Gene Cheung
National Institute of Informatics

28th November, 2017

国立情報学研究所
National Institute of Informatics

Interactive Media Streaming Applications Using Merge Frames

Acknowledgement

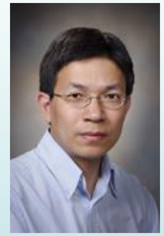
Collaborators:

- M. Kaneko (NII, Japan)
- A. Ortega (USC, USA)
- D. Florencio (MSR, USA)
- P. Frossard (EPFL, Switzerland)
- J. Liang, I. Bajic (SFU, Canada)
- V. Stankovic (U of Strathclyde, UK)
- X. Wu (McMaster U, Canada)
- P. Le Callet (U of Nantes, France)
- X. Liu (HIT, China)
- W. Hu, J. Liu, Z. Guo (Peking U., China)
- L. Fang (Tsinghua, China)
- C.-W. Lin (National Tsing Hua University, Taiwan)

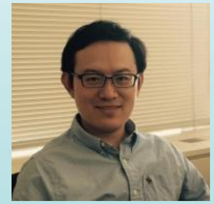


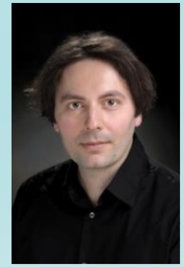








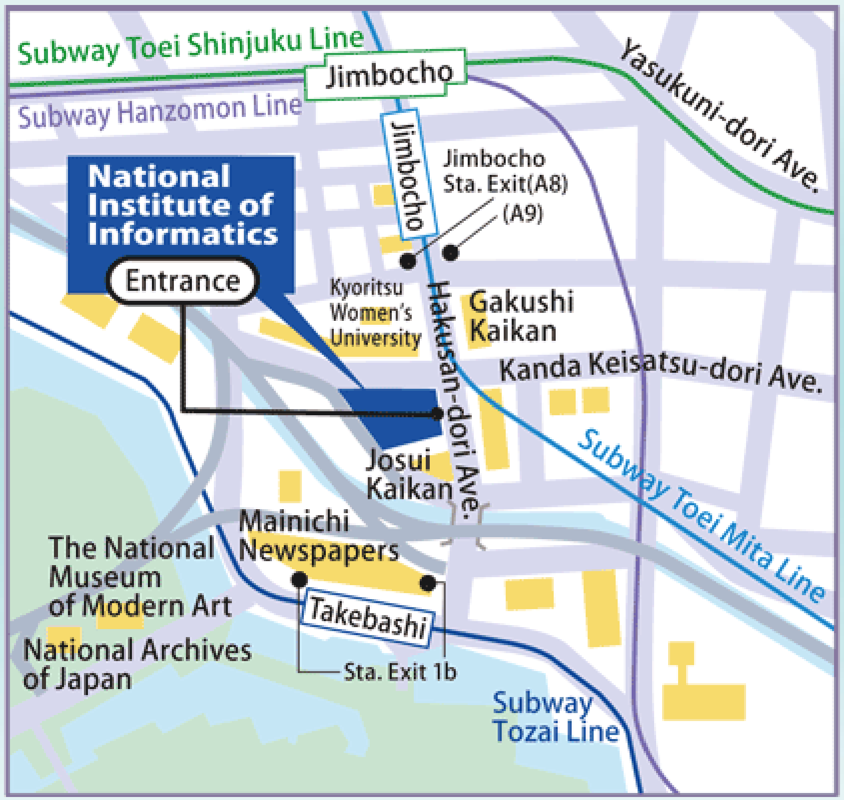




NII Overview

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

- **National Institute of Informatics**
- Chiyoda-ku, Tokyo, Japan.
- Government-funded research lab.



- Offers graduate courses & degrees through **The Graduate University for Advanced Studies** (Sokendai).
- 60+ faculty in "**informatics**": quantum computing, discrete algorithms, database, machine learning, computer vision, speech & audio, image & video processing.

- **Get involved!**
  - 2-6 month Internships.
  - Short-term visits via MOU grant.
  - Lecture series, Sabbatical.

Introduction to APSIPA and APSIPA DL

APSIPA Mission: To promote broad spectrum of research and education activities in signal and information processing in Asia Pacific

APSIPA Conferences: ASPIPA Annual Summit and Conference

APSIPA Publications: Transactions on Signal and Information Processing in partnership with Cambridge Journals since 2012; APSIPA Newsletters

APSIPA Social Network: To link members together and to disseminate valuable information more effectively

APSIPA Distinguished Lectures: An APSIPA educational initiative to reach out to the community

Outline

- What is interactive media navigation?
  - e.g. Multiview / free-viewpoint video
- Merge frame for interactive media navigation
  - Previous works
  - Merge frame / block overview
  - Fixed target merging
  - Optimized target merging
- Interactive Virtual Reality Video Streaming

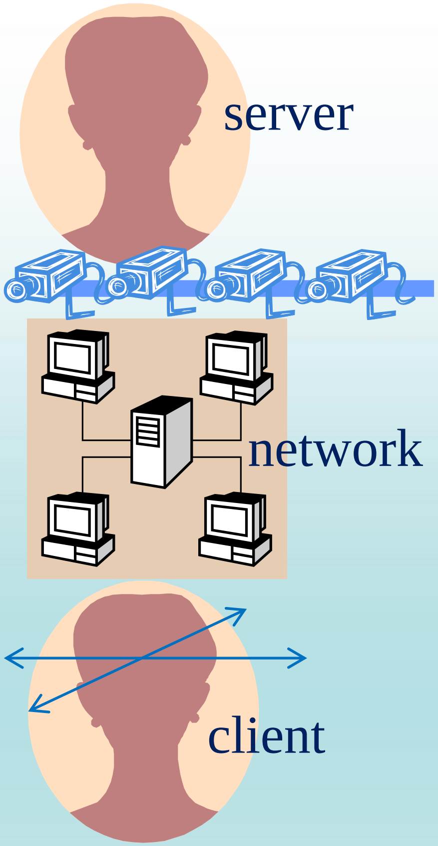
Wei Dai, Gene Cheung, Ngai-Man Cheung, Antonio Ortega, Oscar Au, "**Merge Frame Design for Video Stream Switching using Piecewise Constant Functions**," *IEEE Transactions on Image Processing*, vol. 25, no.8, August 2016

Gene Cheung, Zhi Liu, Zhiyou Ma, Jack Z. G. Tan, "**Multi-Stream Switching for Interactive Virtual Reality Video Streaming**," *IEEE International Conference on Image Processing*, Beijing, China, September, 2017.

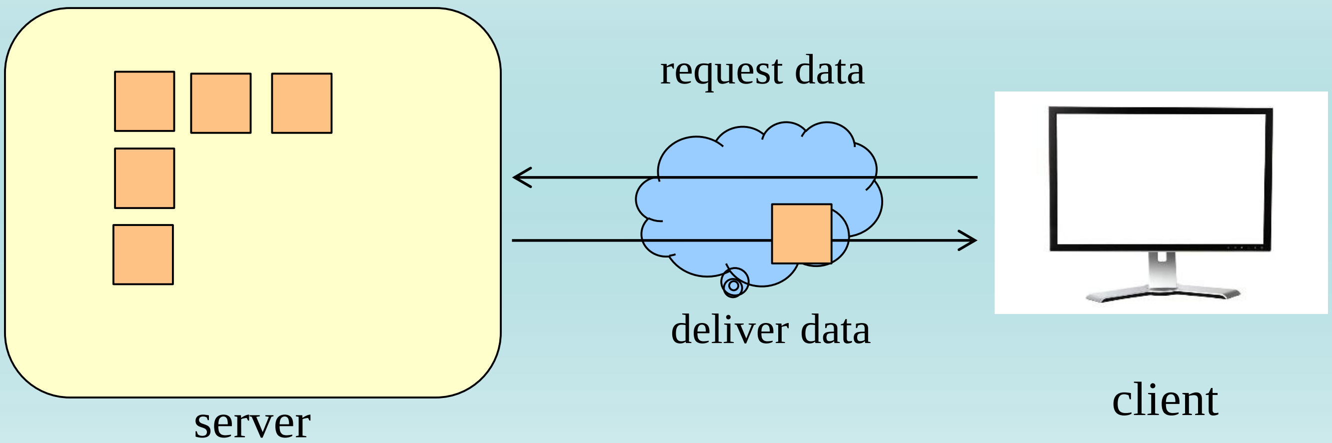
Outline

- What is interactive media navigation?
    - e.g. Multiview / free-viewpoint video
- Merge frame for interactive media navigation
    - Previous works
    - Merge frame / block overview
    - Fixed target merging
    - Optimized target merging
- Interactive Virtual Reality Video Streaming

What is interactive media navigation / streaming?



