Depth Error Induced Virtual View Synthesis Distortion Estimation for 3D Video Coding

Yijian Xiang, Lu Fang, Ren Li
University of Science and Technology of China

Ngai-Man (Man) Cheung
Singapore University of Technology and Design
Multiview Video Plus Depth

- Multiple video sequences captured by an array of cameras
- Each texture image has an associated depth map
- With MVD data, depth-image-based rendering (DIBR) can be used to generate virtual view
With 1-D parallel arrangement of cameras, there is only horizontal disparity $m - m'$.

Based on the disparity, location in the virtual image can be identified to warp the pixel into

$$m - m' = \frac{fb}{z}$$
Effect of Error in Depth Image

- Depth images may contain errors due to lossy compression
- It is not clear how the errors affect the synthesis quality

Depth error: position error in synthesis
- Pixels are warped to slightly shifted positions during synthesis
- Effects are very subtle, e.g., depend on the image contents

Accurate analytical model to estimate the synthesis distortion is very valuable for 3DV system design
Contributions

- Analysis the effect of depth coding error to synthesis distortion
- A model to relate the virtual camera position to the synthesis distortion
- Experiment the model with video sequences and synthesis tools from the 3D-HEVC activities
A local estimation model for sum of squared error (SSE) which is expressed in terms of variance of a video block and an autoregressive model for correlation coefficient, [Kim, Ortega, Lai, Tian, Gomila, 2010]

Cubic synthesis distortion model with simplified synthesis model, [Cheung, Velisavljevic, Ortega, 2011]

Taylor expansion theory based synthesis distortion model, [Yuan, Chang, Huo, Yang, Lu, 2011]

Model based on Power Spectral Density (PSD) and spatial analysis, [Fang, Cheung, Tian, Vetro, Sun, Au, 2014]
- Two reference texture images are captured by the left and right cameras along with their associated depth images.
- Synthesize an image at a certain virtual camera position.
- Two steps: (i) warping and (ii) merging.
Copy the pixels from the reference texture $X_l (X_r)$ to the virtual image $Y_l (Y_r)$.

The pixel copying takes into account the horizontal disparity:

$$m - m' = \frac{D_l(m',n)}{255} (d_{near} - d_{far}) + d_{far}$$
Synthesis Model: Merging

Merging by linear combination:

\[ Y(m,n) = (1-x)Y_l(m,n) + xY_r(m,n) \]

\[ x = \frac{b_l}{b} \]
Synthesis Distortion

The reference texture pixel is warped to a slightly-shifted position in the virtual image.
Synthesis Distortion

\[ D_l(m,n) \]

\[ X_l(m,n) \rightarrow Y_l(m,n) \]

\[ Y(m,n) \]

\[ D_l(m,n) \]

\[ X_l(m,n) \rightarrow Y_l(m,n) \]

\[ Y(m,n) \]

\[ X_l(m,n) \rightarrow Y_r(m,n) \]

\[ Y(r(m,n)) \]

\[ X_l(m,n) \rightarrow W_l(m,n) \]

\[ W(m,n) \]

\[ X_l(m,n) \rightarrow W_r(m,n) \]

\[ W(r(m,n)) \]

\[ X_l(m,n) \rightarrow X_r(m,n) \]

\[ Z(m,n) \]

\[ E[Z^2], Z = Y - W \]

\[ Z \text{ is the synthesis distortion signal} \]

Original depth

Reconstructed depth
Synthesis Distortion

We refine this modeling to take into account certain scene characteristics to obtain better estimation of the synthesis distortion.

Original depth

Reconstructed depth
Refined Synthesis Model

\[ Y(m, n) = \begin{cases} 
(1 - x)Y_l(m, n) + xY_r(m, n) & \text{when both } Y_l(m, n), Y_r(m, n) \text{ are available} \\
Y_l(m, n) & \text{only } Y_l(m, n) \text{ is available} \\
Y_r(m, n) & \text{only } Y_r(m, n) \text{ is available}
\end{cases} \]

\[ x = \frac{b_l}{b} \]

Situation with disocclusion:
Refined Synthesis Model

\[ Y(m, n) = \begin{cases} 
(1 - x)Y_l(m, n) + xY_r(m, n) & \text{when both } Y_l(m, n), \ Y_r(m, n) \text{ are available} \\
Y_l(m, n) & \text{only } Y_l(m, n) \text{ is available} \\
Y_r(m, n) & \text{only } Y_r(m, n) \text{ is available}
\end{cases} \]

\[ x = \frac{b_l}{b} \]

Similar for \( W(m, n) \)

