G. Cheung
National Institute of Informatics
8th March, 2012

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Eye-gaze Prediction via Joint Analysis of Gaze Patterns and Visual Media

.

Outline

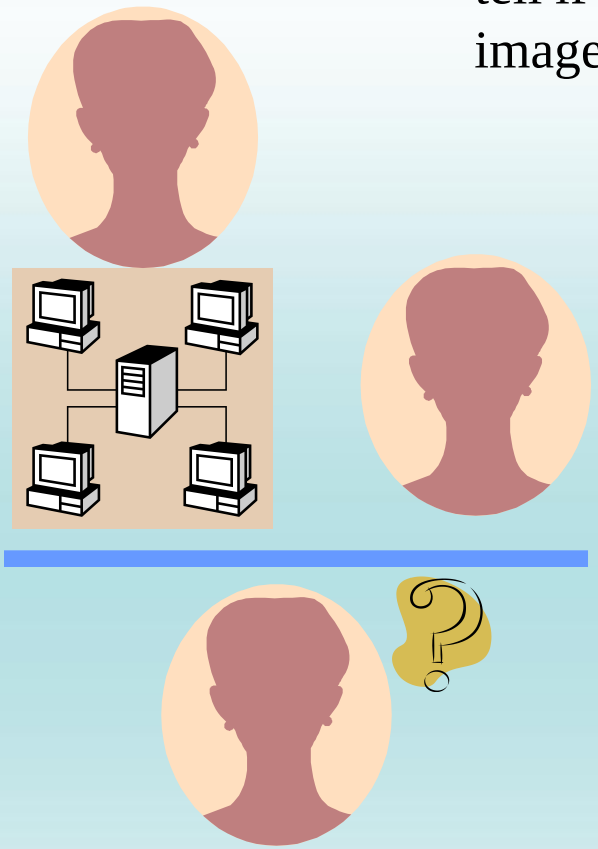
- Overview / update of my research
- Eye-gaze prediction for network video streaming
- Ditto for store-&-playback video
- Visual attention deviation (VAD)
- Conclusion

Outline

- Overview / update of my research
- Eye-gaze prediction for network video streaming
- Ditto for store-&-playback video
- Visual attention deviation (VAD)
- Conclusion

Immersive Visual Communication (IVIC) Test

- **Q**: can a person engage in visual communication, and not be able to tell if participant is actual human (across glass barrier) or rendered images (on digital display)?



Large HQ display w/ life-size images, comparable ambience

Gaze-corrected view

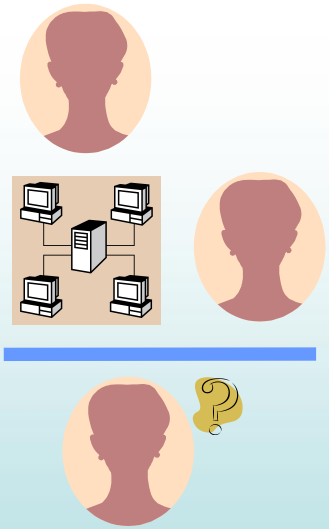
Motion Parallax: **Fast, smooth** interactive view-switching

1. Multiview video coding & View Synthesis

2. Loss/delay tolerant multiview video transport

3. Human-centric visual media interaction

Potential Impact



- Immersive Communication ≠ Skype calls!
  - Non-verbal means (postures, gestures) are important.
  - Eye-contact is important.
- Substitute for face-to-face meetings.
  - Reduce travel cost, improve productivity.
  - Reduce carbon footprints.
  - Example apps: HQ teleconferencing, tele-medicine.
- Enhance Virtual Reality is 1 of 14 grand challenges chosen by ***National Academy of Engineering*** for 21st c.
  - Treatment of social anxieties or phobias.
  - Training & teaching: virtual surgeries, etc.

Microsoft Research

Disney Research SCIENCE AT PLAY

NATIONAL ACADEMY OF ENGINEERING OF THE NATIONAL ACADEMIES

Overview

Multi-camera capture

texture & depth map capture & denoising

Loss-resilient source coding & network coding

Eye-gaze & head movement prediction

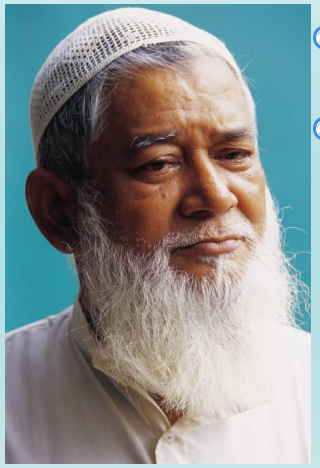
Multi-camera capture

Gaze-corrected

Motion parallax via head tracking

Lossy packet network

1



Multiview video coding & View Synthesis

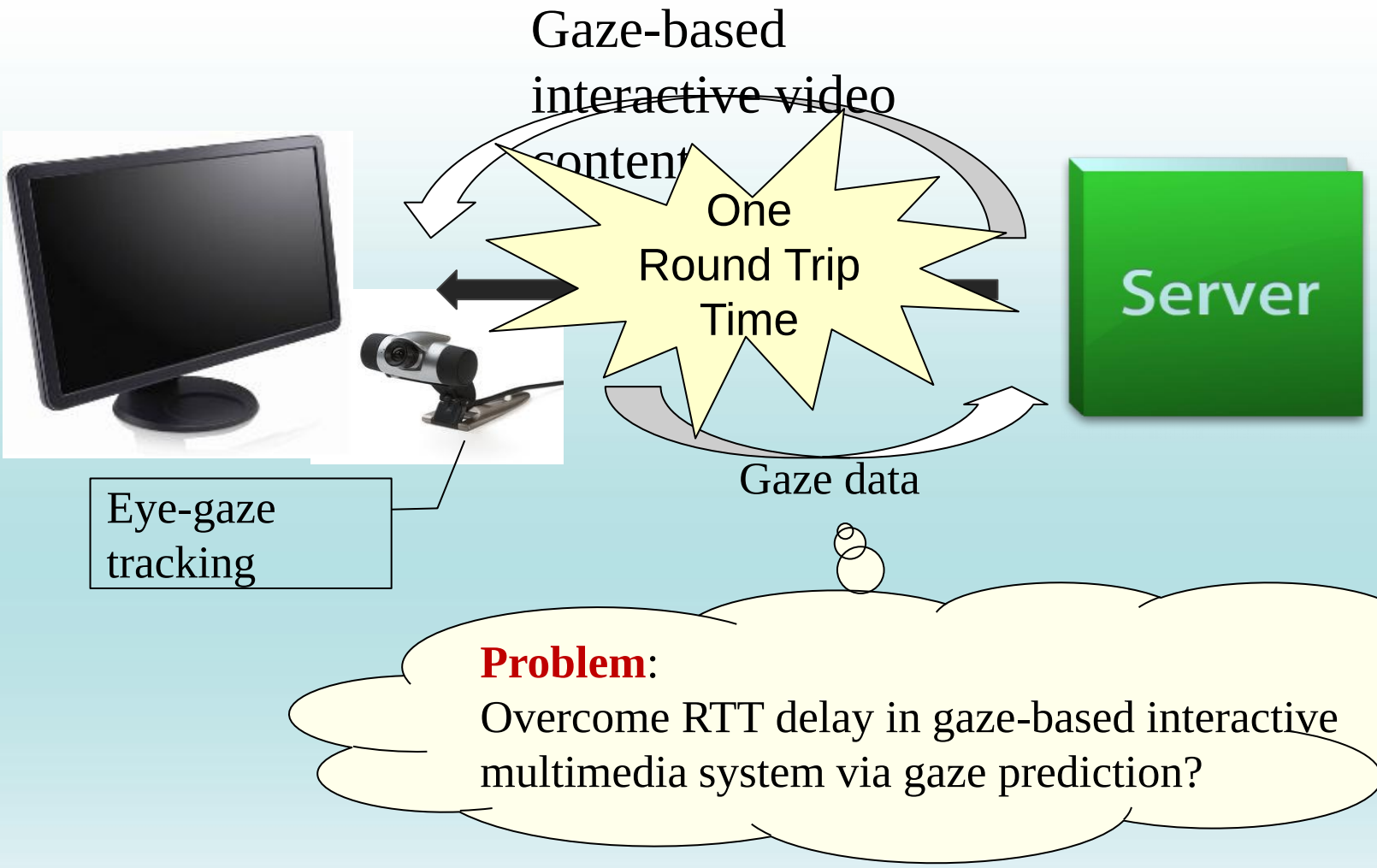
Loss/delay tolerant multiview transmission