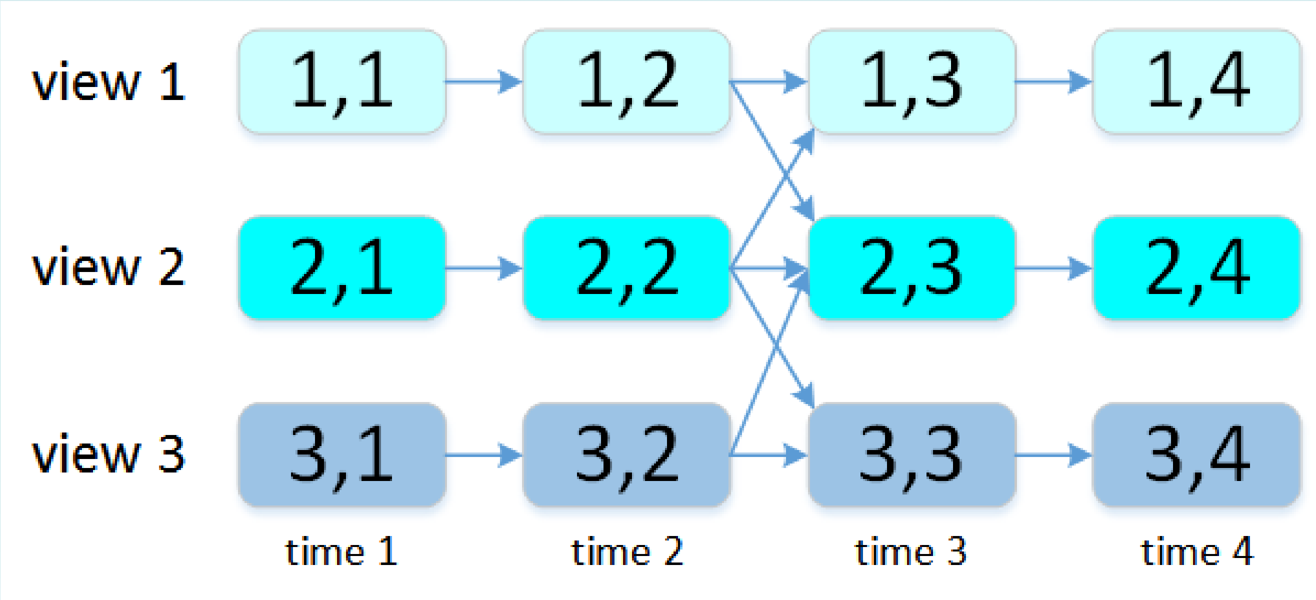
- **Server:** a very large correlated media data set.
  - *e.g.*, multiview video, light field data, etc.
- **Client:** can observe only small data subset at a time.
- **Network**: cannot deliver whole dataset before start of navigation.
- **Interactive navigation**: client requests data, server sends data. Repeat.



G. Cheung, A. Ortega, N.-M. Cheung, B. Girod, "**On Media Data Structures for Interactive Streaming in Immersive Applications**," in *SPIE Visual Communications and Image Processing*, Huang Shan, China, July, 2010.

Interactive Multiview Video Streaming (IMVS)

- **Server:** multiple views of same video captured synchronously in time.
- **Client:** can observe only 1 view at a time.
- **Interactive navigation**:
  - Client plays back video in time uninterrupted.
  - Client requests view, server sends view. Repeat.



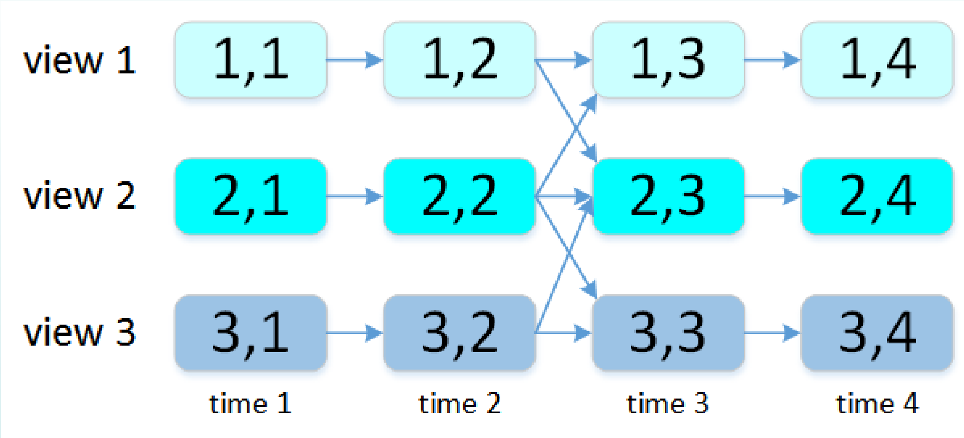
G. Cheung, A. Ortega, N.-M. Cheung, "**Interactive Streaming of Stored Multiview Video using Redundant Frame Structures**," *IEEE Transactions on Image Processing*, vol.20, no.3, pp.744-761, March 2011.

Outline

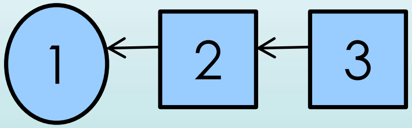
- What is interactive media navigation?
    - e.g. Multiview / free-viewpoint video
- Merge frame for interactive media navigation
    - Previous works
    - Merge frame / block overview
    - Fixed target merging
    - Optimized target merging
- Interactive Virtual Reality Video Streaming

Merge Frame for Media Navigation: conflicting coding requirements

- Inherent tension between coding efficiency & flexible decoding.



- ***Differential coding*** assumes ***single*** order of frame decoding.

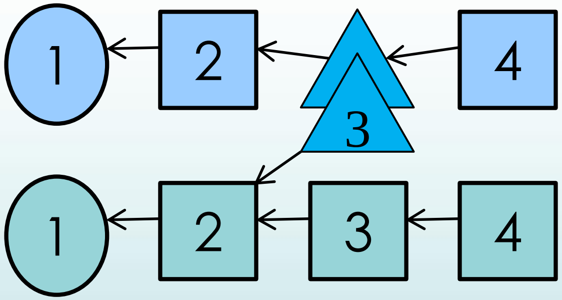


- ***Flexible decoding*** assumes ***several*** orders (paths) of frame decoding.

- **Other examples**:

Research Question: How to enable flexible decoding *without* great sacrifice of coding performance?

Merge Frame for Media Navigation: previous works 1



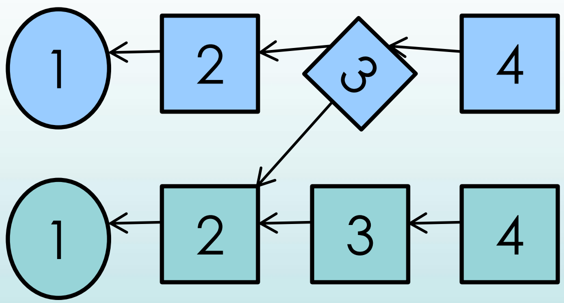
- **SP frames** (H.264 extended profile):
  - ***Primary SP-frame***: motion prediction + extra quantization. (small).
  - ***Secondary SP-frame***: motion prediction + lossless encoding. (large).
- **Pros**: small primary SP-frame.
- **Cons**:
  - very large secondary SP-frames.
  - As many secondary SP-frames as decoding paths.

M. Karczewicz and R. Kurceren, "**The SP- and SI-frames design for H.264/AVC**," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no.7, July 2003, pp. 637–644.

X. Sun, F. Wu, S. Li, G. Shen, and W. Gao, "**Drift-free switching of compressed video bitstreams at predictive frames**," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no.5, May 2006, pp. 565–576.

Merge Frame for Media Navigation: previous works 2

- **DSC frames**:
  - ***Key Idea***: treat merging as *noise removal*.
  - Divide **side information** (SI) frames into block, perform DCT, quantization.
  - Examine *bit-planes* of quantized coefficients.
    - If bit-planes different from target, ***channel coding*** to "denoise" SI bit-planes to target bit-planes.



- **Pros**: one merge frame for many decoding paths.
- **Cons**:
  - Bit-plane / channel coding are complex.
  - Channel coding works well only for *average statistics*.