(warping using reconstructed depth)

\[ W(m, n) = \begin{cases} 
(1 - x)W_l(m, n) + xW_r(m, n) & \text{when both } W_l(m, n), \ W_r(m, n) \text{ are available} \\
W_l(m, n) & \text{only } W_l(m, n) \text{ is available} \\
W_r(m, n) & \text{only } W_r(m, n) \text{ is available}
\end{cases} \]
Refined Synthesis Model

\[ Y(m, n) = \begin{cases} 
(1 - x)Y_l(m, n) + xY_r(m, n) & \text{when both } Y_l(m, n), Y_r(m, n) \text{ are available} \\
Y_l(m, n) & \text{only } Y_l(m, n) \text{ is available} \\
Y_r(m, n) & \text{only } Y_r(m, n) \text{ is available}
\end{cases} \]

\[ x = \frac{b_l}{b} \]

When we compute \( Z = Y - W \), we should consider that \( Y(W) \) could be any of the three situations.

Similar for \( W(m, n) \)

(warping using reconstructed depth)

\[ W(m, n) = \begin{cases} 
(1 - x)W_l(m, n) + xW_r(m, n) & \text{when both } W_l(m, n), W_r(m, n) \text{ are available} \\
W_l(m, n) & \text{only } W_l(m, n) \text{ is available} \\
W_r(m, n) & \text{only } W_r(m, n) \text{ is available}
\end{cases} \]
Synthesis Distortion Estimation

$Y_k$: warping using original depth
$W_k$: warping using reconstructed depth

<table>
<thead>
<tr>
<th>Situation $k$</th>
<th>$Y_k$</th>
<th>$W_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y = (1-x)Y_l + xY_r$</td>
<td>$W = (1-x)W_l + xW_r$</td>
</tr>
<tr>
<td>2</td>
<td>$Y = Y_l$</td>
<td>$W = W_l$</td>
</tr>
<tr>
<td>3</td>
<td>$Y = Y_r$</td>
<td>$W = W_r$</td>
</tr>
<tr>
<td>4</td>
<td>$Y = (1-x)Y_l + xY_r$</td>
<td>$W = W_l$</td>
</tr>
<tr>
<td>5</td>
<td>$Y = Y_l$</td>
<td>$W = (1-x)W_l + xW_r$</td>
</tr>
<tr>
<td>6</td>
<td>$Y = Y_r$</td>
<td>$W = (1-x)W_l + xW_r$</td>
</tr>
<tr>
<td>7</td>
<td>$Y = (1-x)Y_l + xY_r$</td>
<td>$W = W_r$</td>
</tr>
<tr>
<td>8</td>
<td>$Y = Y_l$</td>
<td>$W = W_r$</td>
</tr>
<tr>
<td>9</td>
<td>$Y = Y_r$</td>
<td>$W = W_l$</td>
</tr>
</tbody>
</table>

$E[Z^2] = \sum_{k=1}^{9} E[Z^2 | Z = Z_k] \times p_k$

$= \sum_{k=1}^{9} E[(Y - W)^2 | (Y = Y_k, W = W_k)] \times p_k$

9 different situations that $Z$ should be computed:

Distortion under Different Situations (DDS)

Probability under Different Situations (PDS)
Synthesis Distortion Estimation: Main Results

- Probability under Different Situations (PDS) $p_k$ are linear functions of $x$

- Distortion under Different Situations (DDS):
  
  $$E[(Y-W)^2|Y=Y_k, W=W_k]$$

  are quadratic / biquadratic functions of $x$

\[
E[Z^2] = \sum_{k=1}^{9} E[Z^2|Z = Z_k] \times p_k \\
= \sum_{k=1}^{9} E[(Y - W)^2|(Y = Y_k, W = W_k)] \times p_k,
\]

- Probability under Different Situations (PDS)

- Distortion under Different Situations (DDS)
PDS Modeling

- PDS model – Disocclusion and boundary region (situation 3: \( Y = Y_r, \ W = W_r \))
  - Regions in virtual image that contain information from only one single reference image
  - Caused by dis-occlusion and limitation of camera view

\[
\frac{d_2}{d_1} = \frac{x_2}{x_1}
\]

\[p_3(x) = cx\]
PDS Modeling

- **PDS model – Depth Error Region** (*Situation 6: Y = Y_r, W = (1 - x)W_l + xW_r*)
  - Because of the depth error, some pixel in the left reference may be mistakenly warped into the disocclusion region
PDS Modeling

- PDS model – Depth Error Region (*Situation 6: Y = Y_r, W = (1 - x)W_l + xW_r*)
  - Because of the depth error, some pixel in the left reference may be mistakenly warped into the disocclusion region