Human-centric visual media interaction

Outline

- Overview / update of my research
- Eye-gaze prediction for network video streaming
- Ditto for store-&-playback video
- Visual attention deviation (VAD)
- Conclusion

Motivation 1: Reaction time in interactive video



Motivation 2: Observe (& learn) the observer

- Two processes happening concurrently:
  1. User observing video (physical signals like eye gaze, head movement, pupil size, etc).
  2. Video being playback (object motion, visual saliency maps, etc).
- Deduce higher level semantic info (interest, motive) by correlating the two.

Application:

- Automatic video / photo selection in large media content archive.
- Effective and un-intrusive ad insertion.

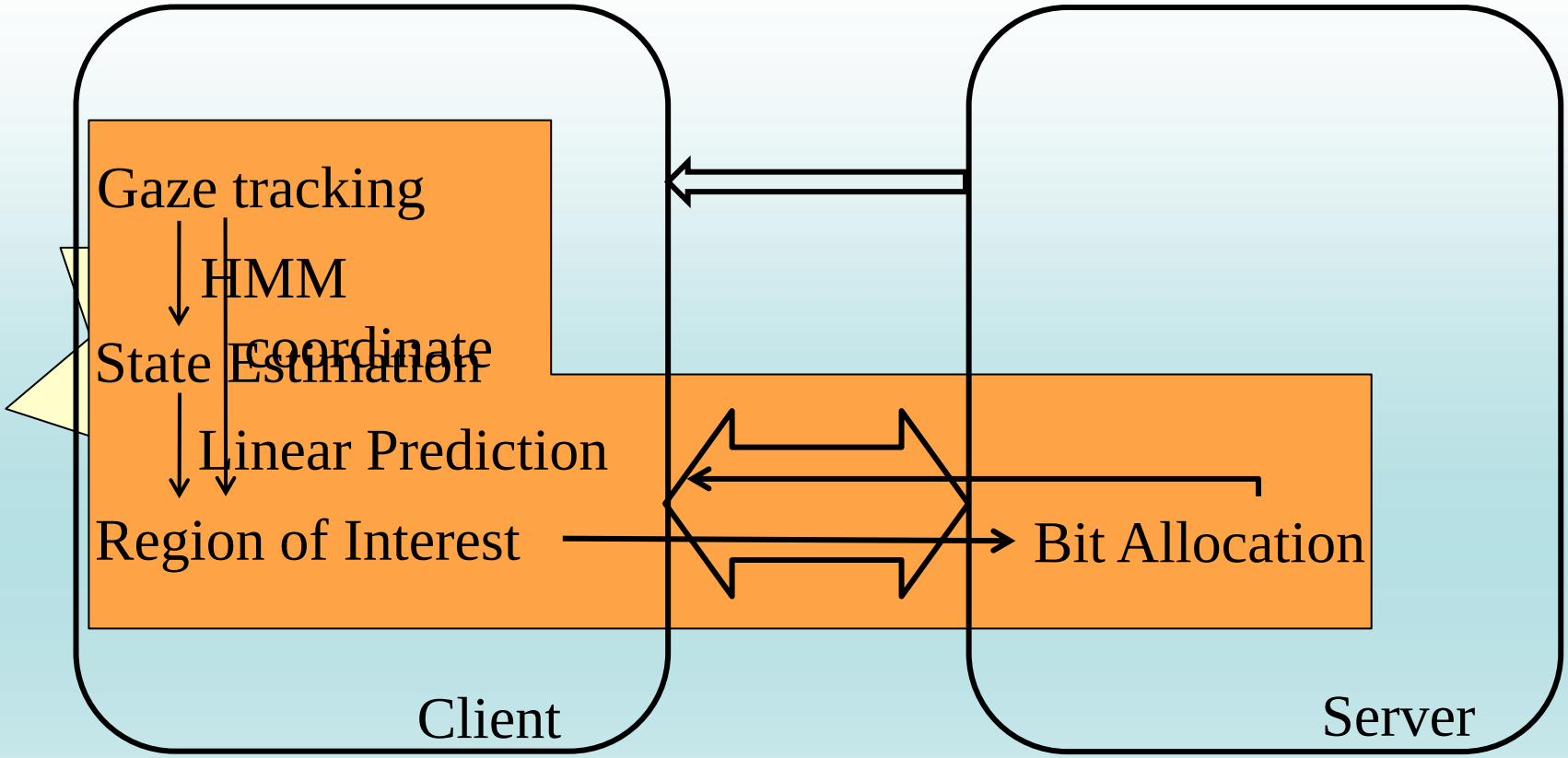
Related Work

- Given ROI, how to allocate video bits to minimize streaming rate [1]
  - **Con**: Assumed ROI not always good.
  - **Our approach**: real-time gaze data & saliency maps of video to infer future ROI.
- Eye-gazed prediction based on mechanics of human eye [2]
  - **Con**: ***content-independent***, but complicated model w/ many parameters.
  - **Our approach**: ***content-dependent***, but simple model w/ few parameters.

[1] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," in *IEEE Trans on CSVT, January 2008, vol. 18, no.1, pp. 134*–139.

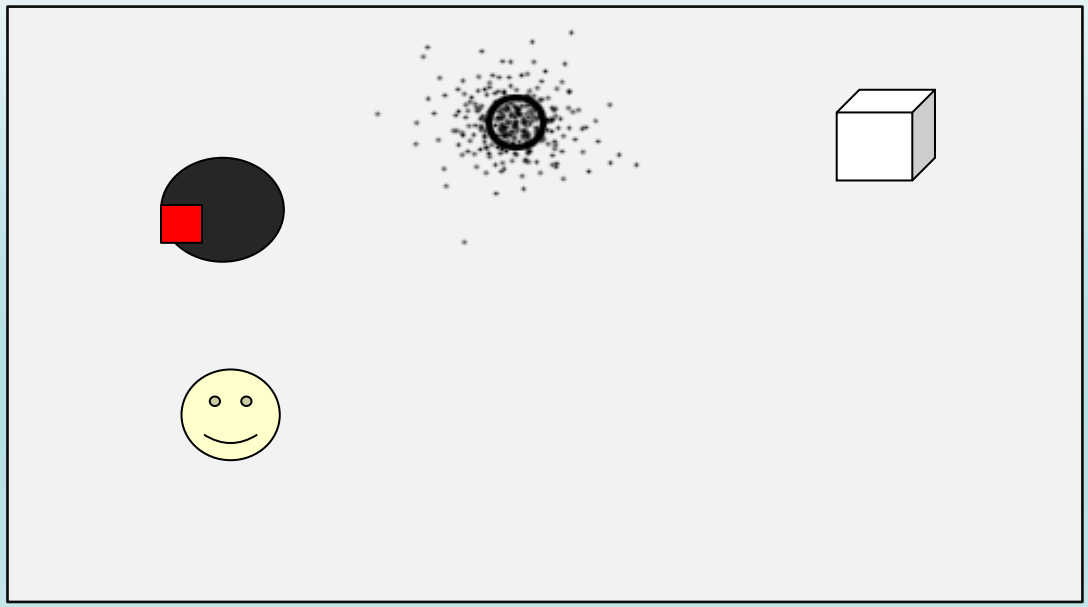
[2] O. Komogortsev, J. Khan, "Eye movement prediction by oculomotor plant Kalman filter with brainstem control," *Journal of Control Theory and Applications*, Jan. 2009, vol.7, no.1.