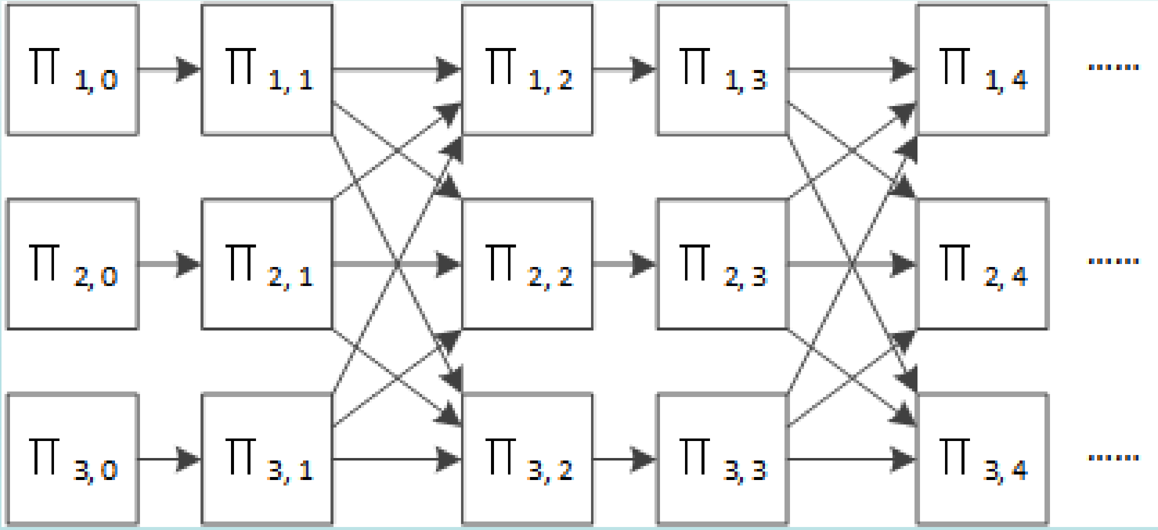
P. Ramanathan, M. Kalman, and B. Girod, "**Rate-distortion optimized interactive light field streaming**," in *IEEE Transactions on Multimedia*, vol. 9, no.4, June 2007, pp. 813–825.

N.-M. Cheung, A. Ortega, and G. Cheung, "**Distributed source coding techniques for interactive multiview video streaming**," in *27th Picture Coding Symposium*, Chicago, IL, May 2009.

Merge Frame for Media Navigation: definition

- **Interactive Video Stream Switching (IVSS)**
  - Multiple ***related*** pre-encoded video streams.
  - Designated ***switching points*** to switch from one to another.

- Picture Interactive Graph
  - Dynamic View Switching: switch to neighboring view of next time instant.
  - No loops in PIG.
  - *Optimized target merging*.



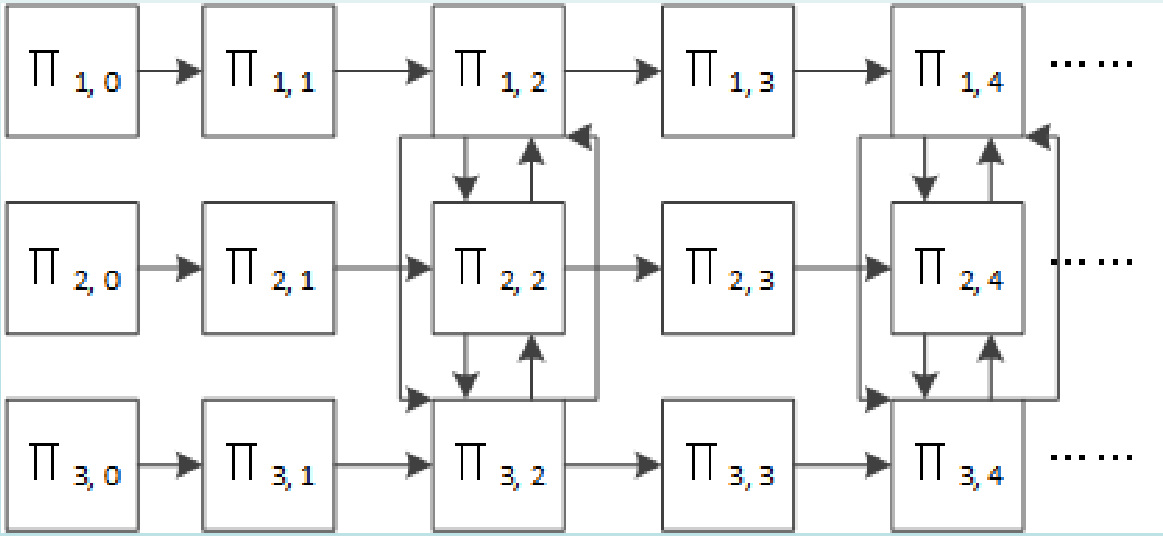
W. Dai, G. Cheung, N.-M. Cheung, A. Ortega, O. Au, "**Rate-distortion Optimized Merge Frame using Piecewise Constant Functions**," *IEEE Int'l Conf. on Image Processing*, Melbourne, Australia, September, 2013. (**Best student paper award**)

Wei Dai, Gene Cheung, Ngai-Man Cheung, Antonio Ortega, Oscar Au, "**Merge Frame Design for Video Stream Switching using Piecewise Constant Functions**," *IEEE Transactions on Image Processing*, vol. 25, no.8, August 2016

Merge Frame for Media Navigation: definition

- **Interactive Video Stream Switching (IVSS)**
  - Multiple ***related*** pre-encoded video streams.
  - Designated ***switching points*** to switch from one to another.

- Picture Interactive Graph
  - Static View Switching: switch to neighboring view of same time instant.
  - **Loops** in PIG.
  - *Fixed target merging*.

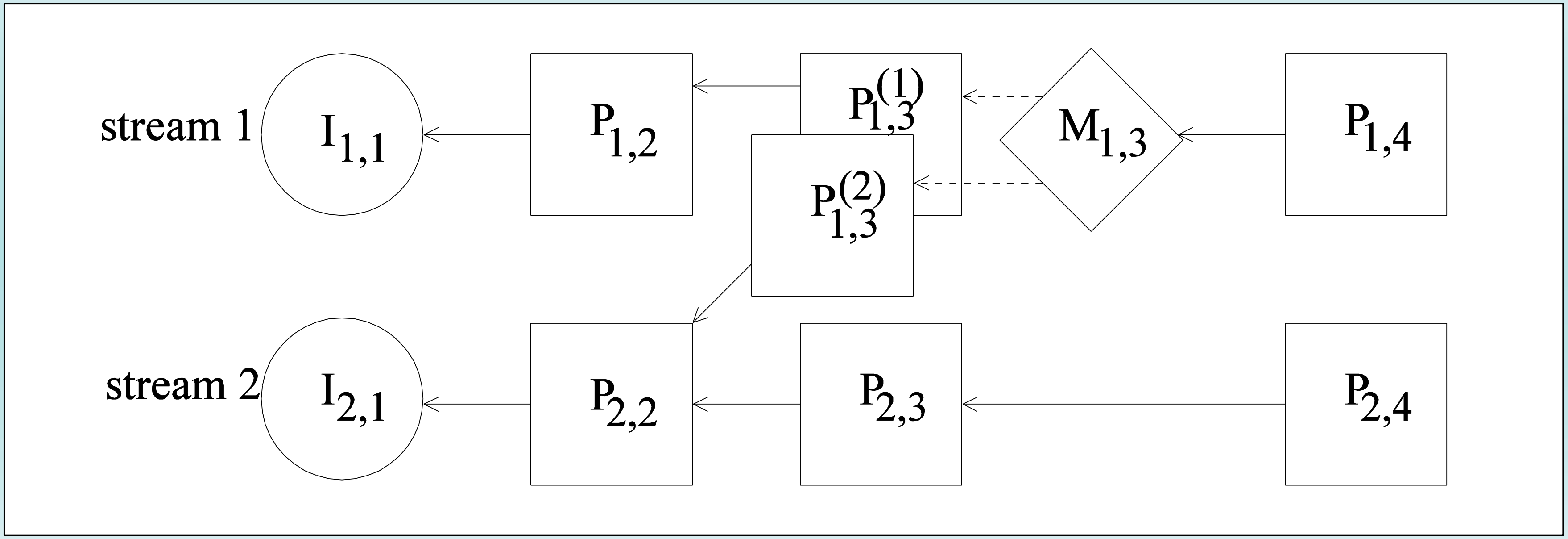


J.-G. Lou, H. Cai, and J. Li, "**A real-time interactive multi-view video system**," in *ACM International Conference on Multimedia*, Singapore, November 2005.

