New Trend and Challenges in 3D Video Coding

3-D 錄像編碼之趨勢及挑戰

Dr. Y.L. Chan, Assistant Professor
Professor W.C. Siu, Chair Professor and Director
Centre for Signal Processing
Department of Electronic and Information Engineering
Hong Kong Polytechnic University

With Contributions from:
Chang-Hong Fu (Dr.) Calvin Cheung(Dr.) Hui Wai Lam
Zhi-Pin Deng Glory Lee Yan-Ho Kam
Outline

- Introduction – New Trends in Video Coding
- Stereoscopic Video Coding
- Multi-view Video Coding standardization in MPEG and JVT
- New Development of 3D Video/Free-viewpoint Video
- Conclusions
Video Technology – Past, Present and Future

**Past**
- B&W TV
- Color TV

**Future**
- Higher frame rate: 60Hz → 200Hz (Movie with high motion)
- Larger resolution: HD → Ultra HD (4K or 8K)
- More views or 3D perception

**Temporal Dimension**
- Higher frame rate

**Spatial Dimension**
- Larger resolution

**View Dimension**
- More views or 3D perception

---

Y.L. Chan, Centre for Signal Processing, POLYU
Video Coding Technologies and Standards

Explore better algorithms for video with higher frame rate and resolution.

H.265 and HVC

Coding Technologies for video with more views.

MVC → 3DV/FTV

Source: http://computer.howstuffworks.com/
Basic H.264 Coding Structure

Input Video Signal

Split into Macroblocks 16x16 pixels

Coder Control

Transform/Scal./Quant.

Decoder

Scaling & Inv. Transform

De-blocking Filter

Intra-frame Prediction

Motion-Compensation

Motion Estimation

Control Data

Quant. Transf. coeffs

Entropy Coding

Output Video Signal

Motion Data

Intra/Inter

Y.L. Chan, Centre for Signal Processing, POLYU
“Conventional” Stereo requires only **two views (L & R)**

Standard format for 3D Cinema and consumer displays

Only color pixel data from two views are coded, but **NO scene geometry information.**
Stereo Video Coding (Simulcast)

- **L/R simulcast**: possible for any MPEG/H.264 standards

  - Bandwidth required to transmit both stereoscopic image streams is doubled
  - ⇒ it is essential to reduce data rate.

Y.L. Chan, Centre for Signal Processing, POLYU
Stereo Video Coding (SEI Message)

Combination of two views into one

- Stereoscopic MAF (Multimedia application format) ISO/IEC 23000-11 based on MPEG-4 part 2 video (L/R packing)
- AVC Stereo Frame Packing Arrangement Supplemental Enhancement Information (SEI) message
  - Allow various methods of L/R packing.
  - In July 2009, Frame Packing Arrangement SEI messages were added to MPEG-4 AVC.
  - The Frame Packing Arrangement SEI message tells the decoder that the left- and right-eye stereo views are packed into a single high-resolution video frame either in a top-to-bottom, side-by-side, checkerboard, or other arrangement.
  - Packing both left- and right-eye stereo views into a single video frame makes it possible to use existing encoders and decoders to distribute 3DTV immediately without having to wait for MVC & Stereo High Profile hardware to be deployed widely.
Example

AVC Stereo Frame Packing arrangement SEI message

L/R Packing

H.264/AVC Encoder Field Coding

Network

H.264/AVC Decoder Field Decoding

L/R Depacking

Y.L. Chan, Centre for Signal Processing, POLYU
Stereo Video Coding (Stereo High Profile)

Exploitation of inter-view redundancy

Each macroblock picks the best reference using R-D optimization.
Block-based Disparity Estimation

Following the method for block-based motion estimation
- Divide the anchor image into regular blocks.
- The disparity function over each small block of the reference view can be described by a constant vector.
- The constant disparity vector is determined by minimizing the error.
- Either exhaustive or other fast search methods are used for block-based motion estimation can be applied.

