First-person perception and interaction

Kristen Grauman
University of Texas at Austin
Facebook AI Research
Visual recognition: significant progress
Visual recognition: significant progress

Big labeled datasets -> Deep learning -> GPU technology

dog  cat  car

Kristen Grauman, FAIR & UT Austin
Visual recognition: significant progress

Big labeled datasets → Deep learning → GPU technology

ImageNet top-5 error (%)

2010: 0.28
2011: 0.25
2012: 0.18
2013: 0.12
2014: 0.07
2015: 0.036
2016: 0.03
2017: 0.023

16.7% ↓ 23.3% ↓

dog cat car
The Web photo perceptual experience

- BSD (2001)
- PASCAL (2007-12)
- LabelMe (2007)
- ImageNet (2009)
- SUN (2010)
- Places (2014)
- MS COCO (2014)
- Visual Genome (2016)
The Web photo perceptual experience

A “disembodied” well-curated moment in time
First-person perception and learning

Status quo:
Learning and inference with “disembodied” photos.
First-person perception and learning

Status quo:
Learning and inference with “disembodied” photos.

On the horizon:
Visual learning in the context of motion, interaction, and multi-sensory observations.
First-person perception and learning

Status quo:
Learning and inference with “disembodied” photos.

On the horizon:
Visual learning in the context of motion, interaction, and multi-sensory observations.
This talk

Main idea:
Towards embodied perception
This talk

Main idea:
Towards embodied perception via agents that learn to anticipate their perceptual experience as a function of their own actions
This talk

Multi-sensory  Motion  Interaction

Towards embodied perception
This talk

Multi-sensory

Motion

Interaction

Audio-visual learning

Towards embodied perception
This talk

Multi-sensory

Motion

Interaction

Audio-visual learning

Navigation policies

Towards embodied perception
This talk

Multi-sensory

Motion

Interaction

Audio-visual learning

Navigation policies

Affordance learning

Towards embodied perception
Spatial effects in audio

source

head shadow (high freq)

path length difference

Kristen Grauman, FAIR & UT Austin

Image Credit: Michael Mandel
Spatial effects in audio

Cues for spatial hearing:
- Interaural time difference (ITD)
- Interaural level difference (ILD)
- Spectral detail (from pinna reflections)
Spatial effects in audio

Cues for spatial hearing:

- Interaural time difference (ITD)
- Interaural level difference (ILD)
- Spectral detail (from pinna reflections)
Our idea: 2.5D visual sound

Monaural

Binaural

Gao & Grauman, CVPR 2019
Our idea: 2.5D visual sound

Monaural → “Lift” → Binaural

Kristen Grauman, FAIR & UT Austin

Gao & Grauman, CVPR 2019
Our idea: 2.5D visual sound

“Lift” mono audio to spatial audio via visual cues

Gao & Grauman, CVPR 2019
Why infer binaural sound?
Why infer binaural sound?

Upgrade audio

Monaural Audio  ➔  Predicted Binaural Audio

Kristen Grauman, FAIR & UT Austin
Why infer binaural sound?

Upgrade audio

Monaural Audio ➔ Predicted Binaural Audio

Improve separation

sound of guitar

sound of saxophone
Our idea: 2.5D visual sound

“Lift” mono audio to spatial audio via visual cues
Our idea: 2.5D visual sound

“Lift” mono audio to spatial audio via visual cues

left channel

right channel

mono audio

spectrogram

visual frame = spatial cues

Gao & Grauman, CVPR 2019
Our idea: 2.5D visual sound

“Lift” mono audio to spatial audio via visual cues

left channel

mono audio

right channel

spectrogram

visual frame = spatial cues

Mono2Binaural

Gao & Grauman, CVPR 2019
Our idea: 2.5D visual sound

“Lift” mono audio to spatial audio via visual cues

left channel

right channel

mono audio

spectrogram

visual frame = spatial cues

Mono2Binaural

predicted left channel

predicted right channel

Kristen Grauman, FAIR & UT Austin

Gao & Grauman, CVPR 2019
Our idea: 2.5D visual sound

“Lift” mono audio to spatial audio via visual cues

mono audio → spectrogram → Mono2Binaural → predicted left channel → predicted right channel

visual frame = spatial cues
FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play
FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Data collection rig

GoPro

3Dio
Binaural Mic

Gao & Grauman, CVPR 2019
FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Binaural microphone rig linked to camera and monaural mic
FAIR-Play dataset

https://github.com/facebookresearch/FAIR-Play

Binaural microphone rig linked to camera and monaural mic

Capture ~5 hours video and binaural sound in a music room
Results: 2.5D visual sound

Listen with headphones!
Results: 2.5D visual sound

Input video

Left channel

Our method

Ground-truth

Mono

Right channel

vision.cs.utexas.edu/projects/2.5D_visual_sound/
Ask listener: where is the drum/piano?

