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

- **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**: APSIPA 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


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.

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.

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?
• **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.


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.

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.*


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.**


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.

![Diagram showing SI1 and SI2 blocks merging into M-frame](image)
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**).

\[ f(x) = \left\lfloor \frac{x + c}{W} \right\rfloor W + \frac{W}{2} - c \]
Merge Frame for Media Navigation: 2 merging problems

**Fixed Target Merging:**
- Find M-frame $M$ to reconstruct any SI frame $S^n$, $n=1,\ldots,N$, identically to a fixed target $T$.
- Difficult to optimize M-frame parameters.

**Optimized Target Merging:**
- Find M-frame $M$ to reconstruct any SI frame $S^n$, $n=1,\ldots,N$, identically to a floating target $\overline{T}(M)$, such that:
  \[
  M^* = \arg\min_M D(T, \overline{T}(M)) + \lambda R(M)
  \]
- Optimize M-frame parameters in RD manner.
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 $b$:
    \[
    Z_b = \max_{n \in \{1,...,N\}} |X_b^0 - X_b^n|
    \]
  - Choose step size $W$ to be roughly $2 \times \text{max diff.}$:
    \[
    W_b^# = 2Z_b + 2
    \]

- **Choosing shift $c$ for each block $b$:**
  - Choose shift: $c_b = W_b^# / 2 - X_{b,2}^0$, where $X_{b,2}^0 = X_b^0 \mod W_b^#$
  - **Lemma V.1:** given this choice of step and shift,
    \[
    f(X_b^n) = X_b^0, \quad \forall n \in \{0,\ldots,N\}
    \]
  - **Merge block group $B_m$:** use a bigger step:
    \[
    Z_{B_M} = \max_{b \in B_M} Z_b \quad W_{B_m}^# = 2Z_{B_m} + 2 \quad X_{b,2}^0 = X_b^0 \mod W_{B_m}^#
    \]

\[\text{pcf:} \quad f(x) = \left[ \frac{x + c}{W} \right]W + \frac{W}{2} - c \]
Merge Frame for Media Navigation:
step $W$, shift $c$ (optimized target merging)

- **Choosing step size $W$ for given freq $k$:**
  - Compute $\text{max diff. bet'n 2 q-coeffs}$ in block $b$, then block-wise max diff.:
    \[ Z_b^* = \max_{i,j \in \{0,...,N\}} X_b^i - X_b^j \quad Z_{BM}^* = \max_{b \in B_M} Z_{b}^* \]
  - Choose step size $W$ to be roughly $\text{max diff}$:
    \[ W_{BM} = Z_{BM}^* + 1 \]

- **Choosing shift $c$ for each block $b$:**
  - Given step $W$, range $F_b$ of shifts $c$ can lead to identical merging.
  - Choose $c$ in $F_b$ to min RD cost:
    \[ \min_{0 \leq c_b \leq W_{BM}} d_b + \lambda( - \log P(c_b) ) \]

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

• **Fixed Target Merging:**
  • Step $W$ is roughly $2Z_b$: $W_b^# = 2Z_b + 2$
  • Shift $c$ given $W$ is remainder of target: $c_b = W_b^# / 2 - X_b^{0,2}$, where $X_b^{0,2} = X_b^0 \mod W_b^#$
  • Expect the same coding rate as coset coding!
Comparison with Coset Coding

- **Optimized Target Merging:**
  - Step $W$ is roughly $Z_b$: $W_b = Z_b^* + 1$, where $Z_b \leq Z_b^*$
  - Compared to **coset size** $W > 2Z_b$, nearly half the step size!
  - Feasible range of shifts to select from via RD optimization:
    
    $$
    \min_{0 \leq c_b \leq W_{BM} \mid c_b \in F_b} d_b + \lambda(-\log P(c_b))
    $$
    
  - 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.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>M-frame vs. D-frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balloons</td>
<td>-31.7%</td>
</tr>
<tr>
<td>Kendo</td>
<td>-40.1%</td>
</tr>
<tr>
<td>Lovebird1</td>
<td>-35.7%</td>
</tr>
<tr>
<td>Newspaper</td>
<td>-31.1%</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>M-frame vs. D-frame</th>
<th>M-frame vs. SP-frame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Case</td>
<td>Worst Case</td>
</tr>
<tr>
<td>BasketballDrive</td>
<td>-63.4%</td>
<td>-63.7%</td>
</tr>
<tr>
<td>Cactus</td>
<td>-63.5%</td>
<td>-63.2%</td>
</tr>
<tr>
<td>Kimono1</td>
<td>-65.6%</td>
<td>-65.4%</td>
</tr>
<tr>
<td>ParkScene</td>
<td>-56.3%</td>
<td>-56.7%</td>
</tr>
</tbody>
</table>

PSNR plots for BasketballDrive and Kimono1.
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.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>M-frame vs. D-frame</th>
<th>M-frame vs. SP-frame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average Case</td>
<td>Worst Case</td>
</tr>
<tr>
<td>Balloons</td>
<td>-63.4%</td>
<td>-63.7%</td>
</tr>
<tr>
<td>Kendo</td>
<td>-63.5%</td>
<td>-63.2%</td>
</tr>
<tr>
<td>Lovebird1</td>
<td>-65.6%</td>
<td>-65.4%</td>
</tr>
<tr>
<td>Newspaper</td>
<td>-56.3%</td>
<td>-56.7%</td>
</tr>
</tbody>
</table>

![Kendo Graph](image1.png)

![Lovebird1 Graph](image2.png)
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,

Round-trip Time Interactive Delay

Server / Client Interaction Model:

- **Client**: transmits head coordinate $\theta$ per frame.
- **Server**: transmits corresponding video stream $f(\theta)$.
Redundant Frame Structure Design

Objective Function:

\[ D(\{d_i\}, f()) = \sum_{k=1}^{K} q_k s_k C_k P^{T_s} d_{f(k)} \]

- \(\{d_i\}\) are distortion vectors for streams i’s
- mapping function from angle to stream
- total number of viewing angles
- steady state prob for angle \(k\)
- canonical row vector for angle \(k\)
- RTT in number of video frames
- angle transition matrix
- distortion vector for diff. angles for stream \(f(k)\)
- binary circulant matrix to account for FoV
Redundant Frame Structure Design

Constraints:

• Storage constraint:

\[ \sum_{i \in S} r(d_i) \leq B/Q \]

- Encoding rate of stream \(i\)
- Given distortion vector \(d_i\)
- Storage budget

• Bandwidth constraint:

\[ \sum_{k=1}^{K} q_k r(d_{f(k)}) \leq C \]

- Video length
- Channel bandwidth
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

![PSNR versus storage graphs for two competing schemes](image)

(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/