Difference from motion estimation
- Constant disparity model is less effective than constant motion model even over small blocks.
- Search range should be much larger, if the depth variation in the imaged scene is relatively large or the baseline separation between cameras is large.
Stereo Video Coding (Stereo High Profile)

- Stereo High Profile does not support
  - N-view displays
  - **Autostereoscopy**: a method of displaying three-dimensional images that can be viewed without the use of special headgear or glasses on the part of the user

For these purposes, either **coding of multiple views** (if available) or **depth-based synthesis** is needed.
Multi-view Video (MVV)

- **MVV representations**: require multiple synchronized video signals that show the same scenery from different viewpoints.
- Compared to other 3D video formats, it is the **least complex**.
- **Huge amount of data**:
  - VGA color video: 8 views, 30 fps, 10 sec.
  - $1024 \times 768 \times 3$ bytes (R, G, and B) $\times$ 8 views $\times$ 30 fps $\times$ 10 sec. $= 5,662$ Gbytes
  - For 1 multi-view video, 8 CDs are required
- Typically, MVV has a larger amount of inter-view statistical dependencies than stereoscopic video.
MVC Standard

- MVC Standard was approved in July 2008
  - Specified as an amendment of H.264/MPEG-4 AVC

- MVC Profiles
  - Profiles determine the subset of coding tools
  - Multiview High
    - Supports same subset of coding tools for inter-view prediction as the existing High Profile of AVC (but no interlaced)
  - Stereo High
    - Includes support for interlaced and limits number of views to stereo only (2 views only)
Key Elements of MVC Standard

**Major Syntax Modification**
- No changes to lower-level AVC syntax (Slice and lower) ⇒ compatible and easily integrated with existing hardware
- Small backward compatible changes to high-level syntax ⇒ to specify view dependency and random access points
- Base layer required and easily extraction from video bit stream

**Enable Inter-view prediction to exploit inter-view redundancy**
- via flexible reference picture management
- Allow decoded pictures from other views to be inserted and removed from reference picture buffer
- Core decoding modules not aware of whether reference picture is a time reference or view reference
Temporal Coding Structure in Simulcast

Temporal prediction using hierarchical B pictures

Y.L. Chan, Centre for Signal Processing, POLYU
Inter-view Prediction for Key Pictures in MVC

Prediction across views to exploit inter-view redundancy

Y.L. Chan, Centre for Signal Processing, POLYU
Inter-view Prediction for ALL Pictures in MVC

Y.L. Chan, Centre for Signal Processing, POLYU
Performance in Coding Efficiency

PSNR Difference Relative to Simulcast

- Key picture prediction
- Prediction for all pictures

*majority of gains due to inter-view prediction at key pictures*
Summary: MVC Standard

- MVC standard was finalized
- Builds on the widely deployed AVC standard;
  - core encoding/decoding processed unchanged
- Offers the base view extraction: a compatible 2D representation from the 3D version

Base view extraction: Backward compatible to H.264/AVC
Challenges in MVV Coding

- Reduce computational complexity
  - Fast motion & disparity estimation
- Improve coding efficiency
  - MVC motion/disparity vector coding
  - Investigation on new modes according to temporal and view prediction structure
- Illumination Compensation (IC)
  - Incorporate illumination change into MC process
- Support other functionalities
  - New prediction structure with better coding efficiency and random access capability

...
Fast motion & disparity estimation

- Takes a huge amount of encoding time
- Speed-up process is the critical factor to make the implementation of MVC standard (JMVM codec) practical.
- Technologies
  - Joint motion and disparity estimation by using coherence condition between motion and disparity in MVC sequences
  - The use of new predictors, based on view/motion dependence, to reduce the size of search window
Prediction structure

Q At most 4 reference frames
Q In JMVM, an exhaustive ME & DE in all the reference frames is performed to search for the best result.
Search strategy - loop constraint

In JMVM, $mv_{r,t}$ and $dv_{r,t}$ are obtained independently with exhaustive full search. It’s time consuming.