System Overview

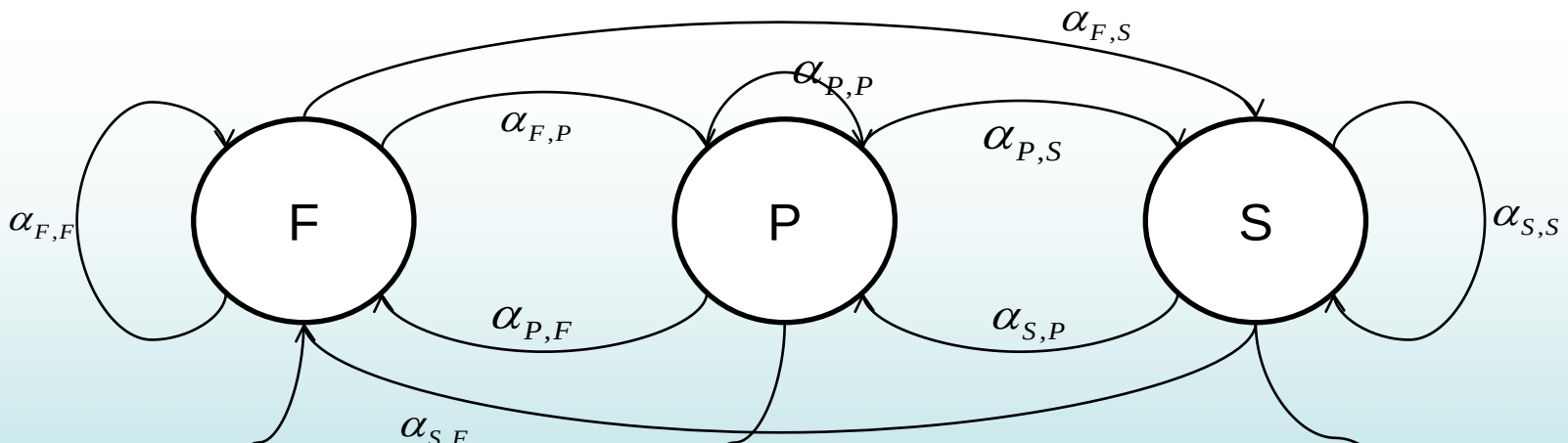


HMM for gaze tracking

Latent States:



HMM for gaze tracking



Three latent states

$Y_{n+1} = Y_n + w_F$

$Y_{n+1} = Y_n + Fixation$

$Y_{n+1} = \begin{cases} Y_n + G & if \quad X_n \neq S \\ Y_n + g_{n-k:n} + W_{S,k} & o.w. \end{cases}$