N.-M. Cheung and A. Ortega, "**Compression algorithms for flexible video decoding**," in *IS&T/SPIE Visual Communications and Image Processing (VCIP'08)*, San Jose, CA, January 2008.

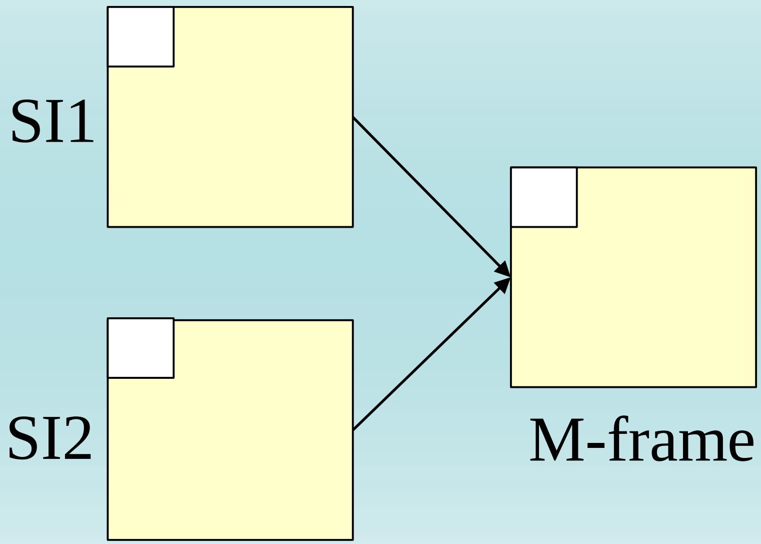
Merge Frame for Media Navigation: framework

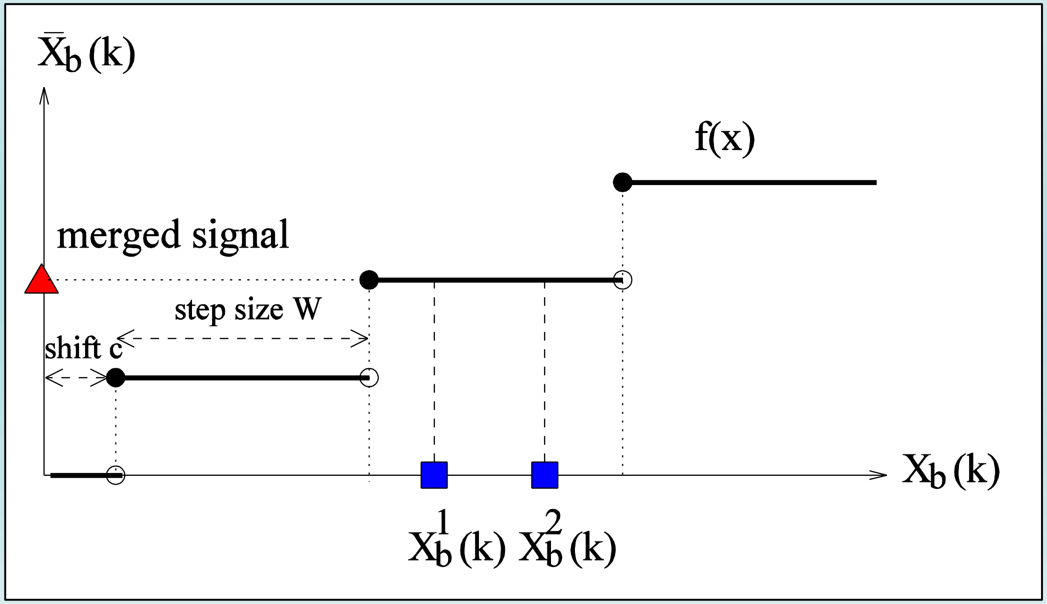
- **Switching Mechanism**
  - **Side Information (SI) frame**: P-frame predicted from diff. streams.
  - **Merge frame:** merge diff. among SI frames into same frame.
  - **Interactive Transmission:** transmit one SI frame + merge frame according to chosen decoding path.



Merge Frame for Media Navigation: merge frame (M-frame) overview

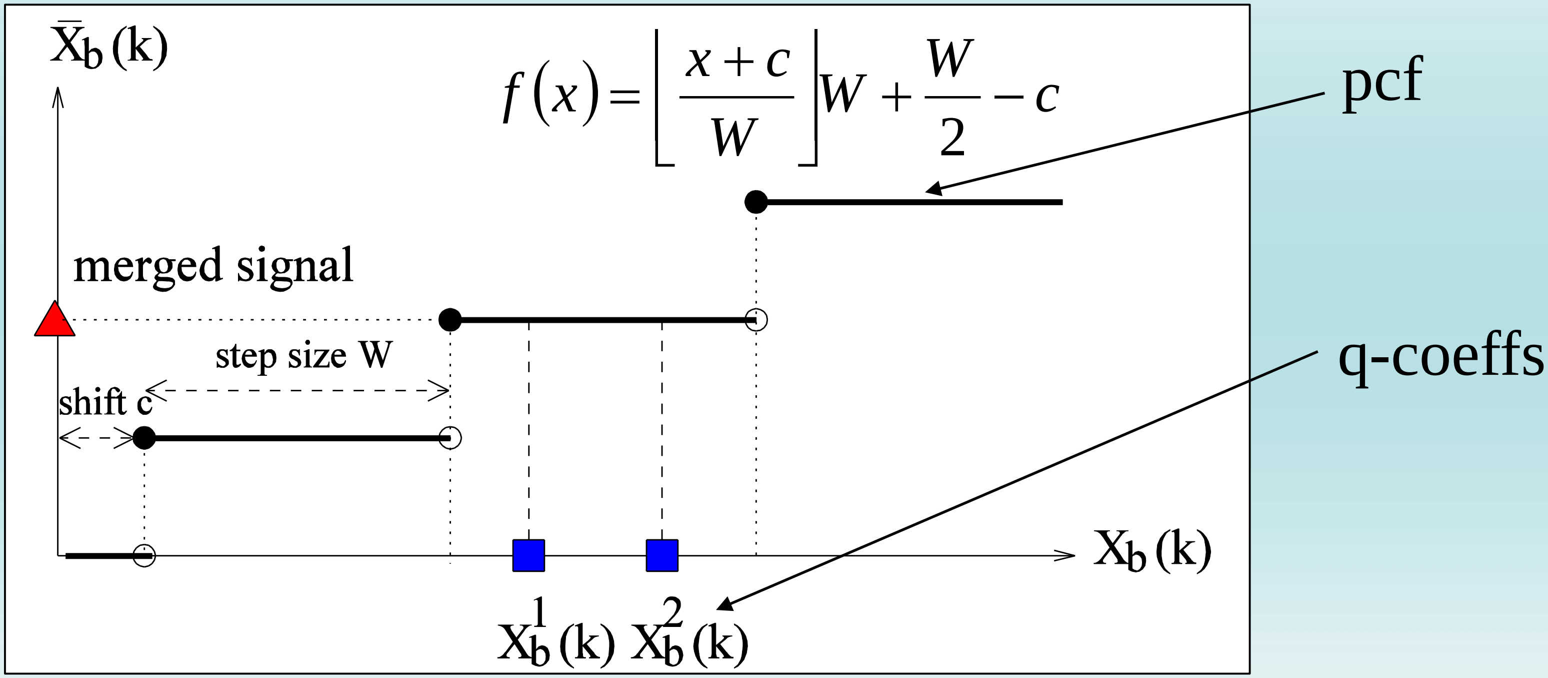
1. Each decoded SI frame is divided into 8x8 blocks, DCT transform and coefficient quantized (**q-coeff**).
2. Given block *b*, if q-coeffs of SI frames very different, use *I*-block.
3. If q-coeffs of SI frames the same, use *skip* block.
4. If q-coeffs of SI frames slightly different, use ***merge block***.





Merge Frame for Media Navigation: merge block overview

- Use ***piecewise constant function*** (pcf) for merging of SI's q-coeffs:
  - Q-coeff's must land on the same "step" for **identical merging**.
- pcf defined by step size *W* and shift c:
  - Choose *W* per frequency of all merge blocks (**cheap**).
  - Choose c per block per frequency (**expensive**).

