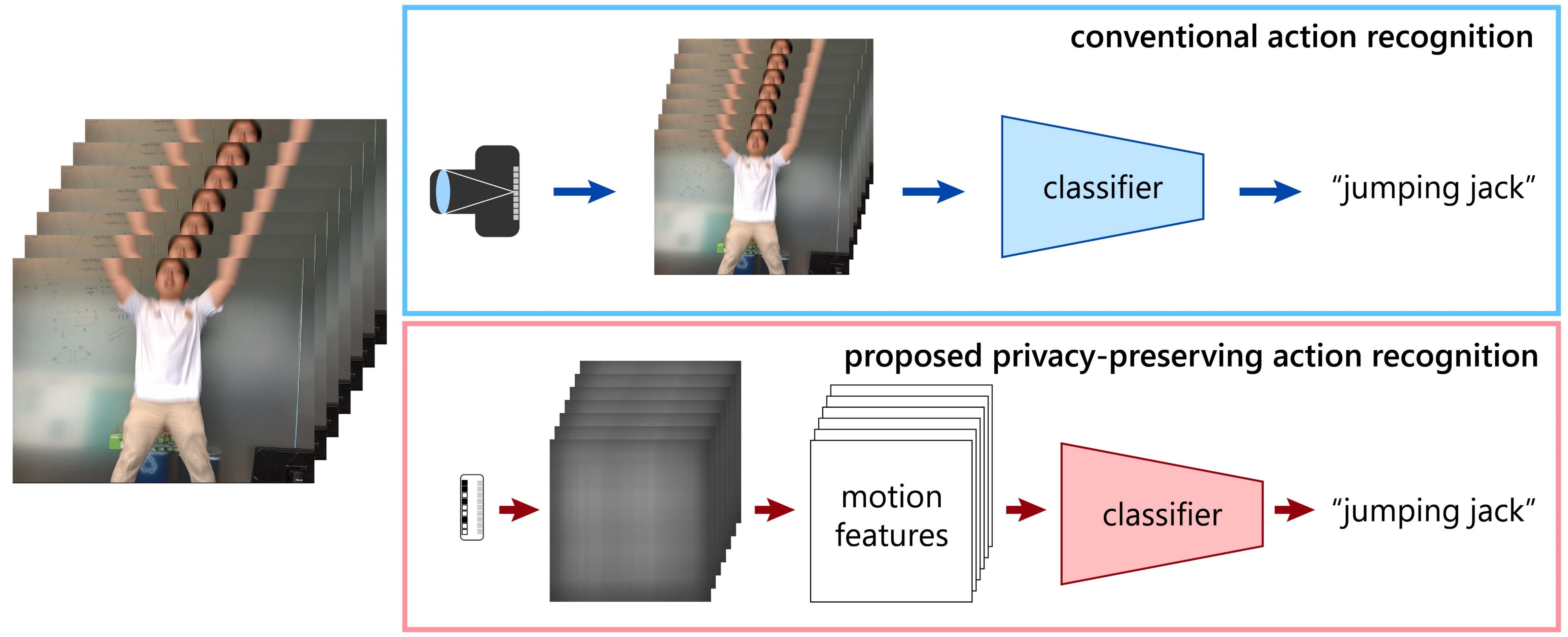
Motion features

Translation (phase correlation), Rotation & Scaling



$$o_2(\mathbf{p}) = o_1(\mathbf{sRp}')$$
$$\mathbf{p}' = \mathbf{p} + \Delta\mathbf{p}$$

$$C(\boldsymbol{v}) = \frac{O_1 \cdot O_2^*}{|O_1 \cdot O_2^*|} = \phi^* \frac{O_1 \cdot O_1^*}{|O_1 \cdot O_1^*|} = \phi(-\Delta\mathbf{p})$$



Goal: executing visual task(s) without looking at privacy-revealing data.

We propose:

1. **Pre-capture privacy**: lens-free coded aperture cameras.
2. **Post-capture privacy**: "mask-invariant" motion features.

\+ **T** features are invariant of mask patterns (A in Fourier space).