**New Search strategy:** integrate ME & DE by making use of loop constraint to get the current $mv_{r,t}$ and $dv_{r,t}$ iteratively.
## Experimental results

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Algorithm in SPIE</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>BD-PSNR (dB)</strong></td>
<td><strong>BD-Bitrate (%)</strong></td>
</tr>
<tr>
<td>Ballroom</td>
<td>-0.21</td>
<td>+6.60</td>
</tr>
<tr>
<td>Exit</td>
<td>-0.07</td>
<td>+3.42</td>
</tr>
<tr>
<td>Vassar</td>
<td>-0.02</td>
<td>+1.04</td>
</tr>
<tr>
<td>Race1</td>
<td>-0.14</td>
<td>+3.46</td>
</tr>
<tr>
<td>Rena</td>
<td>-0.03</td>
<td>+0.71</td>
</tr>
</tbody>
</table>

- JMVM 8.0, Inter 16 × 16 mode, SR = ±96
- **BD measure**
  - Bjontegaard, G.: ‘Calculation of average PSNR differences between RDcurves’, ITU-T SG16 Q.6 VCEG- M33, Austin, TX, USA, April 2001
- **SPIE**
Limitations/Technical Problems in MVC Standard

- Acquisition/production with large camera arrays is not common and is somewhat difficult
  - Color consistency among multiple views.
  - Synchronization among multiple cameras.
- Although more efficient than simulcast, rate of MVC is still proportional to the number of views.
Limitations/ Technical Problems in MVC Standard

- The baseline is **fixed** from capturing
  - Depth perception cannot be adjusted to different types and sizes.
- The number of output views cannot be varied (only decreased).
- Head motion parallax cannot be supported.
3D Video/FTV

- MPEG has actively contributed compression technology for stereo and multi-view video, and is considering to take the next steps towards 3D and free-viewpoint video.

- Main Objectives:
  - Define generic formats that provide the high-quality reconstruction of a limited number of input views for advanced stereoscopic processing functionality and to support auto-stereoscopic displays.
  - Inclusion of depth: decouple number of transmitted views with number of required views for display

Depth-based video formats
High-Level View of 3DV/FTV Technology Chain

Support for high-quality auto-stereoscopic displays, such that the new format enables the generation of many high-quality views from a limited amount of input data, e.g. stereo and depth.

Enabling stereo devices to cope with varying -display types and sizes -different viewing preferences

This includes the ability to vary the baseline distance for stereo video to adjust the depth perception, which could help to avoid fatigue and other viewing discomforts.

Y.L. Chan, Centre for Signal Processing, POLYU
Bit Rate vs. 3D Rendering Capabilities

- Advancing beyond the current MVC standard for 3DV/FTV applications

Simulcast

MVC

3DV

should be compatible with
- existing standards
- Mono and stereo devices
- Existing or planned infrastructure

only a very limited continuum around the available original view since view synthesis artifact increase dramatically with the distance of the virtual viewpoint.
Anticipated 3D Video Format

Due to limitations in the production environment, the 3DV data format is assumed to be based on limited camera inputs.

Stereoscopic displays
- Variable stereo baseline
- Adjust depth perception

Possible for a large number of views to be generated from the reconstructed data format

Auto-stereoscopic N-view displays
- Wide viewing angle
- Large number of output views
Video plus Depth (V+D)

Per pixel depth data can be regarded as a monochromatic, luminance-only video signal.

A rough number is that 10-20% of the bit rate which is necessary to encode the color video is sufficient to encode the depth at good quality. This is due to the specific statistics of depth data, being on average more smooth and less structured than color data.

The stereo impression can be adjusted and customized after transmission. Also more than 2 views can be generated at the decoder enabling support of multiview displays and head motion parallax viewing within practical limits.
The concept of V+D is highly interesting due to the backward compatibility, compression efficiency and extended functionality.