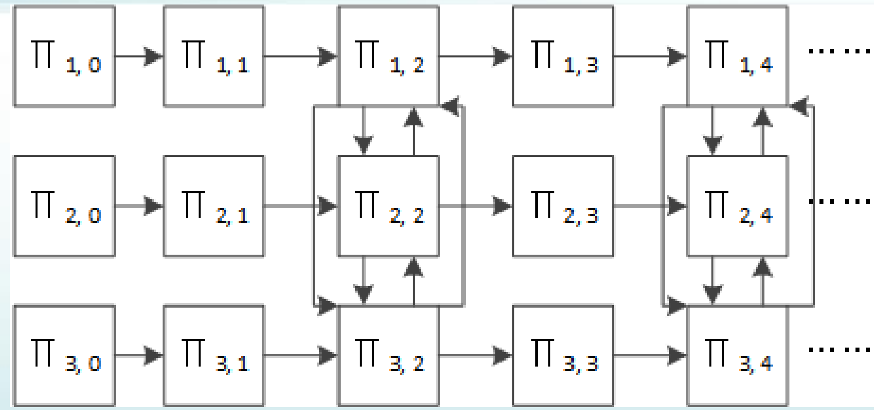


Merge Frame for Media Navigation: 2 merging problems

Fixed Target Merging:

- Find M-frame M to reconstruct any SI frame S^n, $n=1,\ldots,N$, <u>*identically*</u> to a **fixed target** T.
- Difficult to optimize M-frame parameters.

Static view switching



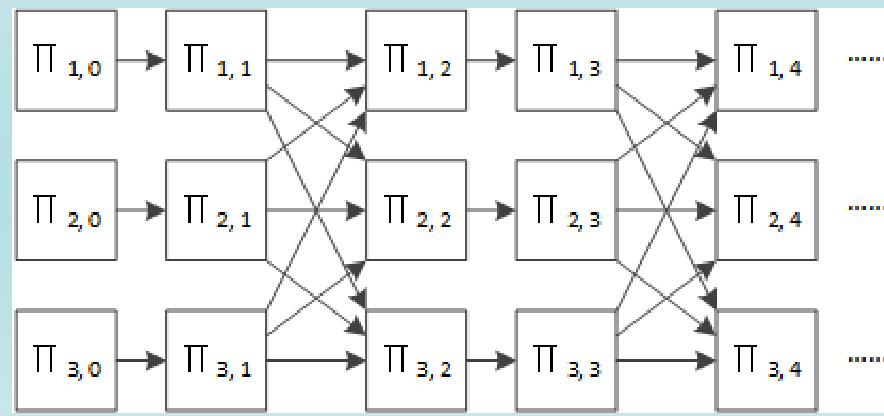
Optimized Target Merging:

- Find M-frame M to reconstruct any SI frame S^n, $n=1,\ldots,N$, <u>*identically*</u> to a **floating target** $\overline{\mathrm{T}}(\mathrm{M})$, such that:

$$\mathrm{M}^* = \arg\min_{\mathrm{M}} D\left(\mathrm{T}, \overline{\mathrm{T}}(\mathrm{M})\right) + \lambda R(\mathrm{M})$$

- Optimize M-frame parameters in RD manner.

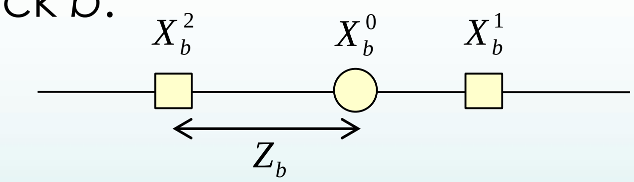
Dynamic view switching



Merge Frame for Media Navigation: step W, shift c (fixed target merging)

- **Choosing step size *W* for given freq *k***:
  - Compute **max diff.** from **target q-coeff** in each block b:

$$Z_b = \max_{n \in \{1,\ldots,N\}} \left| X_b^0 - X_b^n \right|$$



  - Choose step size W to be roughly 2 * **max diff**:

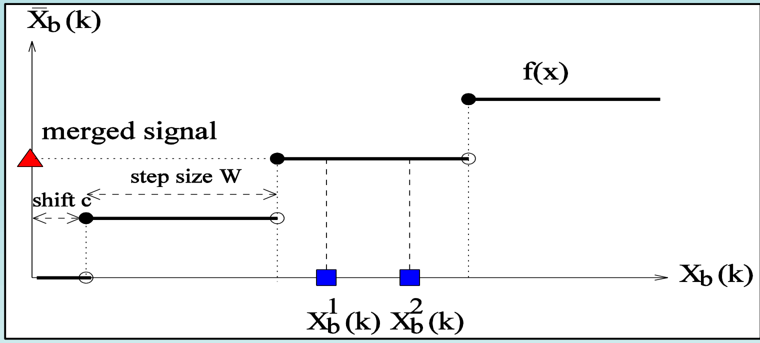
$$W_b^{\#} = 2Z_b + 2$$

pcf:

$$f(x) = \left\lfloor \frac{x+c}{W} \right\rfloor W + \frac{W}{2} - c$$

- **Choosing shift *c* for each block *b***:
  - Choose shift: $c_b = W_b^{\#}/2 - X_{b,2}^0$, where $X_{b,2}^0 = X_b^0 \bmod W_b^{\#}$
  - **Lemma V.1**: given this choice of step and shift,

$$f\left(X_b^n\right) = X_b^0, \quad \forall n \in \{0,\ldots,N\}$$



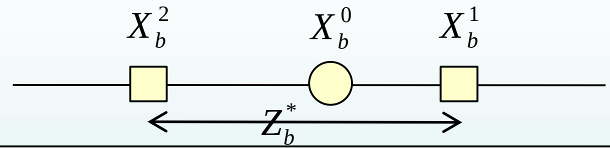
- **Merge block group B_m**, use a bigger step:

$$Z_{B_M} = \max_{b \in B_M} Z_b \qquad W_{B_m}^{\#} = 2Z_{B_m} + 2 \qquad X_{b,2}^0 = X_b^0 \bmod W_{B_m}^{\#}$$

Merge Frame for Media Navigation: step W, shift c (optimized target merging)

- **Choosing step size *W* for given freq *k***:
  - Compute **max diff**. bet'n 2 q-coeffs in block b, then block-wise max diff.:

$$Z_b^* = \max_{i,j \in \{0,\ldots,N\}} X_b^i - X_b^j \qquad Z_{B_M}^* = \max_{b \in B_M} Z_b^*$$

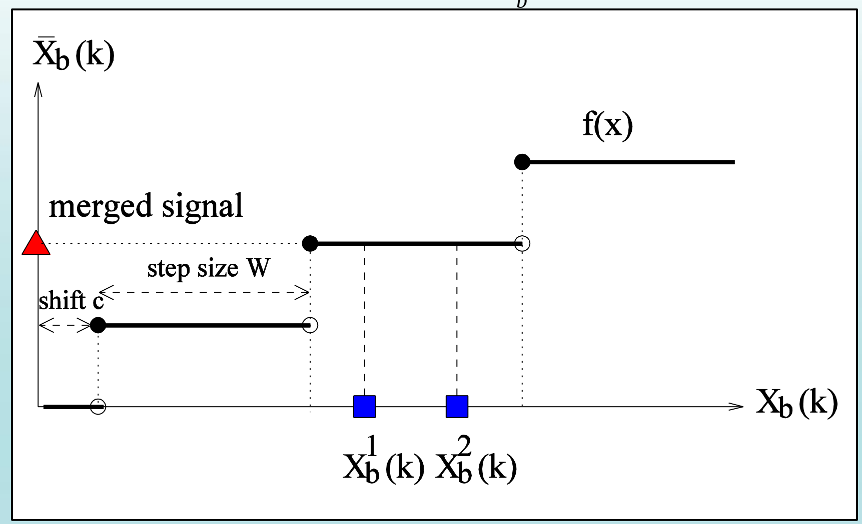


  - Choose step size W to be roughly **max diff**:

