YORK
UNIVERSITÉ
UNIVERSITY

UNIVERSITÉ DE NANTES

KANDAO

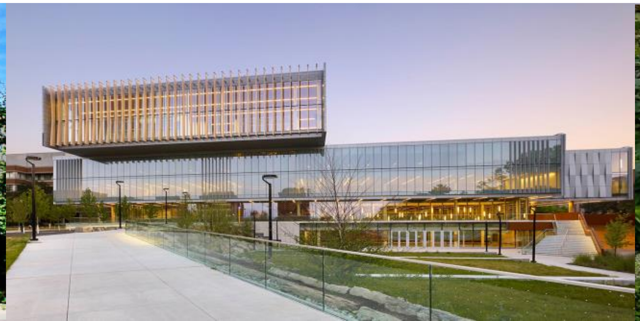
Sparse Directed Graph Learning for Head Movement Prediction in 360 Video Streaming

Xue Zhang, Gene Cheung,
Patrick Le Callet, and Jack Z. G. Tan

ICASSP2020
Barcelona

Session: TH3.PI: Image/Video Processing II

--7th May, 2020

Outline

- Motivation
- Related Works
- Contributions
- Sparse Directed Graph Learning
- Experiments

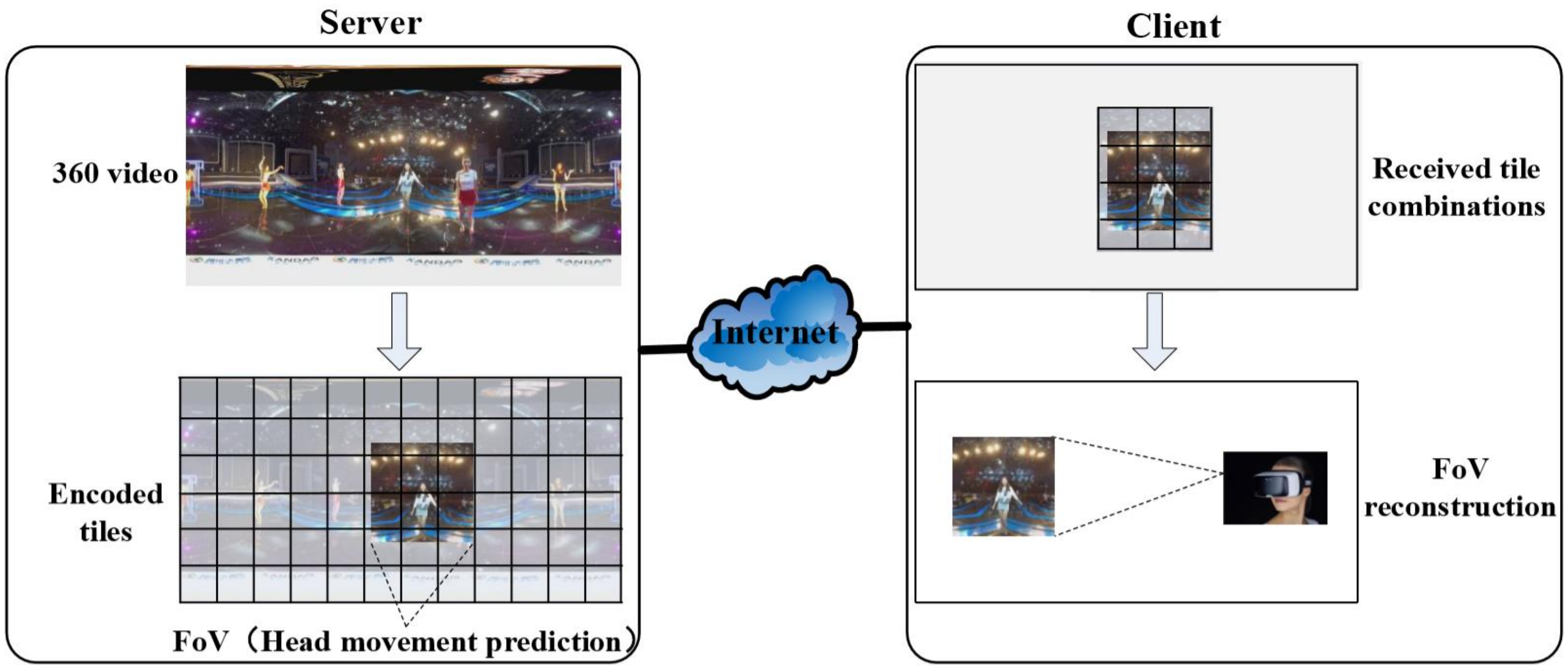
Outline

- Motivation
- Related Works
- Contributions
- Sparse Directed Graph Learning
- Experiments

