First-person Vision

Topic Presentation: Yousi Lin
First person “egocentric” vision:

First Person Vision (FPV) is a transformative system that can monitor, record and assist people in their daily lives at work or at play in a truly symbiotic manner.

- Linked to ongoing experience of the camera wearer
- World seen in context of the camera wearer’s activity and goals
Some of the more important works and commercial announcements in FPV.

SenseCam: released by Microsoft Research in 2006.

GoPro Hero: first one released in 2010.

Google Glasses: released by Google in 2012.
New era for first-person vision

Augmented reality

Health monitoring

Law enforcement

Science

Robotics

Life logging

Kristen Grauman, UT Austin
Guillermo | Tijuana
What can a first person camera tell us about the wearer?

- Personal/social attention
- Human kinematics (object/pose/action)
- Visual sensorimotor behaviors

slide credits: CVPR 2016 Tutorial: First Person Vision
Learning to Predict Gaze in Egocentric Video

Yin Li, Alireza Fathi, James M. Rehg
School of Interactive Computing, Georgia Tech
Proceedings of the 2013 IEEE International Conference on Computer Vision

Goal: Understanding first person’s behavior using gaze

Why: Gaze is a very important signal

How: Predicting the camera wearer’s gaze using egocentric cues
Method

- Modeling the first person’s head-eye / hand-eye coordination
- Only use egocentric cues, e.g. hand pose, head movement
- A temporal dynamic model for fixations
Egocentric Cues

Eye, Head and Hand Coordinations

- Center Prior (Head Orientation)
- Head Motion
- Hand Location

They did not use low-level image features or high level task information
1. Head-Eye Coordination

Center Prior: Head Orientation

Monitor Based Tracking | Egocentric Gaze Tracking

MIT | GTEA Gaze | GTEA Gaze+
1. Head-Eye Coordination

**Head Motion**

Horizontal Direction

- Large head motion is always accompanied by a large gaze shift
- Linear correlation of head motion and gaze shift in horizontal direction
2. Eye-Hand Coordination

Manipulation Point: a control point where the person is most likely to manipulate an object
2. Eye-Hand Coordination

Peak of gaze distributions around hands, where manipulations are most likely to happen
Temporal Models

Temporal Modeling

Saccade or Fixation

Head/Hand Motion

Gaze

Features

Hand Location/Pose

Random Forests

1. \( g_t \): gaze point at frame \( t \).
2. \( m_t \): \( m_t = \{0, 1\} \), where \( m_t = 1 \) denotes \( g_t \) is a fixation.
3. \( z_t \): feature vector for frame \( t \), which contains the manipulation point (2D), the global motion vector (2D), the hand motion vector (2D), the hand configuration (1D categorical).

notation:
The model: \[ P(\{g_t, m_t\}_{t=1}^K | \{z_t\}_{t=1}^K) = \prod_{t=1}^K P(g_t | z_t) \prod_{t=1}^K P(m_t | g_{N(t)}) \]

**Single Frame Gaze Prediction:**

\[ P(g_t | z_t) \propto \exp \left( - \|g_t - \tilde{g}_t\|^2_{\Sigma_s} \right) \]

**Fixations and Gazes:**

\[ P(m_t | g_{N(t)}) \propto \exp \left( -m_t \sum_{i \in N(t)} \|g_i - g_t\|^2_2 \right) \]

\[ m_t = \prod_{i \in N(t)} \frac{-\text{sign}(\|g_i - g_t\|^2_2 - c) + 1}{2} \]

**Inference:**

To get the gaze points and fixations, they applied Maximum Likelihood (ML) estimation of the first equation.

**Learning:**

1. train the single frame random regression tree
2. select the velocity threshold \( c \), the covariance matrix \( \Sigma_s \) and the constant \( \lambda \)

**notation:**

1. \( g_{N(t)} \): the temporal neighbors of \( g_t \).
2. \( \tilde{g}_t \): \( \tilde{g}_t = f(z_t) \)
3. \( \Sigma_s \): covariance matrix
4. \( c \): velocity threshold
Temporal Models

Fixations and Gazes

Temporal Modeling

Single Frame Gaze Prediction

Random Forests
GTEA Gaze Dataset

- 17 subjects
- Free choice meal preparation activities
- 42 objects

The first dataset of its kind
GTEA Gaze+ Dataset

- 6 subjects
- 7 activities (Making pizza, hamburger, breakfast, greek salad, etc.)
- Each activity takes around 10 min, around 100 action in each activity
Results: Gaze Prediction

GTEA Gaze Dataset

Comparison of AUC Scores

Comparison of AAE

ROC Curve (Frame)
Results: Gaze Prediction

Comparison of AUC Scores

Comparison of AAE

ROC Curve (Frame)
Application: Action Recognition

- Action recognition of 25 classes using predicted gazes 29% -> 32.8%
- Action recognition using human gaze s -> 49%
Application: Object Segmentation

Object Segmentation

Mean Best Overlap

Number of Segment Candidates

CPMC
CPMC with gaze prediction
CPMC with human gaze
ActSeg with gaze
ActSeg with gaze prediction
Conclusions

• A small circle of pixels around the point of gaze is sufficient to recognize daily actions in egocentric vision

• They treat gaze as a latent variable and showed that they could predict it when it cannot be measured

• Gaze prediction based on user’s head movement and hand location is surprisingly effective
Thank you!