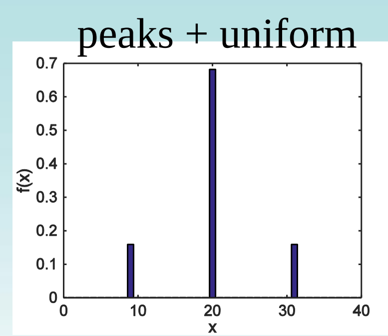
$$W_{B_M} = Z_{B_M}^* + 1$$

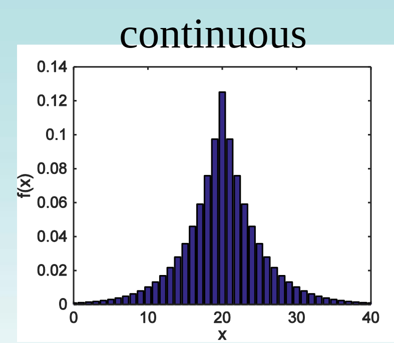
- **Choosing shift *c* for each block *b***:
  - Given step W, ***range*** $\boldsymbol{F_b}$ of shifts c can lead to *identical merging*.
  - Choose c in $\boldsymbol{F_b}$ to min RD cost:

$$\min_{0 \le c_b \le W_{B_M} \mid c_b \in F_b} d_b + \lambda\left(-\log P(c_b)\right)$$

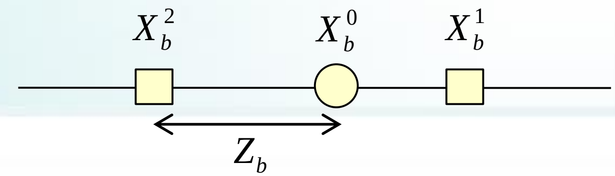


- **Initialize $P(c_b)$**:
  - Initialize a "peaks + uniform" distribution.
  - Rate-constrained LM till convergence.





Comparison with Coset Coding



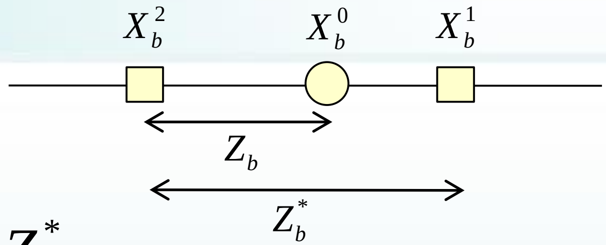
- **Coset Coding**:
  - SI values X_b^n are noisy observations of target X_b^0
  - Compute first largest difference w.r.t. to target:

$$Z_b = \max_n \left| X_b^n - X_b^0 \right|$$

  - **Encoder**: select **coset size** $W > 2Z_b$, transmit **coset index** $i_b = X_b^0 \bmod W$
  - **Decoder**: compute $\hat{X}_b = \arg\min_{X \in Z} \left| X_b^n - X \right| \quad s.t. \quad i_b = X \bmod W$

- **Fixed Target Merging**:
  - Step W is roughly $2Z_b$: $W_b^{\#} = 2Z_b + 2$
  - Shift c given W is remainder of target: $c_b = W_b^{\#} / 2 - X_{b,2}^0$, where $X_{b,2}^0 = X_b^0 \bmod W_b^{\#}$
  - <u>Expect the same coding rate as coset coding!</u>

Comparison with Coset Coding



- **Optimized Target Merging**:
  - Step W is roughly Z_b : $W_b = Z_b^* + 1$, where $Z_b \le Z_b^*$
  - Compared to **coset size** $W > 2Z_b$, nearly half the step size!
  - Feasible range of shifts to select from via RD optimization:

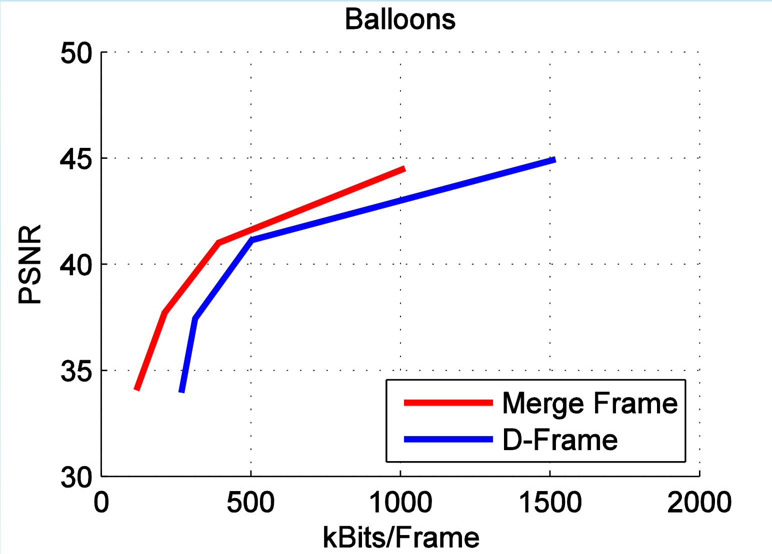
$$\min_{0 \le c_b \le W_{B_M} | c_b \in F_b} d_b + \lambda\left(-\log P(c_b)\right)$$

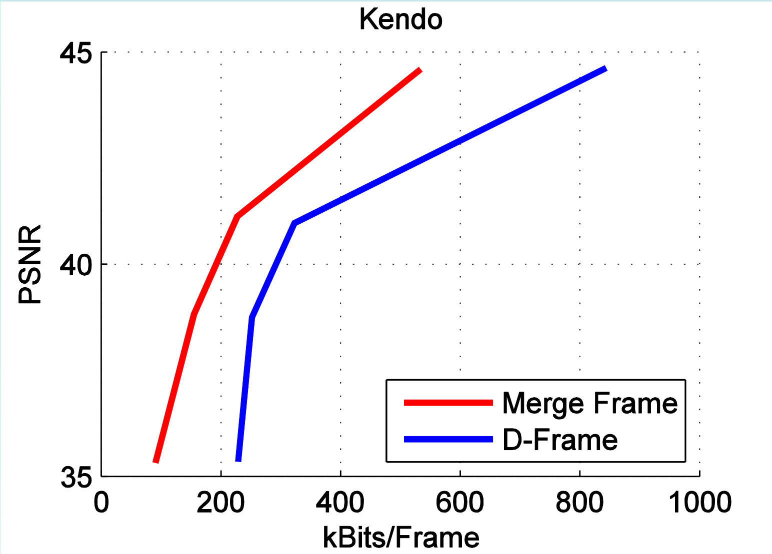
  - Expect significant coding gain, especially at low rates.

Merge Frame for Media Navigation: experiments

- **Exp Setup**: Static view switching
  - **Fixed target merging**: 3 views with the same QP.
  - H.264 for I- and P-frames.
  - Compared w/ DSC frames.

Sequence Name	M-frame *vs.* D-frame
Balloons	-31.7%
Kendo	-40.1%
Lovebird1	-35.7%
Newspaper	-31.1%