Gene Cheung(genec@yorku.ca)

Motivation

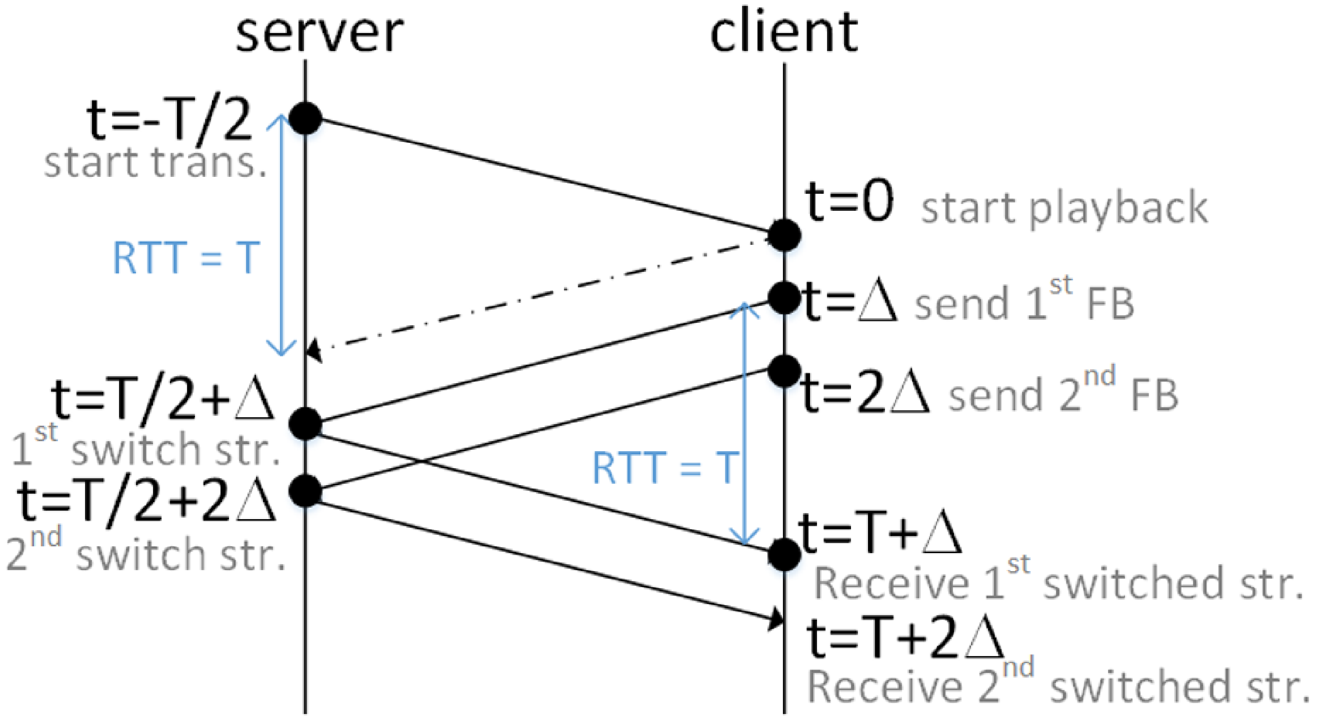
➢ Challenges in interactive 360 video streaming scenario



- 360 videos: high spatial resolution (*e.g.*,10K 10240×4320)
- Bandwidth-limited networks
- Extract and transport only a sub-region corresponding to a viewer's current field-of-view (FoV)
- Round-trip-time (RTT) delay: head movement prediction foretelling a viewer's future FoVs

Motivation (cont'd)

- **What is RTT?**



Interaction between server and client where RTT is T and frame interval is Δ.
A switched stream arrives T seconds after a feedback is sent.