Transition probability

α

P: Pursuit

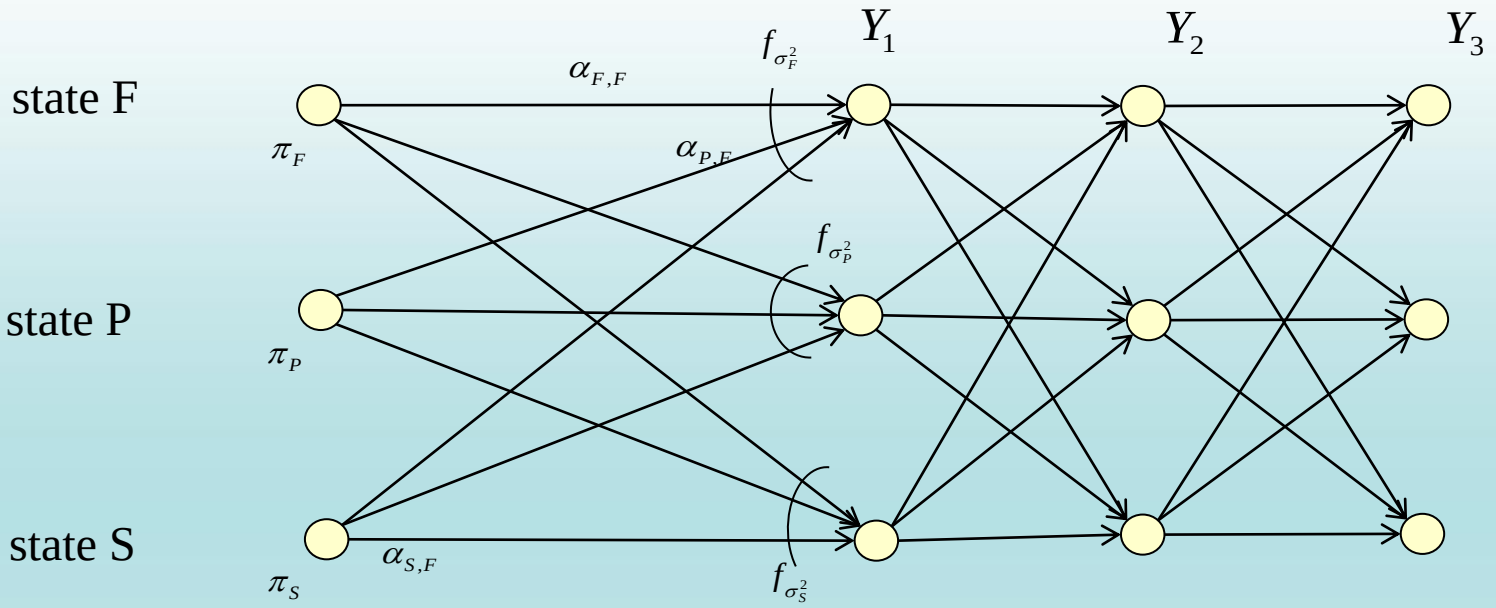
S: Saccade

Observations thru ***emission probabilities***

Y_n : x or y coordinate value

State Estimation

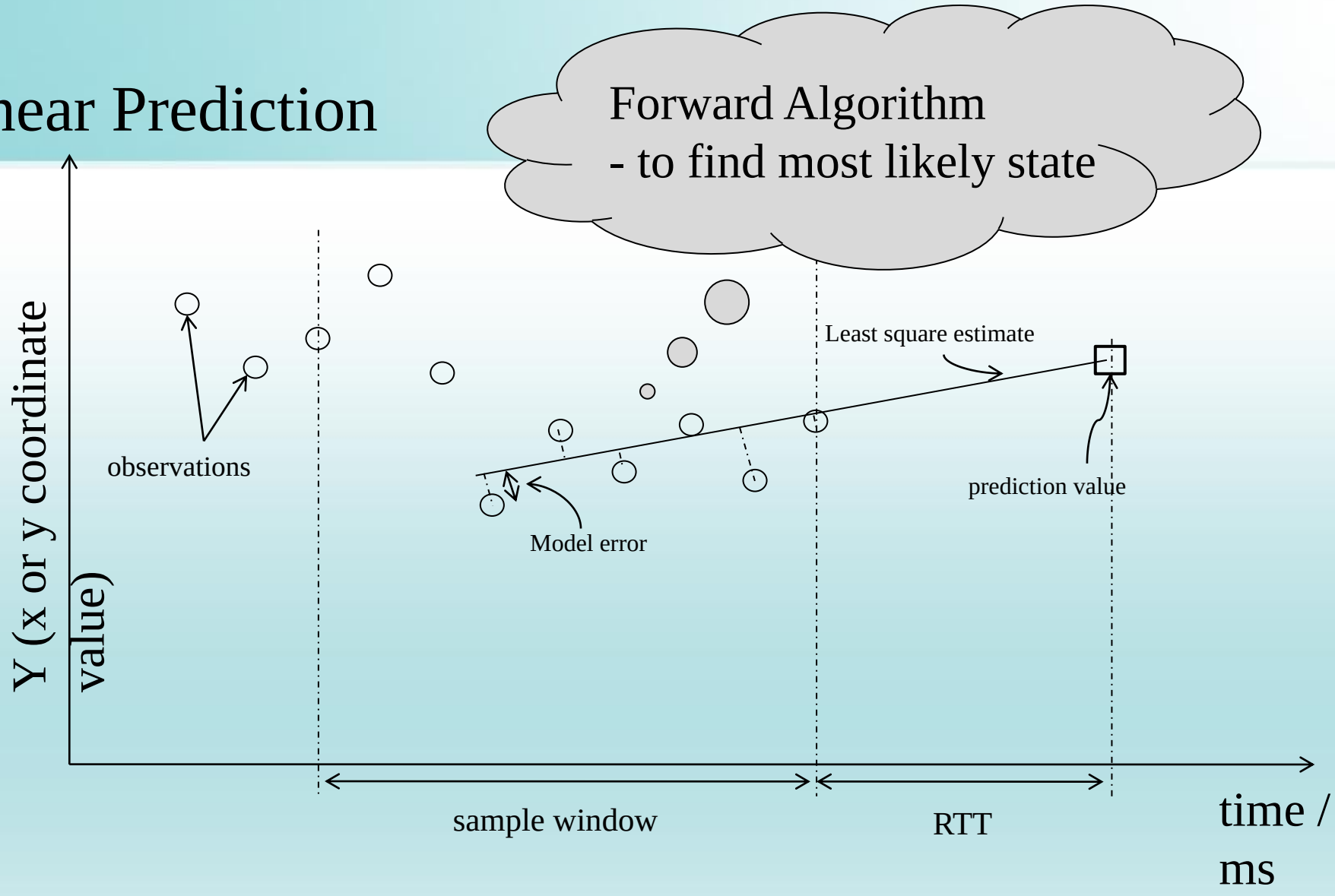
- Find most likely latent states (Forward Algorithm[1])



$$P(X_n = j) = \sum_i P\left(X_{n-1} = i\right)\alpha_{i,j} P\left(Y_n \mid Y_{n-1}, X_n = j\right)$$

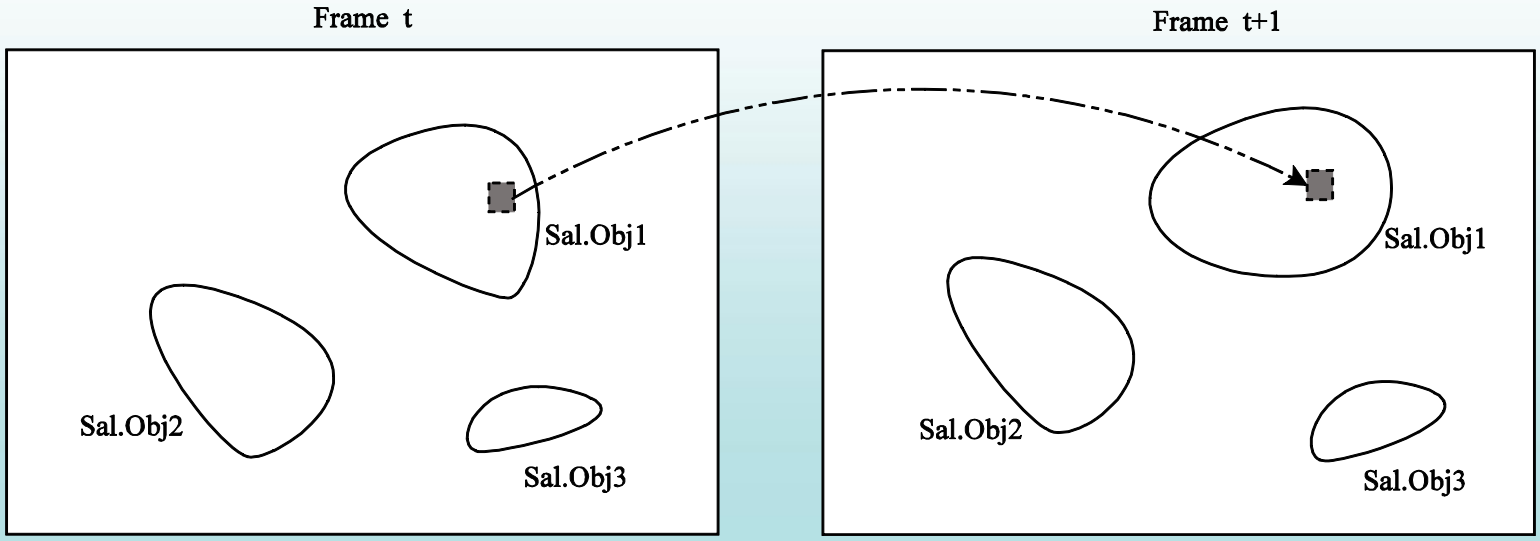
[1] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Linear Prediction



Bit-allocation Strategy

Def'n: ***saliency object*** is contiguous region above threshold.



- Predict only if enough samples are all F or all P, *and* state est. prob $\geq \tau_c$
- Smartly allocate bits iff prediction lands in same saliency object.