Listener does not see any video
Datasets

FAIR-Play
- 10 musical instruments, e.g., cello, guitar, harp, trumpet, etc.
- ~5 hours of performances

YouTube Datasets
[Morgado et al. NeurIPS 2018]
- Streets, random YouTube
- ~1000 360° video clips
- Converted to binaural audio using decoder
## Results: Binaural audio prediction

<table>
<thead>
<tr>
<th></th>
<th>FAIR-Play</th>
<th></th>
<th>RE-C-Street</th>
<th></th>
<th>YT-Clean</th>
<th></th>
<th>YT-Music</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
<td>ENV</td>
</tr>
<tr>
<td>Ambisonics</td>
<td>-</td>
<td>-</td>
<td>0.744</td>
<td>0.126</td>
<td>1.435</td>
<td>0.155</td>
<td>1.885</td>
<td>0.183</td>
</tr>
<tr>
<td>Audio-Only</td>
<td>0.966</td>
<td>0.141</td>
<td>0.590</td>
<td>0.114</td>
<td>1.065</td>
<td>0.131</td>
<td>1.553</td>
<td>0.167</td>
</tr>
<tr>
<td>Flipped-Visual</td>
<td>1.145</td>
<td>0.149</td>
<td>0.658</td>
<td>0.123</td>
<td>1.095</td>
<td>0.132</td>
<td>1.590</td>
<td>0.165</td>
</tr>
<tr>
<td>Mono-Mono</td>
<td>1.155</td>
<td>0.153</td>
<td>0.774</td>
<td>0.136</td>
<td>1.369</td>
<td>0.153</td>
<td>1.853</td>
<td>0.184</td>
</tr>
<tr>
<td>Mono2Binaural (Ours)</td>
<td>0.836</td>
<td>0.132</td>
<td>0.565</td>
<td>0.109</td>
<td>1.027</td>
<td>0.130</td>
<td>1.451</td>
<td>0.156</td>
</tr>
</tbody>
</table>

Ambisonics: Morgado et al. NeurIPS 2018

Kristen Grauman, FAIR & UT Austin

Gao & Grauman, CVPR 2019
## Results: Binaural audio prediction

<table>
<thead>
<tr>
<th></th>
<th>FAIR-Play</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
</tr>
<tr>
<td>Ambisonics</td>
<td>-</td>
<td>-</td>
<td>0.744</td>
<td>0.126</td>
<td>1.435</td>
</tr>
<tr>
<td>Audio-Only</td>
<td>0.966</td>
<td>0.141</td>
<td>0.590</td>
<td>0.114</td>
<td>1.065</td>
</tr>
<tr>
<td>Flipped-Visual</td>
<td>1.145</td>
<td>0.149</td>
<td>0.658</td>
<td>0.123</td>
<td>1.095</td>
</tr>
<tr>
<td>Mono-Mono</td>
<td>1.155</td>
<td>0.153</td>
<td>0.774</td>
<td>0.136</td>
<td>1.369</td>
</tr>
<tr>
<td><strong>Mono2Binaural (Ours)</strong></td>
<td><strong>0.836</strong></td>
<td><strong>0.132</strong></td>
<td><strong>0.565</strong></td>
<td><strong>0.109</strong></td>
<td><strong>1.027</strong></td>
</tr>
</tbody>
</table>

Best binaural prediction results on all four datasets

Ambisonics: Morgado et al. NeurIPS 2018

Gao & Grauman, CVPR 2019
Results: Audio-visual source separation

original video
(before separation)