Gene Cheung(genec@yorku.ca)
LASSONDE SCHOOL OF ENGINEERING | YORK UNIVERSITÉ UNIVERSITY

Outline

- Motivation
- Related Works
- Contributions
- Sparse Directed Graph Learning
- Experiments

Related works

- **Linear regression models** [1][2]
  - pro: historical samples and dead-reckoning algorithms to extrapolate the trends
  - con: prediction accuracy drops precipitously for larger RTTs

- **Pure data-driven model learning**
  - pro: using neural networks [3] or reinforcement learning scheme [4]
  - con: 1) a huge dataset of traces for training a large number of network parameters;
    2) training is typically specific to particular setups (*e.g.*, RTT mean and variance).

[1] L. Xie, Z. Xu, Y. Ban, X. Zhang, and Z. Guo, “360probdash: Improving QOE of 360 video streaming using tile-based http adaptive streaming,” *ACM MM’17*, pp. 315–323.
[2] S. Petrangeli, V. Swaminathan, M. Hosseini, and F. De Turck, “An HTTP/2-based adaptive streaming framework for 360 virtual reality videos,” *ACM MM’17*, pp. 306–314.
[3] C.-L. Fan, S.-C. Yen, C.-Y. Huang, and C.-H. Hsu, “Optimizing fixation prediction using recurrent neural networks for 360 video streaming in head-mounted virtual reality, *TMM*, vol.22, no.3, pp. 744 – 759, March 2020.
[4] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, “Predicting head movement in panoramic video: A deep reinforcement learning approach,” *TPAMI*, vol. 41, no. 11, pp. 2693–2708, July 2018.

Related works (cont'd)

- **Visual attention (VA) detection** (*e.g.*, ICME Grand Challenge "salient360!")

  - **pro**: 1) datasets [5];
    2) toolbox to facilitate the development of VA models [6];
    3) framework to evaluate VA models [7];
    4) ad-hoc VA models for 360 contents [8].

  - **con**: 1) more an "aggregate" behavior rather than an individual behavior;
    2) target prediction is in time horizon of typically 10s to 15s viewing time not the typical RTT.

[5] Y. Rai, J. Gutierrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," *ACM MMSys'17*, pp. 205–210.
[6] J. Gutierrez, E. David, Y. Rai, and P. Le Callet, "Toolbox and dataset for the development of saliency and scanpath models for omnidirectional / 360◦ still images," *Signal Processing: Image Communication*, vol. 69, pp. 35–42, November 2018
[7] M. Silva, J. Gutierrez, A. Coutrot and P. Le Callet, "Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360◦ contents," *IEEE QoMEX'18*, Italy.
[8] Y. Zhu, G. Zhai, and X. Min, "The prediction of head and eye movement for 360 degree images," *Signal Processing: Image Communication*, vol. 69, pp. 15–25, 2018.

Outline

- Motivation
- Related Works
- Contributions
- Sparse Directed Graph Learning
- Experiments

Contributions

Sparse directed graph learning

discrete angles in the 360 view are nodes in a graph

a 360 image saliency map + collected viewers' head movement traces + a biological head rotation model

an estimate of stationary probability distribution

instantiation of the state transitions

physical constraints on the state transitions

a unified Markov model → **a probability transition matrix $\boldsymbol{P}$**

One can evaluate the view probability distribution $\boldsymbol{v}_{t+T}$ one RTT T later as $\boldsymbol{v}_{t+T} = \boldsymbol{v}_t \boldsymbol{P}^T$ given original distribution $\boldsymbol{v}_t$ at time t.

Gene Cheung(genec@yorku.ca)

Outline

Gene Cheung(genec@yorku.ca)

Define two variables:

- $\boldsymbol{P}$: $K \times K$ *view transition probability matrix* ($360°$ sphere is discretized uniformly into K angles)
- $\boldsymbol{q}$: stationary *view probability vector*
- $\boldsymbol{qP} = \boldsymbol{q}$

A *maximum a posteriori* (MAP) optimization problem to find a Markov model for head movement prediction