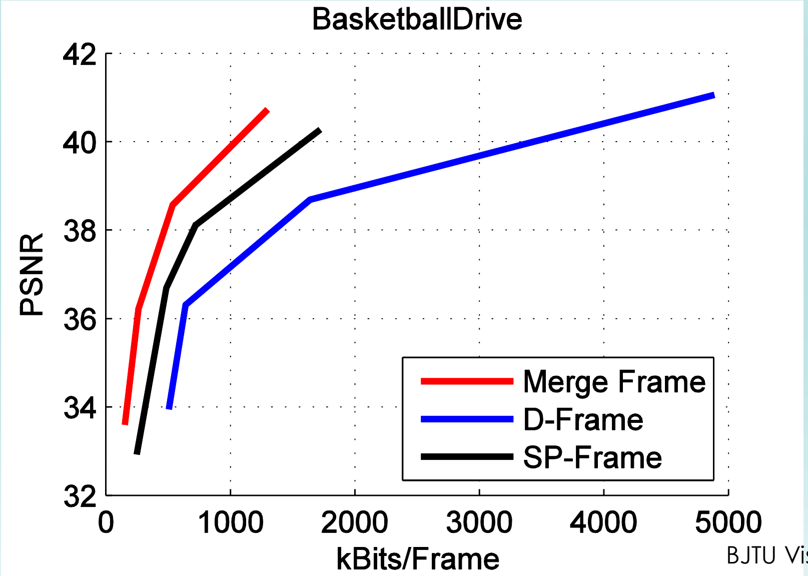


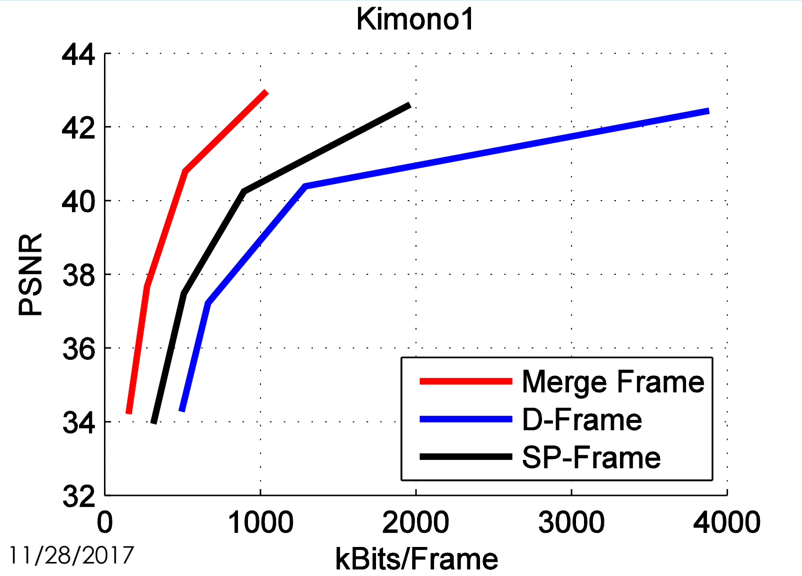


Merge Frame for Media Navigation: experiments 2

- **Exp Setup**: Bit-rate adaptation
  - **Optimized target merging**: 3 streams of same sequence at diff. rates (TFRC).
  - H.264 for I- and P-frames.
  - vs. DSC frames, SP-frames.
  - Worst case plots.

Sequence Name	M-frame *vs.* D-frame		M-frame *vs.* SP-frame	
	Average Case	Worst Case	Average Case	Worst Case
BasketballDrive	-63.4%	-63.7%	-17.0%	-39.4%
Cactus	-63.5%	-63.2%	-18.8%	-42.1%
Kimono1	-65.6%	-65.4%	-36.3%	-49.9%
ParkScene	-56.3%	-56.7%	-19.5%	-43.8%

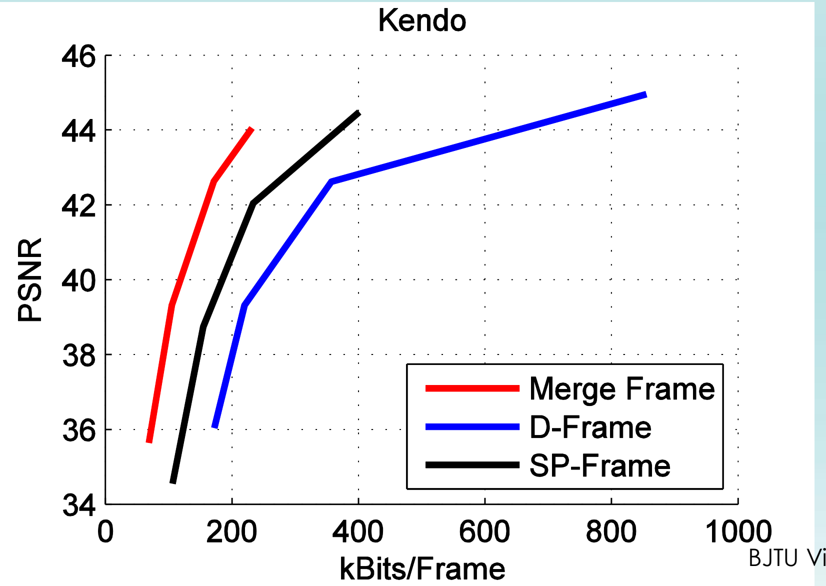


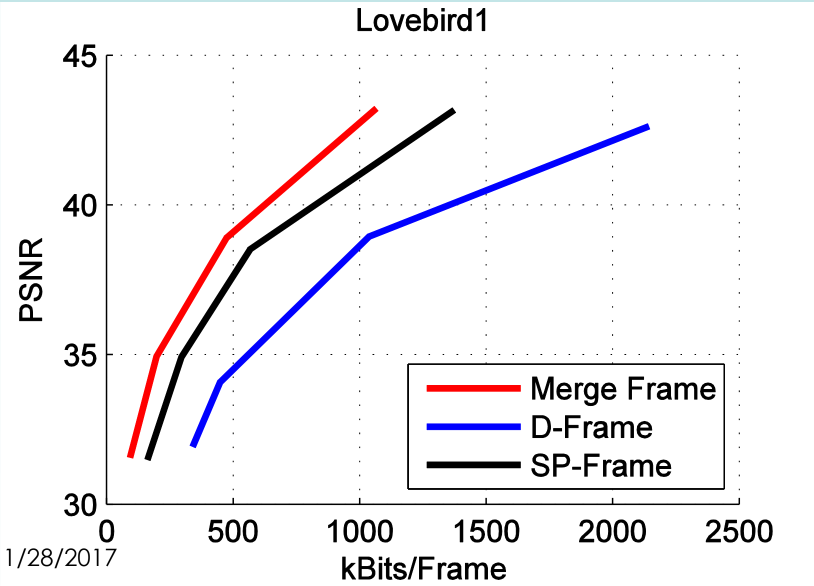


Merge Frame for Media Navigation: experiments 3

- **Exp Setup**: Dynamic view switching
  - **Optimized target merging**: 3 views with the same QP.
  - H.264 for I- and P-frames.
  - vs. DSC frames, SP-frames.
  - Worst case plots.

Sequence Name	M-frame *vs.* D-frame		M-frame *vs.* SP-frame	
	Average Case	Worst Case	Average Case	Worst Case
Balloons	-63.4%	-63.7%	-17.0%	-39.4%
Kendo	-63.5%	-63.2%	-18.8%	-42.1%
Lovebird1	-65.6%	-65.4%	-36.3%	-49.9%
Newspaper	-56.3%	-56.7%	-19.5%	-43.8%



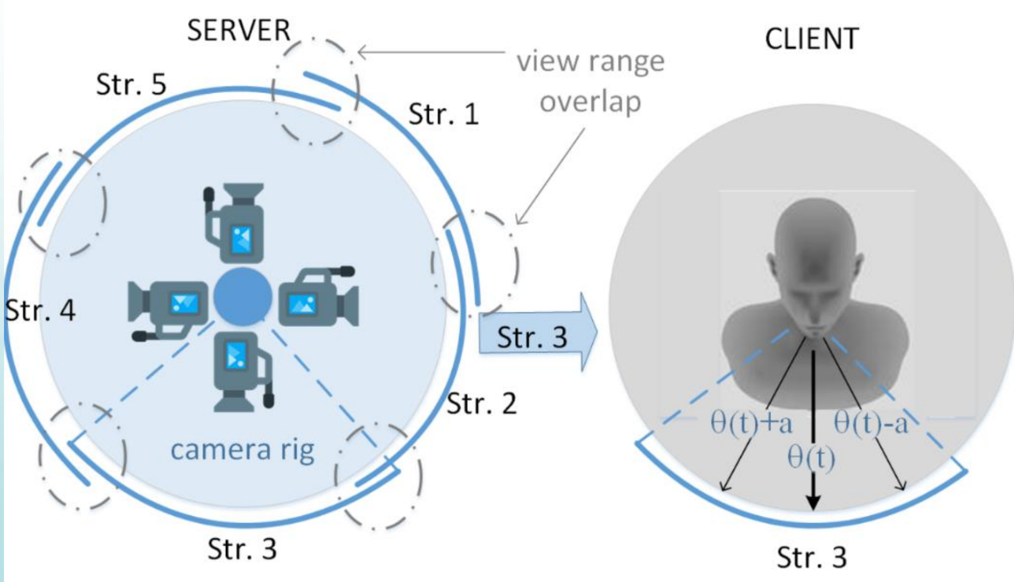


Outline

- What is interactive media navigation?
    - e.g. Multiview / free-viewpoint video
- Merge frame for interactive media navigation
    - Previous works
    - Merge frame / block overview
    - Fixed target merging
    - Optimized target merging
- Interactive Virtual Reality Video Streaming

Interactive Virtual Reality Video Streaming

- **Virtual reality** (VR): immersive 360 video w/ headsets.
- Diff. ***fields-of-view*** (FoV) rendered on headset, as user's head rotates L / R.
- Transmit only FoV:
  - reduces BW, but
  - results in stream-switch delay due to server-to-client RTT.