Experimental Setup

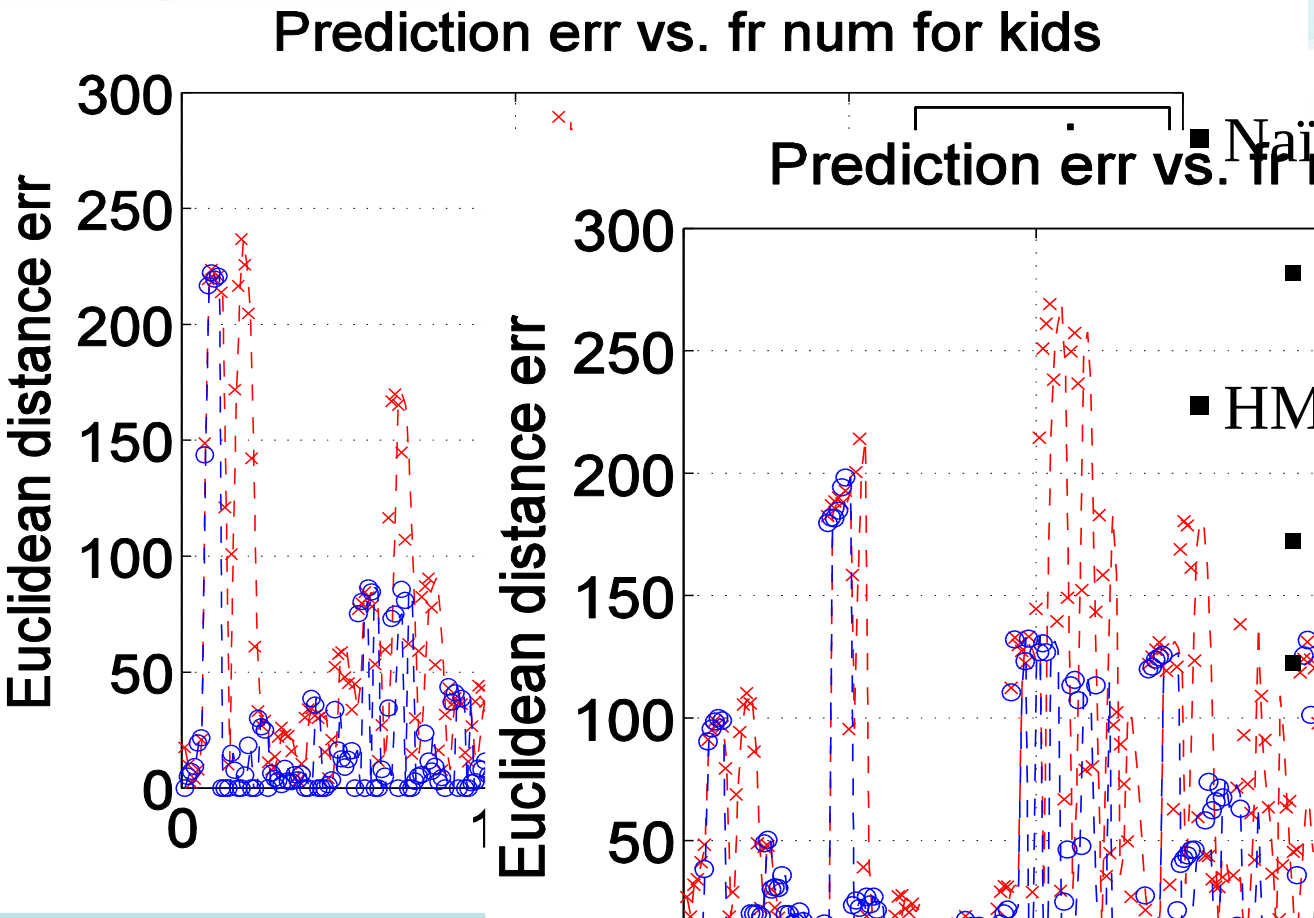
- Ga
  - 3
  - V
    - or: 22 inch (
- MP
  - C
  - 300 frames per video
  - Display rate: 30 fps

kids_cif

table_cif

http://www.inference.phy.cam.ac.jk/opengazer/

Experimental Results

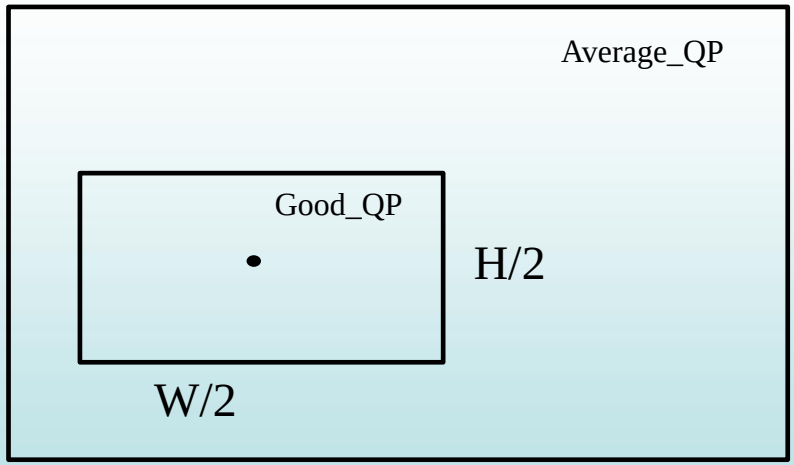


- Naïve linear prediction (nlp)
- Predict using prev. 2 gaze pts.
- HMM
- Predict only when confident.
- Different pred. method for F/P.