➢ **Likelihood Term** (depends on data traces)

$$P(\mathcal{X}|\boldsymbol{\theta}) = \prod_{k=1}^{K} q_k^{N_k} \prod_{l=1}^{K} p_{kl}^{N_{kl}}$$

number of occurrences of angle k in set $\mathcal{X}$

number of occurrences of switching from angle k to angle l in set $\mathcal{X}$

where $\mathcal{X}$ is the training set of observed angle switches in traces

$$\boldsymbol{\theta} = \{\{q_k\}, \{p_{kl}\}\}$$

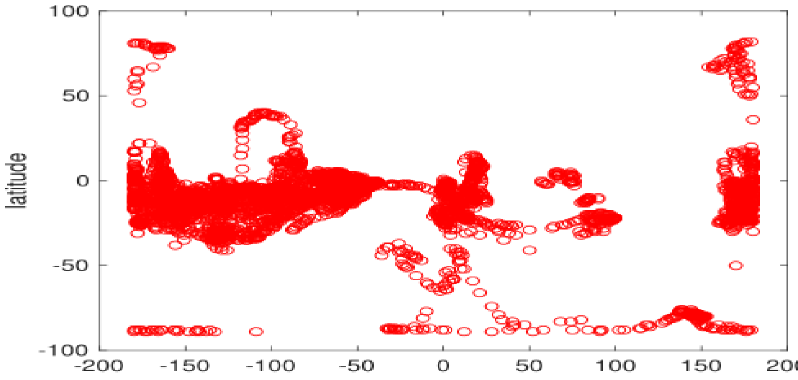
Sparse Directed Graph Learning (cont'd)

➢ Prior Term

- The prior for $\boldsymbol{q}$ depends on a computed *360 saliency* map [6]

$$P(\mathbf{q}) = \prod_{k=1}^{K} \exp\left(\frac{-(q_k - \hat{q}_k)^2}{\sigma_q^2}\right)$$

the normalized saliency of angle k



user scanpath

- The prior for $\boldsymbol{P}$ depends on a *sparse graph assumption*

$$P(\mathbf{P}) = \exp\left(\frac{- \| \mathbf{P} \|_0}{\sigma_p^2}\right)$$

[6] J. Gutierrez, E. David, Y. Rai, and P. Le Callet, "Toolbox and dataset for the development of saliency and scanpath models for omnidirectional / 360◦ still images," *Signal Processing: Image Communication*, vol. 69, pp. 35–42, November 2018.

Sparse Directed Graph Learning (cont'd)

- **MAP Estimation**

$$\arg\min_{\{\{q_k\},\{p_{kl}\}\}} -\sum_{k=1}^{K}\left(N_k \ln q_k + \sum_{l=1}^{K} N_{kl} \ln p_{kl}\right) + \sum_{k=1}^{K}\frac{(q_k-\hat{q}_k)^2}{\sigma_q^2} + \frac{1}{\sigma_p^2}\sum_{k,l}\omega_{kl}\,(p_{kl})^2 \quad (10)$$

$$\text{s.t.} \quad \sum_{k=1}^{K} q_k p_{kl} = q_l, \quad \forall l \quad (8c)$$

$$\sum_{k=1}^{K} q_k = 1, \quad \sum_{l=1}^{K} p_{kl} = 1, \quad \forall k \quad (8d)$$

$$q_k \geq \epsilon_q, \quad \forall k, \quad p_{kl} \geq \begin{cases} \epsilon_p, & \text{if } \forall k, \forall l \in \mathcal{N}(k) \\ 0, & \text{otherwise} \end{cases} \quad (8e)$$

Iterative reweighted least square (IRLS) [9]

$$\omega_{kl} = \frac{1}{\left(\tilde{p}_{kl}^2 + \varepsilon_s\right)}$$

using previous estimate $\tilde{p}_{kl}$ to promote sparsity in $\boldsymbol{P}$

the neighborhood of K, to ensure that transition probabilities between adjacent angles are non-zero based on a biological head movement model.

[9] I. Daubechies, R. DeVore, M. Fornasier, and C S. Gunturk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 63, no. 1, pp. 1–38, 2010.