Does not need to specify any coding algorithms
- only necessary to specify high-level syntax that allows a decoder to interpret two incoming video streams correctly as color and depth. Additionally, information about depth range (Znear and Zfar) needs to be transmitted.

MPEG has already defined MPEG-C Part 3 “ISO/IEC 23002-3 Representation of Auxiliary Video and Supplemental Information” in early 2007
- Format enabling simple stereoscopic application using standard video codecs.
- Allows one video plus depth data from which a second view is generated.
- Rate not significant increased compared to monoscopic video plus depth.
Video plus Depth (V+D)

Disadvantage:

- increased complexity for both sender side and receiver side: view synthesis has to be performed after decoding to generate the 2nd view of the stereo pair.
- only a very limited continuum around the available original view since view synthesis artifact increase dramatically with the distance of the virtual viewpoint.

Before encoding the depth data have to be generated. This is usually done by depth/disparity estimation from a captured stereo pair.
- Such algorithms are highly complex and still error prone.

In the near future V+D might be more suitable for applications with playback functionality, where depth estimation can be performed offline on powerful machines
- in a production studio
- home 3D editing suite
- enabling viewing of downloaded 3D video clips and 3D-TV
MVD involves a number of highly complex and error prone processing steps.
- Depth has to be estimated for the N views and the sender.
- N color with N depth videos have to be encoded and transmitted.
- At the receiver the data have to be decoded and the virtual views have to be rendered.
Multiview Video plus Depth (MVD)

- General Approach: N input views with only K views encoded (K<N).
  - At minimum: One video, one depth map
- M output views with quality depending on
  - Number of K encoded views
  - Compactness of representation/data rate R
  - Accuracy of depth information
  - Quality of view interpolation
- Example: **N-view displays** and only low amount of views (1-3) and associated depth maps are encoded
  - Generate synthesized views using video and depth
- Technologies required (**New Challenges**):
  - Depth estimation (non-normative)
  - Depth encoding (normative)
  - View synthesis (non-normative or with minimum normative quality requirements)

Capture → Correction → Depth Search → Encoder → Transmission /Storage → Decoder → Interpolation → Display

N × K × video + depth → M
Efficient support of multiview auto-stereoscopic displays based on MVD

The remaining views can be synthesized by depth image based rendering (DIBR)

Only 3 original views V1, V5, and V9 are in the decoded stream + corresponding depth maps D1, D5, and D9.
Focus of work in MPEG Video Subgroup/JVT activities

- Future 3D displays requiring different views (up to 20 – 50 views) supporting relatively narrow variation of view angles.
- Generate synthesized views by using video and depth.

**Technologies required (New Challenges):**

- Depth estimation (non-normative)
- Depth encoding (normative)
- View synthesis (non-normative or with minimum normative quality requirements)

Under assessment: Depth estimation and view synthesis algorithms.

- Coding tools of video and depth not yet considered;
- Definition of date rate/combination points (video, depth)
- To provide reference anchors (MVC video+MVC depth+synthesis) in upcoming CfP.
Summary and Future Challenges

- 3D Video Technology is maturing due to worldwide development from capturing to display.

- MPEG has actively contributed compression technology for stereo and multi-view video, and is considering to take the next steps towards 3D and free-viewpoint video.

- Trying to define generic formats that
  - Consider capturing technology, i.e. maximal 2-3 input views
  - Break linear dependency of coding bitrate from number of target views
  - Provide scene geometry data in general form (depth)

- H.264 only optimized for 2D color video, but not for depth information
  - New 3D video coding beyond AVC/MVC is required and being developed.
  - Depth coding technique is attracting more attention.
  - It should consider statistical properties of depth and color video to improve coding efficiency.

- New quality evaluation methods for intermediate views

- Provide high-quality view synthesis for continuous view range