Bit Allocation Results

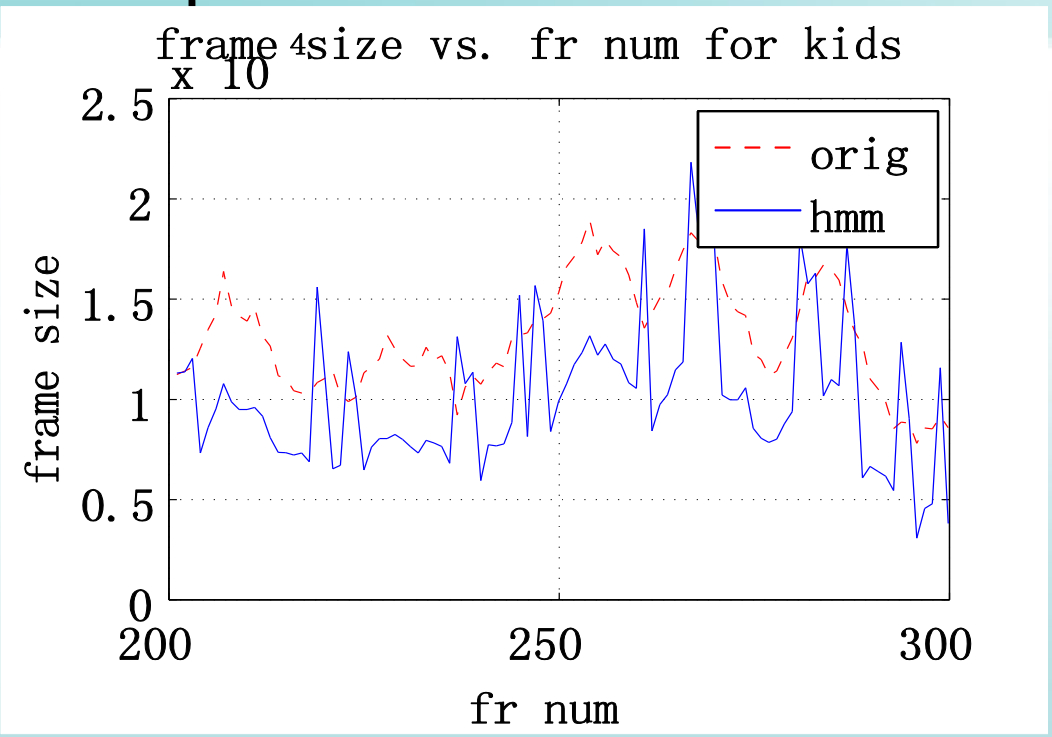
- QP=10 for a desired reference quality

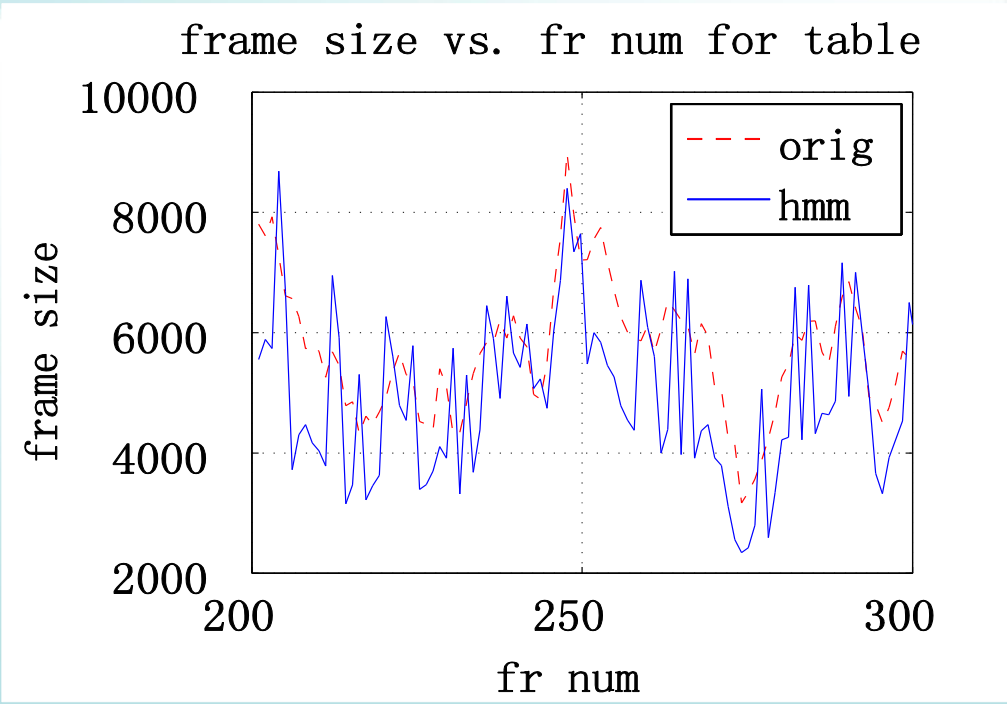
  QP=15 for outside quality



- Two levels of quality for ROI / non-ROI differentiation.
- More levels possible in general for more graceful degradation.

Experimental Results





- **hmm** saves 21%, 17% compared to **orig** for kids and table.
- **Subjective testing**: (real-time ROI encoding system, RTT = 200ms)
  - 1 of 3 videos is encoded by HMM. Test subjects could not identify HMM sequence.

Outline

- Overview / update of my research
- Eye-gaze prediction for network video streaming
- Ditto for store-&-playback video
- Visual attention deviation (VAD)
- Conclusion

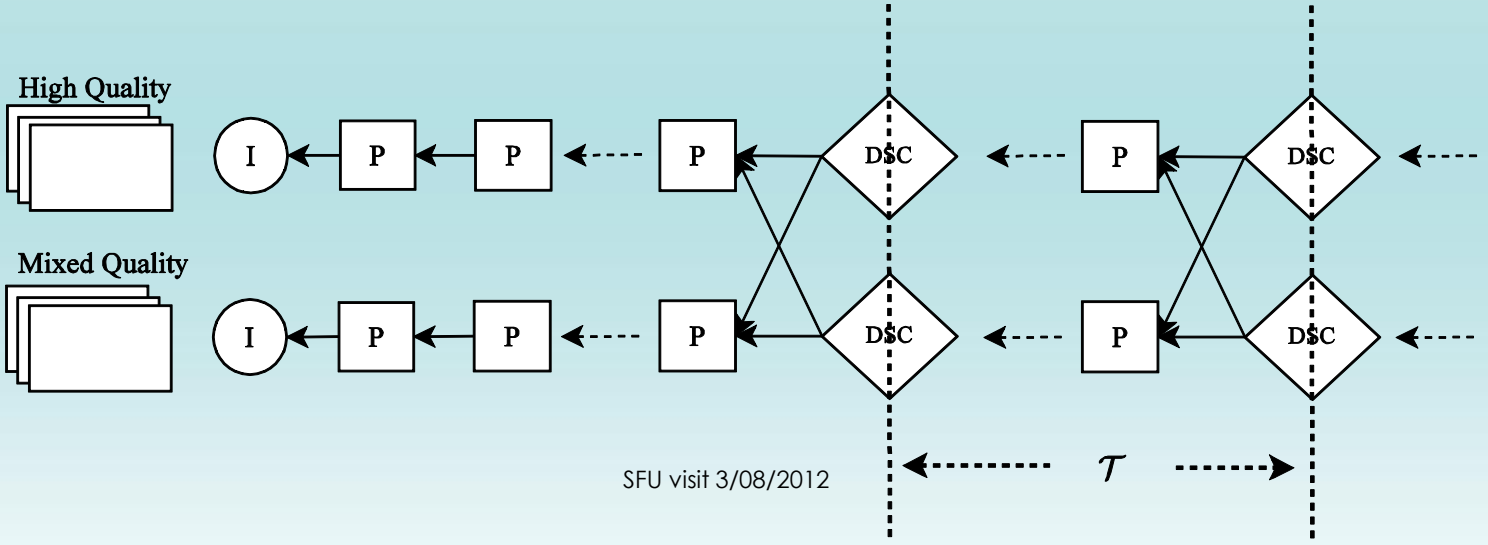
Store-&-playback video

- Real-time encoding per-client is expensive.

Question: Video adaptation for stored video?

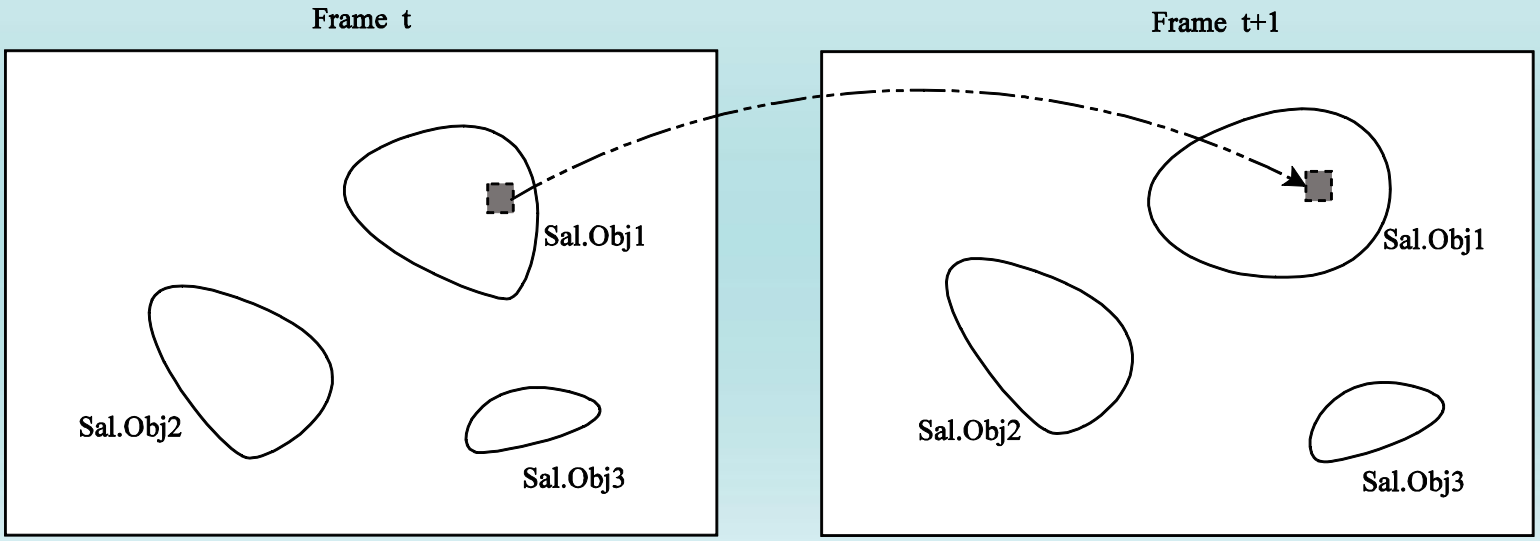
Answer: Yes.

1. Prepare two sub-streams:
   1. **HQ**: HQ everywhere.
   2. **MQ**: HQ for saliency objects, LQ elsewhere.
2. Periodically insert DSC frames for stream-switching.