visual predictions:
dog & violin

2.5d visual sound $\rightarrow$ better audio separation

[Gao, Feris, & Grauman, ECCV 2018]
Results: Binaural audio prediction

<table>
<thead>
<tr>
<th></th>
<th>FAIR-Play</th>
<th></th>
<th>REC-STREET</th>
<th></th>
<th>YT-CLEAN</th>
<th></th>
<th>YT-MUSIC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
<td>ENV</td>
<td>STFT</td>
<td>ENV</td>
</tr>
<tr>
<td>Ambisonics</td>
<td>-</td>
<td>-</td>
<td>0.744</td>
<td>0.126</td>
<td>1.435</td>
<td>0.155</td>
<td>1.885</td>
<td>0.183</td>
</tr>
<tr>
<td>Audio-Only</td>
<td>0.966</td>
<td>0.141</td>
<td>0.590</td>
<td>0.114</td>
<td>1.065</td>
<td>0.131</td>
<td>1.553</td>
<td>0.167</td>
</tr>
<tr>
<td>Flipped-Visual</td>
<td>1.145</td>
<td>0.149</td>
<td>0.658</td>
<td>0.123</td>
<td>1.095</td>
<td>0.132</td>
<td>1.590</td>
<td>0.165</td>
</tr>
<tr>
<td>Mono-Mono</td>
<td>1.155</td>
<td>0.153</td>
<td>0.774</td>
<td>0.136</td>
<td>1.369</td>
<td>0.153</td>
<td>1.853</td>
<td>0.184</td>
</tr>
<tr>
<td>Mono2Binaural</td>
<td><strong>0.836</strong></td>
<td><strong>0.132</strong></td>
<td><strong>0.565</strong></td>
<td><strong>0.109</strong></td>
<td><strong>1.027</strong></td>
<td><strong>0.130</strong></td>
<td><strong>1.451</strong></td>
<td><strong>0.156</strong></td>
</tr>
</tbody>
</table>

Best binaural prediction results on all four datasets

Ambisonics: Morgado et al. NeurIPS 2018

Gao & Grauman, CVPR 2019
Datasets

**FAIR-Play**
- 10 musical instruments, e.g., cello, guitar, harp, trumpet, etc.
- ~5 hours of performances

**YouTube Datasets**
[Morgado et al. NeurIPS 2018]
- Streets, random YouTube
- ~1000 360° video clips
- Converted to binaural audio using decoder

Gao & Grauman, CVPR 2019
Results: Audio-visual source separation

original video (before separation)

visual predictions: dog & violin

2.5d visual sound → better audio separation

[Gao, Feris, & Grauman, ECCV 2018]
This talk

Multi-sensory

Audio-visual learning

Navigation policies

Motion

Interaction

Affordance learning

Towards embodied perception
Visual navigation in novel unmapped environments

Where is the telephone?

Kristen Grauman, FAIR & UT Austin
Our idea: Audio-visual navigation
Our idea: Audio-visual navigation

Sound informs navigating agent about...

Target location
Our idea: Audio-visual navigation

Sound informs navigating agent about...

Target location

Safety
Our idea: Audio-visual navigation

Sound informs navigating agent about...

- Target location
- Semantics
- Safety
Our idea: Audio-visual navigation

Sound informs navigating agent about...

- Target location
- Semantics
- Safety
- Materials
Our idea: Audio-visual navigation

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio simulation platform

We introduce audio simulation platform

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
We introduce audio simulation platform

- Visually realistic 3D environments (Facebook Replica scenes)

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio simulation platform

We introduce audio simulation platform

- Visually realistic 3D environments (Facebook Replica scenes)
- Room impulse response (RIR) for all source x receiver locs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio simulation platform

We introduce audio simulation platform

- Visually realistic 3D environments (Facebook Replica scenes)
- Room impulse response (RIR) for all source x receiver locs
- Convolve with arbitrary waveform to render binaural sound heard by agent

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio simulation platform

We introduce audio simulation platform

- Visually realistic 3D environments (Facebook Replica scenes)
- Room impulse response (RIR) for all source x receiver locs
- Convolve with arbitrary waveform to render binaural sound heard by agent

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio-visual navigation task

Navigate to an audio-emitting goal (e.g., phone ringing)
Audio-visual navigation model
Reinforcement learning for agent’s motion policy from multi-modal inputs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio-visual navigation model
Reinforcement learning for agent’s motion policy from multi-modal inputs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio-visual navigation model

Reinforcement learning for agent’s motion policy from multi-modal inputs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio-visual navigation model
Reinforcement learning for agent’s motion policy from multi-modal inputs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Audio-visual navigation model

Reinforcement learning for agent’s motion policy from multi-modal inputs

[Chen et al., Audio-visual embodied navigation, arXiv 2019]
Does audio help navigation?

2D t-SNE projection of audio features learned by our agent
Does audio help navigation?
Does audio help navigation?

<table>
<thead>
<tr>
<th></th>
<th>PointGoal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind</td>
<td>0.451</td>
</tr>
<tr>
<td>RGB</td>
<td>0.465</td>
</tr>
<tr>
<td>Depth</td>
<td>0.592</td>
</tr>
</tbody>
</table>

success rate normalized by path length (SPL)
Does audio help navigation?