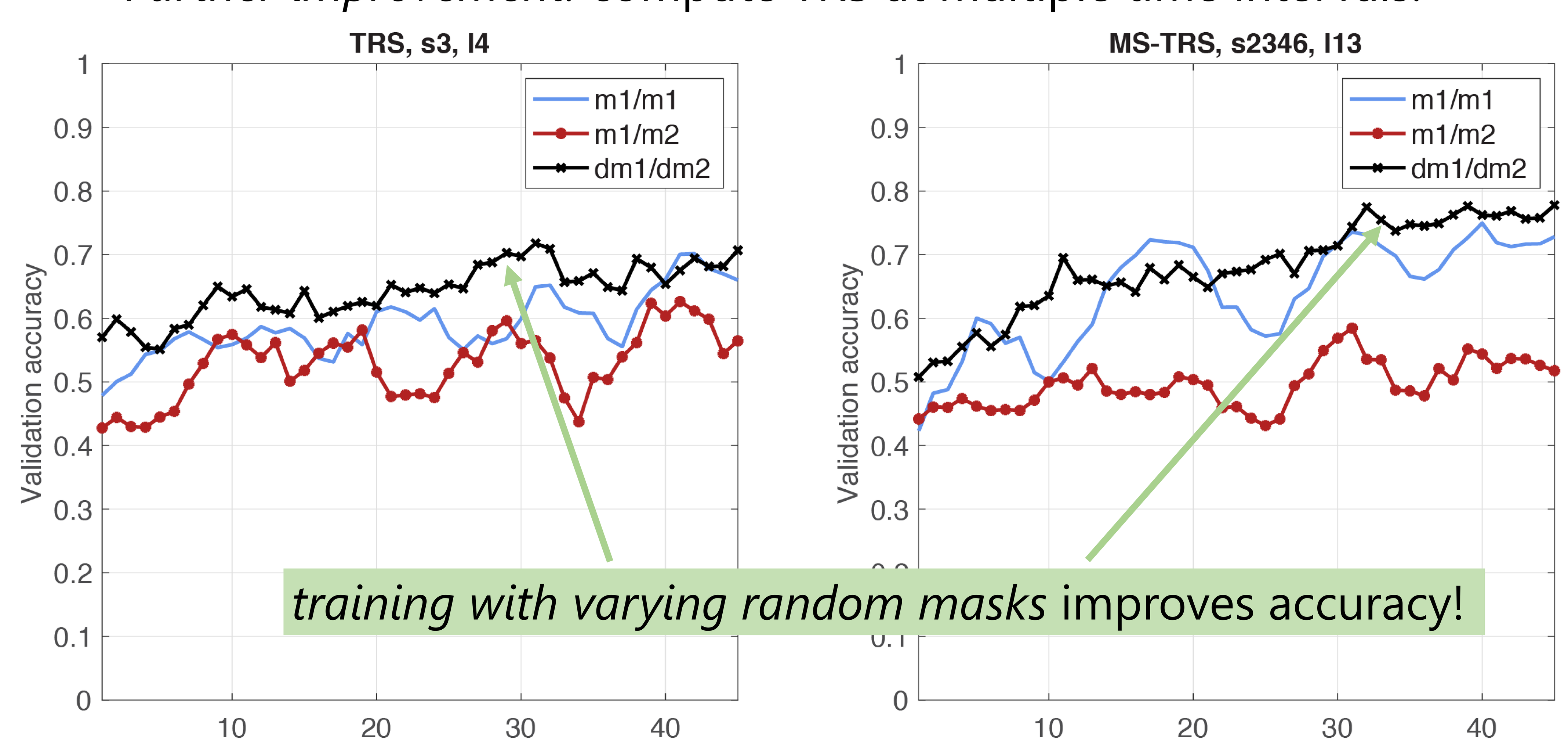
$$C_d(\mathbf{v}) = \frac{D_1 \cdot D_2^*}{|D_1 \cdot D_2^*|} = \phi^* \frac{O_1 \cdot A \cdot A^* \cdot O_1^*}{|O_1 \cdot A \cdot A^* \cdot O_1^*|} \approx \phi^*$$

Cross power spectrum of two CA images in Fourier space.

\- **RS** features do *not* share mask-invariant property.

\+ *Solution:* shuffle masks during training.

\+ *Further improvement:* compute TRS at multiple time intervals.



Results in simulation

predicted class (columns) / true class (rows)

	1	2	3	4	5	6	7	8	9	10
1	97.1	0.0	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	94.3	0.0	0.0	0.0	0.0	2.9	2.9	0.0	0.0
3	0.0	8.6	91.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	10.8	2.7	81.1	5.4	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	3.3	20.0	76.7	0.0	0.0	0.0	0.0	0.0
6	0.0	28.6	8.6	0.0	0.0	57.1	5.7	0.0	0.0	0.0
7	0.0	37.1	5.7	0.0	0.0	5.7	51.4	0.0	0.0	0.0
8	2.9	51.4	2.9	0.0	0.0	0.0	11.4	31.4	0.0	0.0
9	0.0	65.6	0.0	0.0	0.0	0.0	15.6	6.3	12.5	0.0
10	0.0	31.4	2.9	0.0	0.0	0.0	42.9	8.6	8.6	5.7

1. Hopping 2. Staggering 3. Jumping up 4. Jumping jack 5. Body weight squat 6. Standing up 7. Sitting down 8. Throw 9. Clapping 10. Handwaving

Salient motion > subtle motion

Benefits of mask-invariant property

Application: private/public surveillance

User: a generic classifier to only monitor/respond to actions.

Manufacturer: relaxed mask design, less calibration effort.

Hacker: more challenging to recover the scenes w/o mask info.

Reconstruction with PSF info?

Non-trivial and expensive