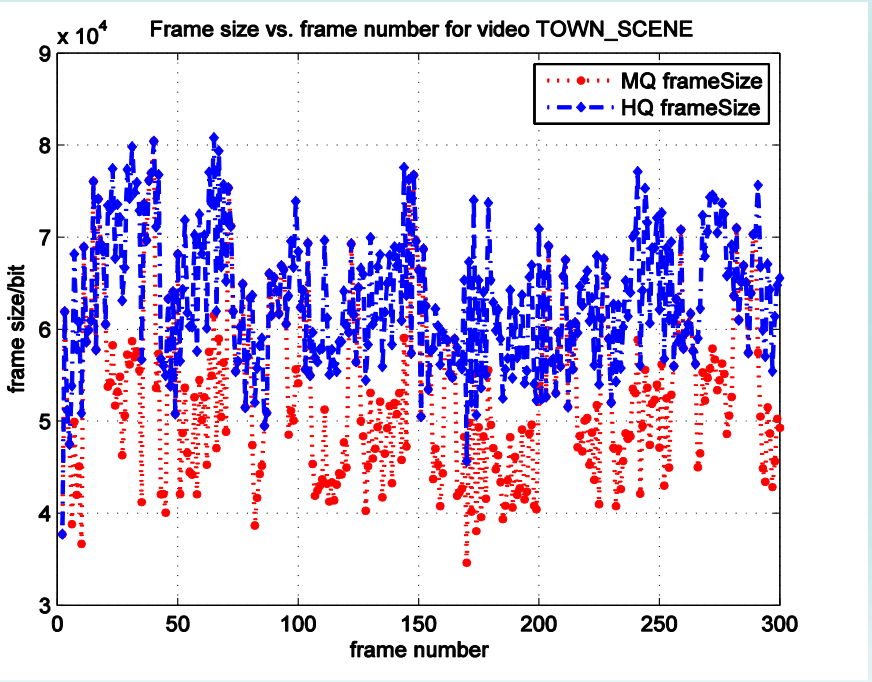
Store-&-playback video

3. If observer's gaze outside Sal. Obj at view-switching point, use HQ.
4. If observer's gaze inside Sal. Obj, *AND* prob leaving Sal. Obj till next switch point $\leq \varepsilon$, use MQ.



Experimental Results

- H.264 coding two HD seqs. HQ (QP = 10), MQ (QP = 10, 12)
- Subjective testing: (5 test subjects)
  - HQ, MQ, LQ randomly shown. Which one is poor quality?
  - 3 identified LQ.

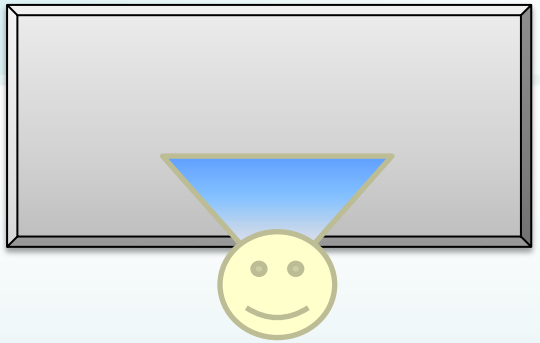


Park_joy	Town_scene
20.41%	24.08%

Outline

- Overview / update of my research
- Eye-gaze prediction for network video streaming
- Ditto for store-&-playback video
- Visual attention deviation (VAD)
- Conclusion

Introduction



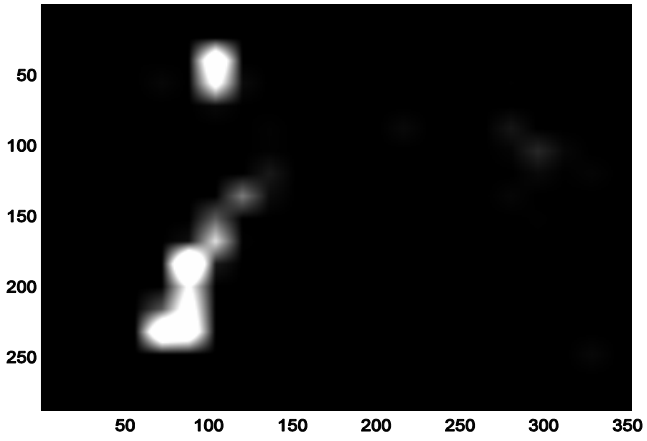
- A viewer sitting at a comfortably close screen cannot observe all spatial regions.
  - Gaze shifts.
- Different videos induce different amount of gaze shifts.
  - Example: presidential address vs. music video.
- What we benefit from ROI prediction
  - ROI-based bit allocation scheme.
  - Real-time customization of video content.

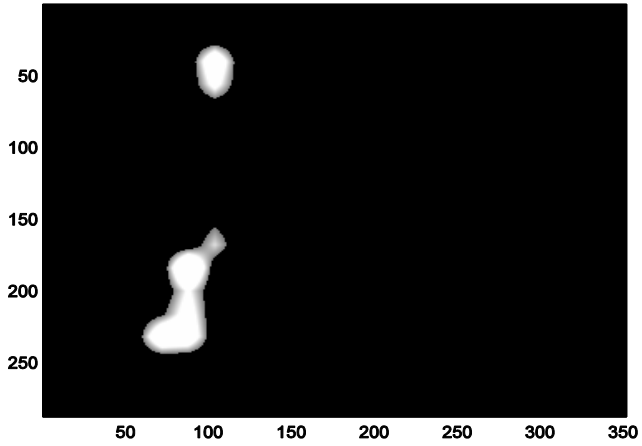
Goal: Measure of how often observer shifts visual attention?

VAD: visual attention deviation.