<table>
<thead>
<tr>
<th></th>
<th>PointGoal</th>
<th>AudioPointGoal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind</td>
<td>0.451</td>
<td>0.647</td>
</tr>
<tr>
<td>RGB</td>
<td>0.465</td>
<td>0.735</td>
</tr>
<tr>
<td>Depth</td>
<td>0.592</td>
<td>0.749</td>
</tr>
</tbody>
</table>

success rate normalized by path length (SPL)
Can audio supplant GPS?

![Graph showing navigation success rate vs. sigma (m). The graph compares four different audio solutions against a baseline of perfect GPS and a noisier GPS.](image-url)
Can audio supplant GPS?

![Graph showing navigation success rate vs. Sigma (m) for different audio and GPS conditions.]

- **Navigation success rate**
- **Sigma (m)**
- **PointGoal-Depth**
- **AudioPointGoal-Depth**
- **AudioGoal-Depth**

**Perfect GPS** ————————> **Noisier GPS**

No GPS!
This talk

Multi-sensory

Motion

Interaction

Audio-visual learning

Navigation policies

Affordance learning

Towards embodied perception
From *naming* objects to *using* them

Embodied perception system

Object manipulation

- Turn on
- Increase height
- Move lamp
- Replace lightbulb

Kristen Grauman, FAIR & UT Austin
From *naming* objects to *using* them

Affordances

- Toggle-able
- Adjustable
- Replaceable
- Movable

Embodied perception system

Object manipulation

Kristen Grauman, FAIR & UT Austin
Current approaches: affordance as semantic segmentation

Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …
Current approaches: affordance as semantic segmentation

Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …
Current approaches: affordance as semantic segmentation

Captures annotators’ expectations of what is important

Sawatzky et al. (CVPR 17), Nguyen et al. (IROS 17), Roy et al. (ECCV 16), Myers et al. (ICRA 15), …
Learning affordances from video

[Nagarajan et al. ICCV 2019]
Learning affordances from video

LSTM

Aggregated state for the action

Action Classifier

“open” $\mathcal{L}_{cls}$

[Nagarajan et al. ICCV 2019]
Extracting interaction hotspot maps

Activation mapping to identify responsible spatial regions

Hypothesize for action $a = \text{“pullable”}$

[Nagarajan et al. ICCV 2019]
Extracting interaction hotspot maps

Activation mapping to identify responsible spatial regions

$H_0 = \text{ReLU} \left( \sum_k a_k x^k \right)$

Hypothesize for action $a = \text{“pullable”}$

[Nagarajan et al. ICCV 2019]
Extracting interaction hotspot maps

Activation mapping to identify responsible spatial regions

[Nagarajan et al. ICCV 2019]
Evaluating interaction hotspots

OPRA
(Fang et al., CVPR 18)

EPIC Kitchens
(Damen et al., ECCV 18)
Evaluating interaction hotspots

OPRA
(Fang et al., CVPR 18)

EPIC Kitchens
(Damen et al., ECCV 18)

Train on video datasets, generate heatmaps on novel images---even from unseen categories
Results: interaction hotspots

Given static image of object at rest, infer affordance regions

[Nagarajan et al. ICCV 2019]
Results: interaction hotspots

Given static image of object at rest, infer affordance regions

<table>
<thead>
<tr>
<th>Method</th>
<th>OPRA data</th>
<th></th>
<th>EPIC data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KLD ↓</td>
<td>SIM ↑</td>
<td>AUC-J ↑</td>
<td>KLD ↓</td>
</tr>
<tr>
<td>CENTER BIAS</td>
<td>11.132</td>
<td>0.205</td>
<td>0.625</td>
<td>10.660</td>
</tr>
<tr>
<td>LSTM+GRAD-CAM</td>
<td>8.573</td>
<td>0.209</td>
<td>0.620</td>
<td>6.470</td>
</tr>
<tr>
<td>EGOGAZE [27]</td>
<td>2.428</td>
<td>0.245</td>
<td>0.646</td>
<td>2.241</td>
</tr>
<tr>
<td>MLNET [6]</td>
<td>4.022</td>
<td>0.284</td>
<td>0.763</td>
<td>6.116</td>
</tr>
<tr>
<td>DEEPIGAZE [33]</td>
<td>1.897</td>
<td>0.296</td>
<td>0.720</td>
<td>1.352</td>
</tr>
<tr>
<td>SALGAN [40]</td>
<td>2.116</td>
<td>0.309</td>
<td>0.769</td>
<td>1.508</td>
</tr>
<tr>
<td>OURS</td>
<td>1.427</td>
<td>0.362</td>
<td>0.806</td>
<td>1.258</td>
</tr>
<tr>
<td>IMG2HEATMAP</td>
<td>1.473</td>
<td>0.355</td>
<td>0.821</td>
<td>1.400</td>
</tr>
<tr>
<td>DEMO2VEC [12]</td>
<td>1.197</td>
<td>0.482</td>
<td>0.847</td>
<td>–</td>
</tr>
</tbody>
</table>

