Learning egocentric policies for where to look

Kristen Grauman
Department of Computer Science
University of Texas at Austin
Human-taken photos

A well-framed, well-curated moment in time

BSD (2001)


PASCAL (2007-12)

LabelMe (2007)

ImageNet (2009)

SUN (2010)

Places (2014)

MS COCO (2014)

Visual Genome (2016)
Passively-captured video

A tangle of relevant and irrelevant information

First-person video

360 video
This talk

Egocentric policies for where to look

1. **Where to look** for object/scene recognition? Intelligent view selection and manipulations

2. **Where to look** when dynamically exploring? Learning to look around for active exploration

3. **Where to look** in a wide field of view video? Automatic cinematography in 360 video
Actively moving to recognize


Jayaraman and Grauman, ECCV 2016
Actively moving to recognize

Our idea: End-to-end active recognition + 3D motion look-ahead

Jayaraman and Grauman, ECCV 2016
End-to-end active recognition: tasks

1. Look around scene
2. Manipulate object
3. Move around object
End-to-end active recognition: results

SUN 360

GERMS

ModelNet-10

Faster recognition via intelligent view selection

Jayaraman and Grauman, ECCV 2016
End-to-end active recognition: example

[Jayaraman and Grauman, ECCV 2016]
End-to-end active recognition: example

Predicted label:

GERMS dataset: Malmir et al. BMVC 2015

[Jayaraman and Grauman, ECCV 2016]
Next-active-object prediction

What object will the camera wearer interact with next?

Next-active-object prediction

Approach: learn properties of active object trajectories

Next-active-object prediction

Egomotion and implied body pose

Learn relationship between egocentric scene motion and 3D human body pose

Input: egocentric video

Output: sequence of 3d joint positions

[Jiang & Grauman, CVPR 2017]
Egomotion and implied body pose

Learn relationship between egocentric scene motion and 3D human body pose

Input: egocentric video

Wearable camera video
Inferred pose of camera wearer

[Jiang & Grauman, CVPR 2017]
This talk

Egocentric policies for where to look

1. **Where to look** for object/scene recognition?
   Intelligent view selection and manipulations

2. **Where to look** when dynamically exploring?
   Learning to look around for active exploration

3. **Where to look** in a wide field of view video?
   Automatic cinematography in 360 video
Goal: Learn to “look around”

Can we learn **look-around policies** for visual agents that are curiosity-driven, exploratory, and generic?
Key idea: Active observation completion

Completion objective: Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment

Agent must choose where to look before looking there.

Jayaraman and Grauman, arXiv 2017
Key idea: Active observation completion

Completion objective: Learn policy for efficiently inferring (pixels of) all yet-unseen portions of environment

Agent must choose where to look before looking there.

Jayaraman and Grauman, arXiv 2017
Approach: Active observation completion

Non-myopic: Train to target a budget of observation time

Jayaraman and Grauman, arXiv 2017
Datasets: Two scenarios

Where to look next?

SUN 360 panoramas
[Xiao 2012]

How to manipulate?
Active “look around” results

*Harel et al, Graph based Visual Saliency, NIPS’07

Jayaraman and Grauman, arXiv 2017
Active “look around” results

Learned active look-around policy: quickly grasp environment independent of a specific task

*Harel et al, Graph based Visual Saliency, NIPS’07

Jayaraman and Grauman, arXiv 2017
Active “look around” visualization

![Observed view](image1)

Ground truth

Visualized internal model over time

---

$t=1$  $t=2$  $t=3$

Jayaraman and Grauman, arXiv 2017
Active “look around” visualization

Ground truth  Visualized internal model over time

Jayaraman and Grauman, arXiv 2017
Active “look around” visualization

Ground truth

observed view

Visualized internal model over time

Jayaraman and Grauman, arXiv 2017
Unsupervised observation completion

Encoder

Decoder

Supervised recognition

[Jayaraman et al, ECCV 16]

“beach”

Plug observation completion policy in for new task
Motion policy transfer

Unsupervised exploratory policy approaches supervised task-specific policy accuracy!
This talk

Egocentric policies for where to look

1. **Where to look** for object/scene recognition?
   Intelligent view selection and manipulations

2. **Where to look** when dynamically exploring?
   Learning to look around for active exploration

3. **Where to look** in a wide field of view video?
   Automatic cinematography in 360 video
Challenge of viewing 360° videos

Control by mouse

How to find the right direction to watch?
Proposed problem: Pano2Vid automatic videography

Definition

Input: 360° video
Output: “natural-looking” normal FOV video
Task: control virtual camera direction and FOV

[Su et al. ACCV 2016, CVPR 2017]
Our approach – AutoCam

Learn videography tendencies from unlabeled Web videos

- Diverse capture-worthy content
- Proper composition

Human-captured NFOV videos ("HumanCam")

Unlabeled video

[Su et al. ACCV 2016, CVPR 2017]
Example spatio-temporal glimpses

High capture-worthiness

Low capture-worthiness

First frame of glimpses scored high/low by our approach
Construct virtual camera trajectory

Densely sample and score glimpses

Pose selection as shortest path(s) problem

Optimize for *multiple diverse* hypotheses

Output smooth view path maximizing capture-worthiness
360 Pano2Vid Dataset

http://vision.cs.utexas.edu/projects/watchable360

• All videos crawled from YouTube using keywords:
  “Hiking”, “Mountain climbing”, “Parade”, “Soccer”

<table>
<thead>
<tr>
<th></th>
<th># videos</th>
<th>Total length</th>
</tr>
</thead>
<tbody>
<tr>
<td>360° videos</td>
<td>86</td>
<td>7.3 hours</td>
</tr>
<tr>
<td>HumanCam</td>
<td>9,171</td>
<td>343 hours</td>
</tr>
</tbody>
</table>

• **For evaluation:** 480 trajectories / 12 hours of human edited video
AutoCam results

http://vision.cs.utexas.edu/projects/watchable360/

Automatically select FOV and viewing direction

[Su & Grauman, CVPR 2017]
AutoCam results

http://vision.cs.utexas.edu/projects/watchable360/

Automatically select FOV and viewing direction

[Su & Grauman, CVPR 2017]
AutoCam results:
Multiple diverse hypotheses
http://vision.cs.utexas.edu/projects/watchable360/

Input Video & Cam. Trajectory

Output Videos

Hypothesis 1

Hypothesis 2
Results: Quantitative evaluation

Similarity to user-uploaded standard web videos

Similarity to human-selected camera trajectories

Create plausible videos by learning “where to look” from unlabeled video

[Su et al. ACCV 2016, CVPR 2017]
Summary

- From curated images to egocentric video: challenges in knowing where to look next.
  - End-to-end active recognition
  - Next-active-object prediction
  - First person body pose estimation
  - Learning generic “look around” behavior
  - Automatic cinematography for 360 video

Dinesh Jayaraman  Yu-Chuan Su  Hao Jiang  Antonino Furnari  Giovanni Maria Farinella
Papers


- **Learning to look around.** [Dinesh Jayaraman](http://example.com), Kristen Grauman, arXiv Sept 2017.