Define Saliency Objects

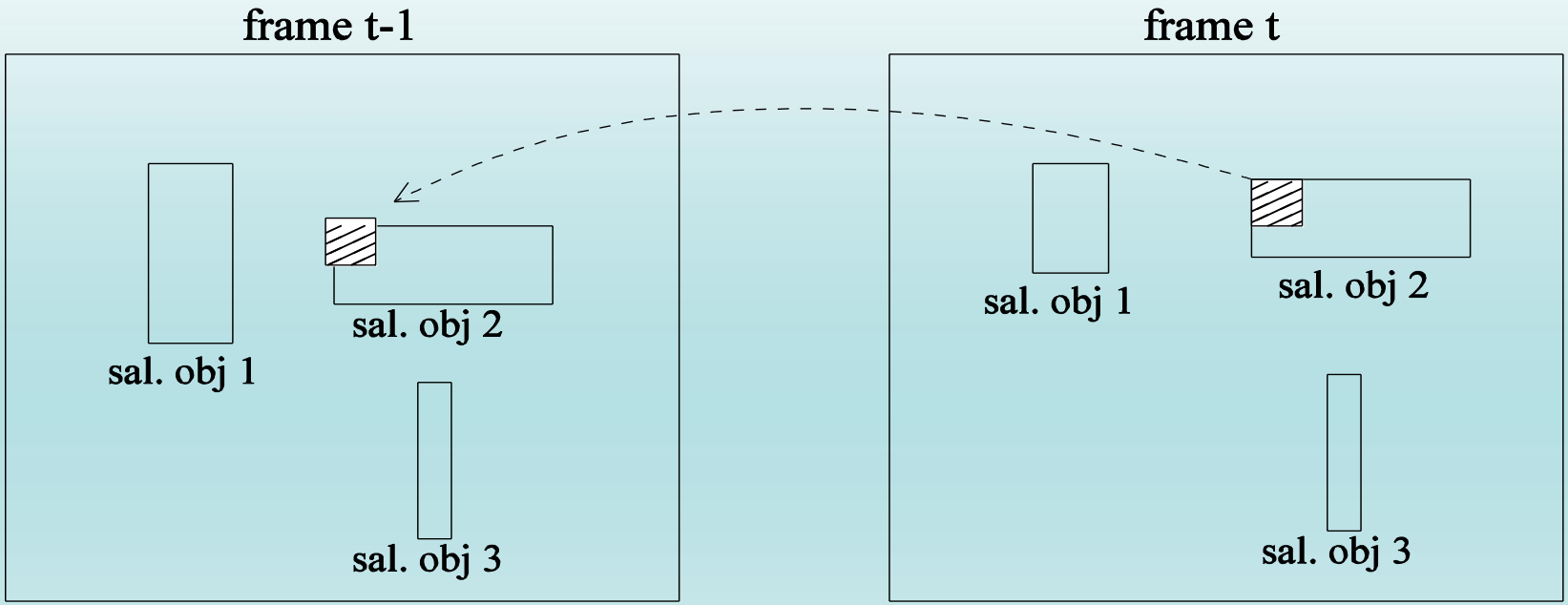
- Analysis of Visual Saliency Maps





Find similar saliency objects

- **Assumption**: saliency objects are only possible ROIs in frames.
- Establish correspondence among saliency objects in consecutive frames using motion estimation.

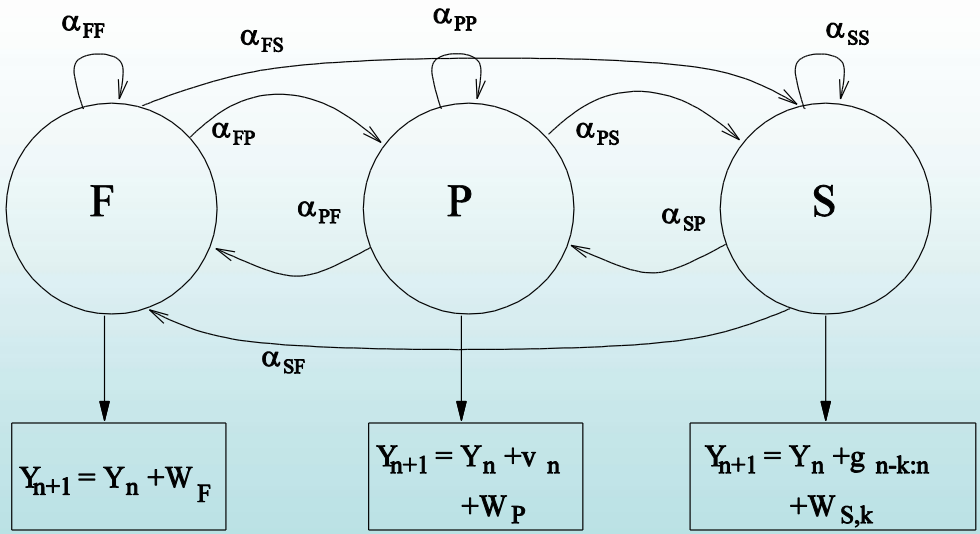


Derive VAD from Saliency Maps

- VAD is steady-state saccade probability:
  1. Define prob of saliency objects.
  2. Write consistency equations for consecutive frames.

$$p_{t+1,1} = p_{t,1}\alpha_{FF} + s_t\alpha_{SF}\left(\frac{p_{t+1,1}}{p_{t+1,1} + p_{t+1,2}}\right)$$

  3. Together w/ total prob theorem, compute HMM parameters, steady state prob.



- "Stationarity" of gaze statistics:
  1. Compute motion-compensated saliency maps.
  2. Compute Kullback-Leibler (KL) Divergence between maps.

Experimental Results

- 4 test sequences:
  - i) two 300-frame standard MPEG video test sequences, at CIF resolution (352x288).
  - Ii) two 250-frame higher resolution video sequences, at SD resolution (720x576).

All videos have 30 fps.

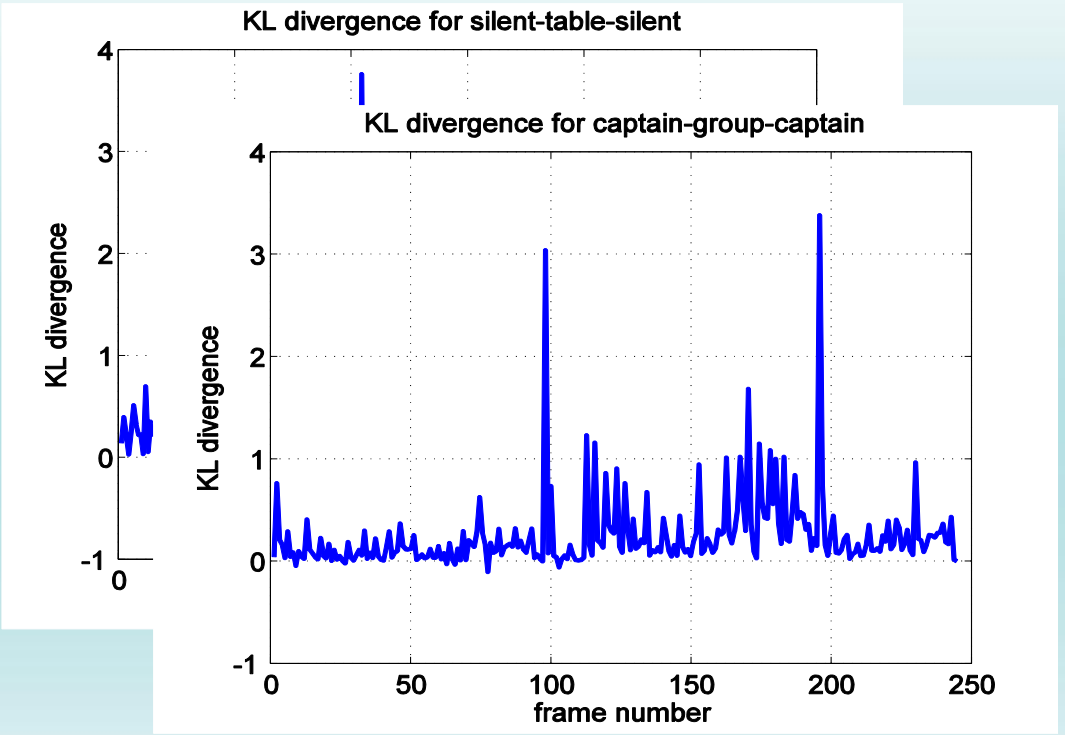
Seq	Gaze data	Saliency map analysis
Cif_silent (300 frames)	0.063	0.089
Cif_table (300 frames)	0.432	0.442
SD_captain (250 frames)	0.152	0.181
SD_parkjoy (250 frames)	0.439	0.457

Experimental Results

Use computed KL divergence to segment video into clips of "stationary" gaze statistics.

100-frame silent
+ 100-frame table
+ 100-frame silent

100-frame captain
+ 100-frame group
+ 50-frame captain



Outline

- Overview / update of my research
- Eye-gaze prediction for network video streaming
- Ditto for store-&-playback video
- Visual attention deviation (VAD)
- Conclusion

Conclusion & Future Work

- Computer watching u watching media.
- Gaze prediction for ROI-based bit allocation.
  - Real-time encoding.
  - Store-and-playback video.
- VAD: a measure of predictability of gaze.
- Can we do more?
  - Pre-attentive stimulus vs. top-down motives.
  - Deduce semantic information w/ implicit cues.