Sparse Directed Graph Learning (cont'd)

- **Optimizing P when q is fixed:**

$$\min_{\mathbf{P}} -\sum_{k=1}^{K}\sum_{l=1}^{K} N_{kl}\ln p_{kl} + \frac{1}{\sigma_p^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\omega_{kl}\left(p_{kl}\right)^2$$

$$\text{s.t.} \sum_{k=1}^{K} q_k p_{kl} = q_l,\ \ \forall l, \quad \sum_{l=1}^{K} p_{kl} = 1,\ \ \forall k,$$

$$p_{kl} \geq \begin{cases} \epsilon_p, & \text{if } \forall k, \forall l \in \mathcal{N}(k) \\ 0, & \text{otherwise} \end{cases}$$

- **Optimizing q when P is fixed:**

$$\min_{\mathbf{q}} -\sum_{k=1}^{K} N_k \ln q_k + \sum_{k=1}^{K}\frac{(q_k-\hat{q}_k)^2}{\sigma_q^2}$$

$$\text{s.t.} \quad \max\{\epsilon_q, \tilde{q}_k - \delta\} \leq q_k \leq \max\{\epsilon_q, \tilde{q}_k + \delta\}, \quad \forall k$$

$$\sum_{k=1}^{K} q_k = 1$$

$\tilde{q}P = \tilde{q}$

the eigenvector of P corresponding to eigenvalue 1.

❖ Frank-Wolfe optimization strategy (projection-free):

linear approximation in each iteration → Find the optimal direction $\mathbf{s}$ by solving:

$$\min_{\mathbf{s}} \ \mathbf{s}^\top \nabla f(\mathbf{P}(t)) \quad \text{such that } \mathbf{s} \in \mathcal{R}$$

← *Linear Program* benefits from warm start.

Gene Cheung(genec@yorku.ca)

Outline

- Motivation
- Related Works
- Contributions
- Sparse Directed Graph Learning
- Experiments

Experiments

- **Six 360 VR sequences** at 30fps with length around 60 seconds [10]
  - two 8K resolution (7680 × 3840) with around 110 traces
  - four 4K resolution (3840 × 1920) with around 50 traces

- **Comparison algorithms**
  - regression models:
    linear regression "LR" [1], weighted linear regression "WLR" [11] and "Heuristic" [2].
  - a naive approach: "Saliency".

- **Prediction error of each trace**: $Er = -\frac{\sum_{t=1}^{L} \ln g_t(T)}{L}$

Where $g_t(T)$ is the view probability of correct prediction for each instant t.

L is the length of each collected trace.

[1] L. Xie, Z. Xu, Y. Ban, X. Zhang, and Z. Guo, "360probdash: Improving QOE of 360 video streaming using tile-based http adaptive streaming," *ACM MM'17*, pp. 315–323.
[2] S. Petrangeli, V. Swaminathan, M. Hosseini, and F. De Turck, "An HTTP/2-based adaptive streaming framework for 360 virtual reality videos," *ACM MM'17*, pp. 306–314.
[10] https://www.kandaovr.com/
[11] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan, "Optimizing 360 video delivery over cellular networks," *All Things Cellular: Operations, Applications and Challenges*, 2016, pp. 1–6.

Experiments (cont'd)

Results

Table 1. The average Er of different models when $T = 0.5$s.

Seq.	LR	WLR	Heuristic	Saliency	Proposed
On the hill	5.61	5.58	6.29	4.49	**0.16**
Beijing	3.73	3.69	6.27	4.71	**0.15**
Guangzhou	1.35	1.34	3.62	4.87	**0.07**
Huizhou	3.07	2.97	5.67	4.19	**0.20**
Concert	9.83	9.51	6.46	4.31	**0.19**
Lamborghini	5.76	5.71	5.22	4.61	**0.13**