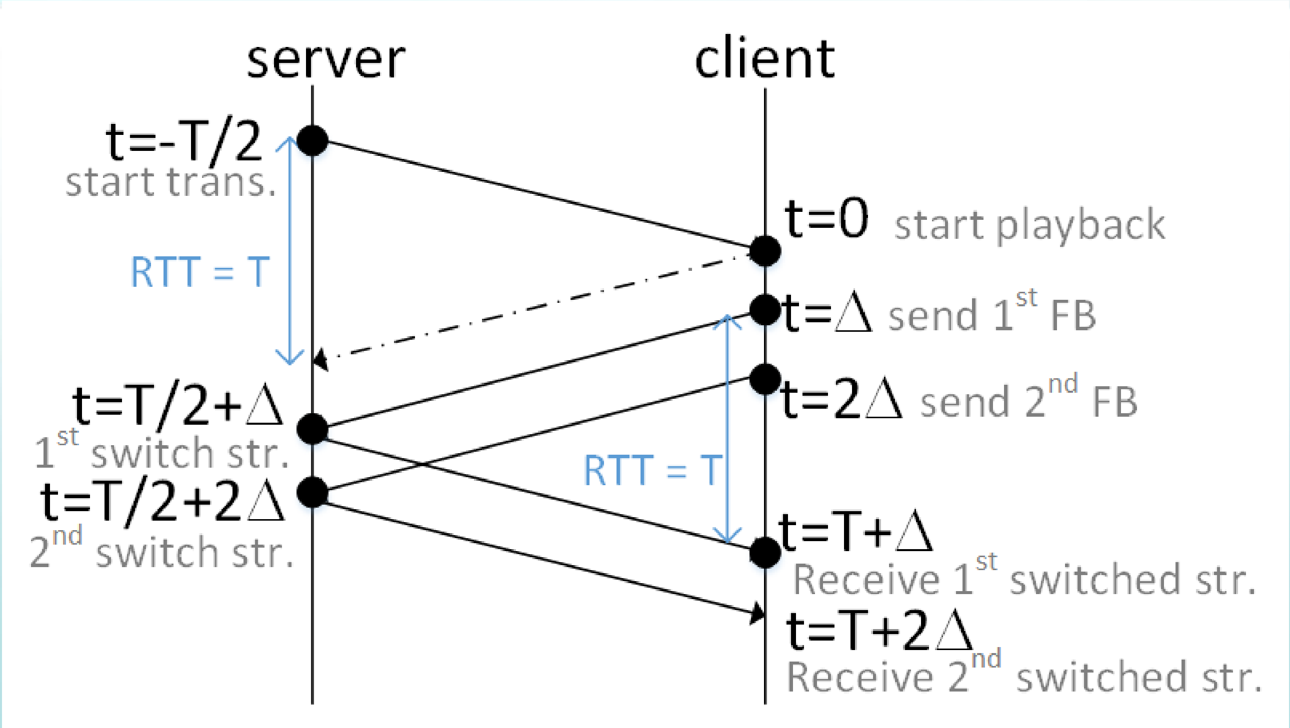


Research problem:

Design redundant video streams covering diff. viewing ranges, accounting for RTT EXPLICITLY, given storage and network constraints,

Gene Cheung, Zhi Liu, Zhiyou Ma, Jack Z. G. Tan, "**Multi-Stream Switching for Interactive Virtual Reality Video Streaming**," *IEEE International Conference on Image Processing*, Beijing, China, September, 2017.

Round-trip Time Interactive Delay



Sever / Client Interaction Model:

- **Client**: transmits head coordinate θ per frame.
- **Server**: transmits corresponding video stream f(θ).

Redundant Frame Structure Design

Objective Function:

{di} are distortion vectors for streams i's

mapping function from angle to stream

total number of viewing angles

steady state prob for angle *k*

angle transition matrix

RTT in number of video frames

$$D(\{\mathbf{d}_i\}, f()) = \sum_{k=1}^{K} q_k \, \mathbf{1}_k \mathbf{C}_a \mathbf{P}^{T_s} \mathbf{d}_{f(k)}$$

canonical row vector for angle k

distortion vector for diff. angles for stream *f(k)*

binary circulant matrix to account for FoV

Redundant Frame Structure Design

Constraints:

encoding rate of stream *i* given distortion vector di

storage budget

- Storage constraint:

$$\sum_{i \in \mathcal{S}} r(\mathbf{d}_i) \leq B/Q$$

video length

channel bandwidth

- Bandwidth constraint:

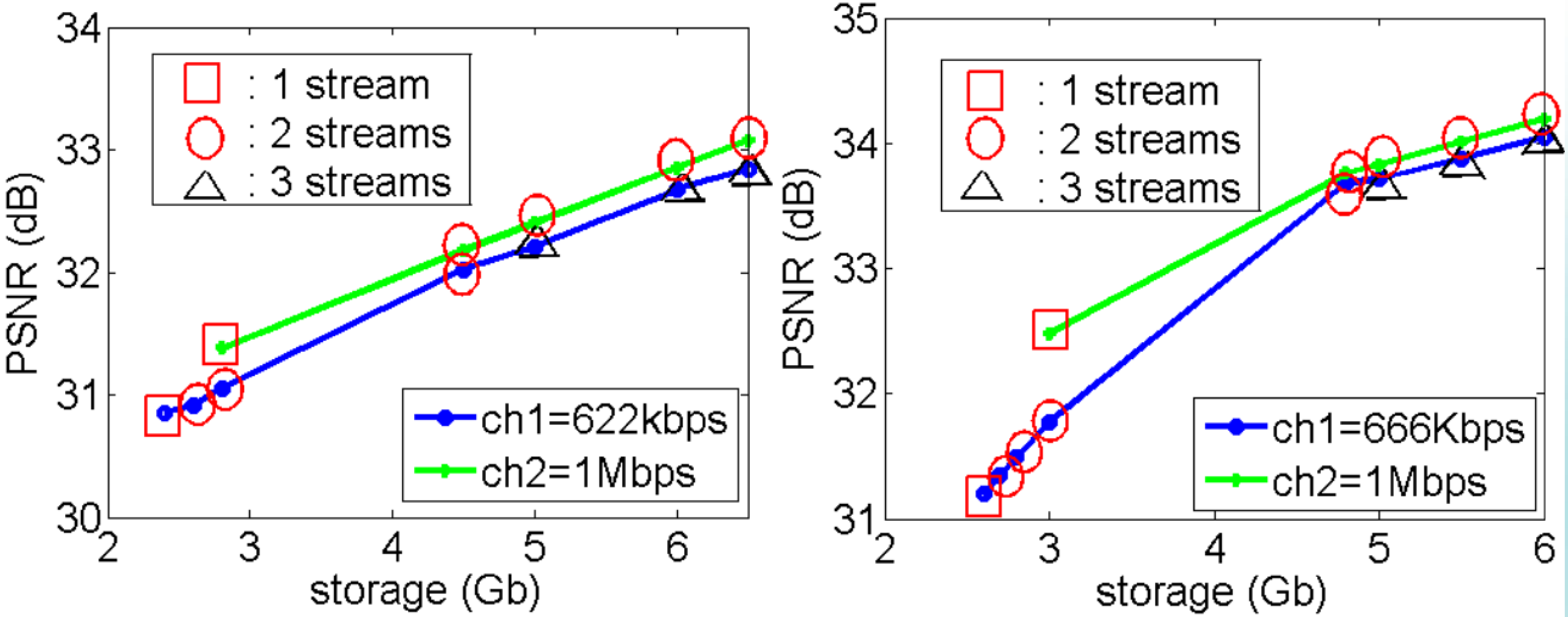
$$\sum_{k=1}^{K} q_k r(\mathbf{d}_{f(k)}) \leq C$$

Experimental Setup

- 2 video sequences: indoor concert and outdoor walking
- FoV is 90◦, maximum switch each time is 5◦
- FoV resolution is 512×512, two quantization parameters used, frame rate is 30fps
- #discrete view angles K=60, #switch during each T is 3
- **View popularity**: transition probabilities from i to j linearly decreases with |i − j|, the slope of decrease is steeper at π/2 and 3π/2
- **Comparison scheme**: ‘static’, a non-switching scheme, which always sends an encoded video covering the entire 360 angles

Simulation Results

- outperforms ‘static’ by up to 2.9dB in PSNR



(a) indoor concert (b) outdoor walking

Fig. 4. PSNR versus storage for two competing schemes.

Summary

- Interactive media navigation
  - Difficult to achieve to good compression efficiency & flexible decoding.
- Merge frame to facilitate interactive navigation
  - Fixed target merging
  - Optimized target merging
- Interactive virtual reality video streaming
  - Redundancy to overcome stream-switch delay due to RTT

Q&A

- Email: cheung@nii.ac.jp
- Homepage: http://research.nii.ac.jp/~cheung/