Nagarajan et al. ICCV 2019
Results: interaction hotspots

Given static image of object at rest, infer affordance regions

<table>
<thead>
<tr>
<th>Model</th>
<th>OPRA data</th>
<th>EPIC data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KLD ↓</td>
<td>SIM ↑</td>
</tr>
<tr>
<td>CENTER BIAS</td>
<td>11.132</td>
<td>0.205</td>
</tr>
<tr>
<td>LSTM+GRAD-CAM</td>
<td>8.573</td>
<td>0.209</td>
</tr>
<tr>
<td>EGOGAZE [27]</td>
<td>2.428</td>
<td>0.245</td>
</tr>
<tr>
<td>MLNET [6]</td>
<td>4.022</td>
<td>0.284</td>
</tr>
<tr>
<td>DEEPGAZEII [33]</td>
<td>1.897</td>
<td>0.296</td>
</tr>
<tr>
<td>SALGAN [40]</td>
<td>2.116</td>
<td>0.309</td>
</tr>
<tr>
<td>OURS</td>
<td><strong>1.427</strong></td>
<td><strong>0.362</strong></td>
</tr>
<tr>
<td>IMG2HEATMAP</td>
<td>1.473</td>
<td>0.355</td>
</tr>
<tr>
<td>DEMO2VEC [12]</td>
<td>1.197</td>
<td>0.482</td>
</tr>
</tbody>
</table>

Weakly Supervised

[Nagarajan et al. ICCV 2019]
Interaction hotspots for object recognition
Interaction hotspots for object recognition

ResNet-50 predictions on COCO objects

refrigerator - 0.997  dishwasher - 0.001
refrigerator - 0.454  ATM machine - 0.210
mailbox - 0.404  refrigerator - 0.139
bookstore - 0.747  refrigerator - 0.009
switchbox - 0.511  refrigerator - 0.005
## Interaction hotspots for object recognition

**COCO**

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>25</th>
<th>100</th>
<th>3300 (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VANILLA</td>
<td>44.3 ± 0.3</td>
<td>56.6 ± 0.2</td>
<td>65.6 ± 0.4</td>
<td>75.2 ± 0.1</td>
</tr>
<tr>
<td>AUTOENCODER</td>
<td>39.4 ± 0.4</td>
<td>51.2 ± 0.2</td>
<td>59.1 ± 0.2</td>
<td>72.8 ± 0.3</td>
</tr>
<tr>
<td>OURS</td>
<td><strong>46.8 ± 0.3</strong></td>
<td><strong>57.9 ± 0.1</strong></td>
<td><strong>63.2 ± 0.2</strong></td>
<td><strong>73.9 ± 0.3</strong></td>
</tr>
</tbody>
</table>

**Legend:**
- **Holdable**
- **Openable**
Interaction hotspots for object recognition

Better low-shot object recognition by anticipating object function
Interaction hotspots for robot grasping

**Without watching people**

<table>
<thead>
<tr>
<th></th>
<th>Total Reward</th>
<th>Height Reward</th>
<th>Position Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body Centre Distance</td>
<td>0.36</td>
<td>0.36</td>
<td>-0.34</td>
</tr>
<tr>
<td>Max Contact Distance</td>
<td>0.400</td>
<td>-0.05</td>
<td>-0.36</td>
</tr>
<tr>
<td>Multi Contact Distance</td>
<td>0.34</td>
<td>0.32</td>
<td></td>
</tr>
</tbody>
</table>

```
-0.34
-0.36
 0 50 100 150 200
```

```
0.36
0.425
 0 50 100 150 200
```

```
0.34
0.400
 0 50 100 150 200
```

```
0.32
```

---

[Images of robot hand on surface]
Interaction hotspots for robot grasping

Without watching people

Learn grasping policy for 24 DoF dexterous hand that rewards closeness to hotspots
Summary

Towards first-person perception

- self-supervised learning via anticipation
- learning to autonomously direct the camera
- multi-sensory observations (audio, motion, visual)
- object interaction from video