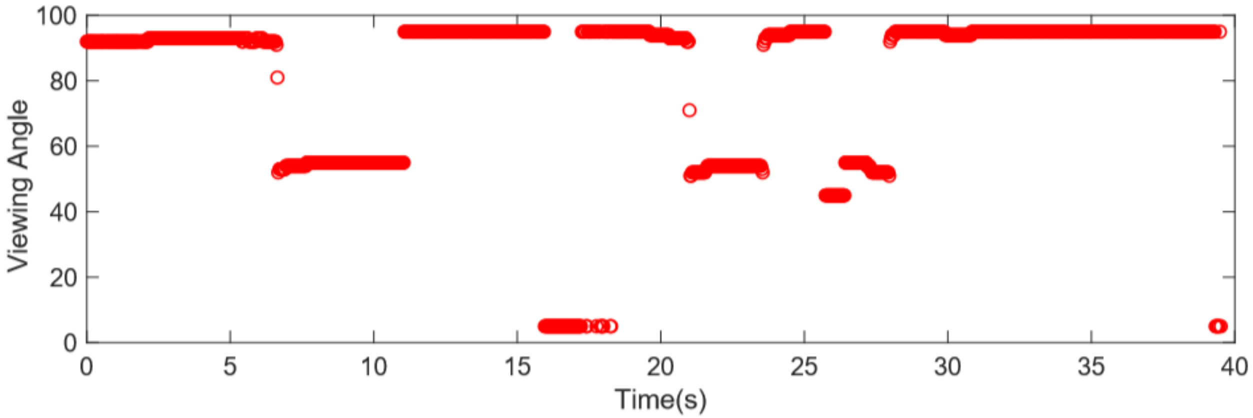
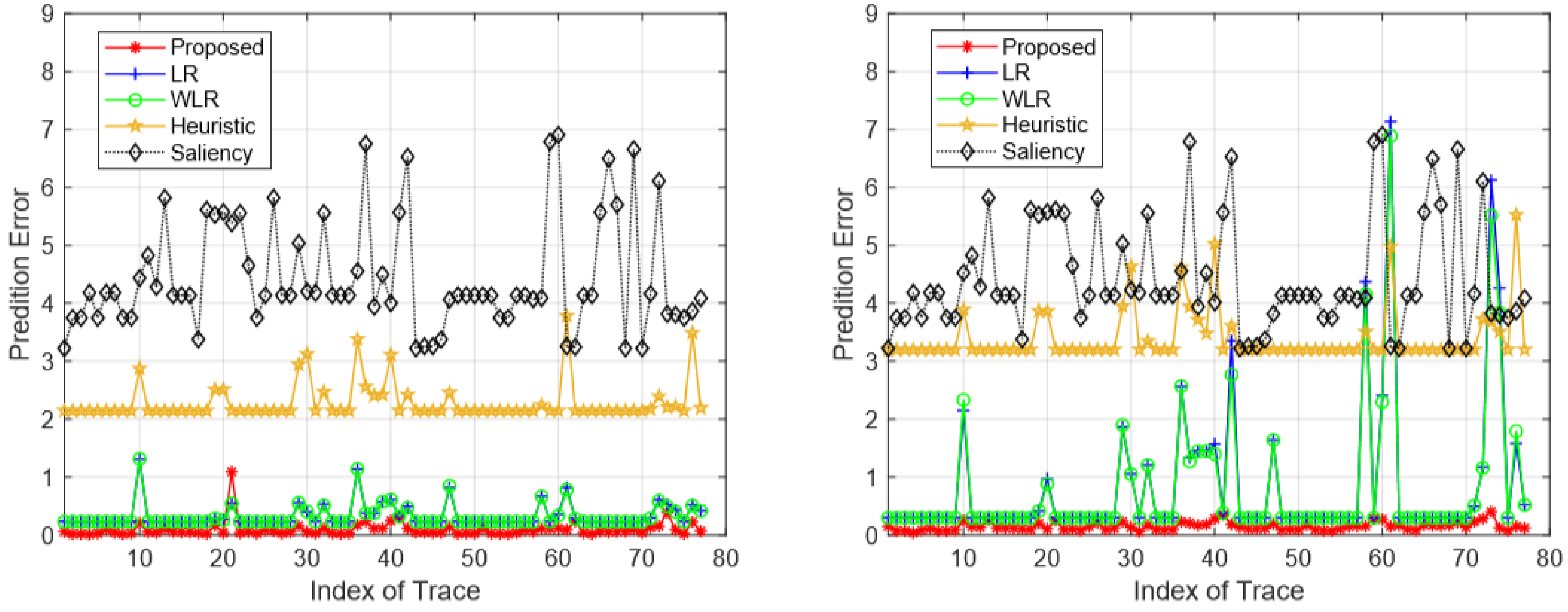


Fig. 2. The variation of user's angles over time in one trace of *On the hill*.

Note that when users have large and frequent head motions, it is difficult for competitors to predict accurately but not for our proposed model.

Gene Cheung(genec@yorku.ca)

Experiments (cont'd)



(a) $T = 0.5s$ (b) $T = 2s$

Fig. 3. Prediction error smaller than 10 for *On the hill*.

Benefiting from projection-free FW and warm start in LP, our strategy has reduced complexity.

Q&A

Thank You !

Gene Cheung(genec@yorku.ca)