With reconstructed depth

- Probability for a given pixel \( a \) to be shifted into the disocclusion region, \( \gamma(a) \):
  \[
  \gamma(a) = P(m_0 - m < \Delta m < m_1 - m)
  = \int_{m_0 - m}^{m_1 - m} f(\Delta m)d\Delta m
  \]
  - Here \( f(\Delta m) \) is the pdf of the disparity error \( \Delta m \).
  - Expected number of pixels shifted into the disocclusion region:
    \[
    E[N] = \sum_a \gamma(a)
    \]
  - Approximately, \( E[N] \) is directly proportional to \( x = b_1 / b \)

\[
p_6(x) = c'x
\]
Synthesis Distortion Estimation

\[ E[Z^2] = \sum_{k=1}^{9} E[Z^2 | Z = Z_k] \times p_k \]

\[ = \sum_{k=1}^{9} E[(Y - W)^2 | (Y = Y_k, W = W_k)] \times p_k, \]

Linear combination of DDS model (quadratic/biquadratic) and PDS model (linear)

<table>
<thead>
<tr>
<th>Situation</th>
<th>DDS model</th>
<th>PDS model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(E(Y_l - W_l)^2 : c_{11}x + c_{12}x^2; E(Y_r - W_r)^2 : c_{13}(1 - x) + c_{14}(1 - x)^2)</td>
<td>(c_{15})</td>
</tr>
<tr>
<td>2</td>
<td>(c_{21}x + c_{22}x^2)</td>
<td>(c_{23}(1 - x))</td>
</tr>
<tr>
<td>3</td>
<td>(c_{31}(1 - x) + c_{32}(1 - x)^2)</td>
<td>(c_{33}x)</td>
</tr>
<tr>
<td>4</td>
<td>(c_{41}x + c_{42}x^2)</td>
<td>(c_{43}(1 - x))</td>
</tr>
<tr>
<td>5</td>
<td>(c_{51}x + c_{52}x^2 + c_{53}x^3 + c_{54}x^4)</td>
<td>(c_{55}(1 - x))</td>
</tr>
<tr>
<td>6</td>
<td>(c_{61}(1 - x) + c_{62}(1 - x)^2 + c_{63}(1 - x)^3 + c_{64}(1 - x)^4)</td>
<td>(c_{65}x)</td>
</tr>
<tr>
<td>7</td>
<td>(c_{71}(1 - x) + c_{72}(1 - x)^2)</td>
<td>(c_{73}x)</td>
</tr>
</tbody>
</table>

Distortion under Different Situations (DDS)

Probability under Different Situations (PDS)
Experiment Results

- Test video sequences: kendo, balloons, champagne, etc.
- Depth images in reference views are quantized with several QP: 36, 40, 44, 48

Verification of linear model for PDS with kendo sequence under QP=48:

\[ p_1(x) \]

\[ p_2(x) \]

\[ p_3(x) \]

\[ p_4(x) \]

\[ p_5(x) \]

\[ p_6(x) \]

\[ p_7(x) \]

Empirical  Model
Experiment Results

- Verification of synthesis distortion model with balloons sequence under different QP
- $c_0$ to $c_5$ are estimated from empirical synthesis distortions of several synthesized views

![Graphs showing MSE for different QPs](image)

- MSE, depth QP=36
- MSE, depth QP=40
- MSE, depth QP=44
- MSE, depth QP=48
Conclusions

- Investigated the model to relate the virtual camera position to the synthesis distortion
- Based on detailed analysis, a polynomial function of degree 5 can characterize the synthesis distortion at different virtual camera positions
- Performed experiments to verify the model
Thank you
Contributions

- Propose an analytical model to estimate the depth-error-induced virtual view synthesis distortion (VVSD) in 3D video

- Analyze the merging operations under different situations that affect pixel availability: overlapping region, disocclusion and boundary region, disparity error region, and infrequent region

- Show that VVSD is the linear combination of Distortion under Different Situations (DDS) weighted by and Probability under Different Situations (PDS)

- Prove that quadratic/biquadratic models and linear models are capable of estimating DDS and PDS respectively.
Approach

- NINE situations

<table>
<thead>
<tr>
<th>Situation $k$</th>
<th>$Y_k$</th>
<th>$W_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y = (1-x)Y_l + xY_r$</td>
<td>$W = (1-x)W_l + xW_r$</td>
</tr>
<tr>
<td>2</td>
<td>$Y = Y_l$</td>
<td>$W = W_l$</td>
</tr>
<tr>
<td>3</td>
<td>$Y = Y_r$</td>
<td>$W = W_r$</td>
</tr>
<tr>
<td>4</td>
<td>$Y = (1-x)Y_l + xY_r$</td>
<td>$W = W_l$</td>
</tr>
<tr>
<td>5</td>
<td>$Y = Y_l$</td>
<td>$W = (1-x)W_l + xW_r$</td>
</tr>
<tr>
<td>6</td>
<td>$Y = Y_r$</td>
<td>$W = (1-x)W_l + xW_r$</td>
</tr>
<tr>
<td>7</td>
<td>$Y = (1-x)Y_l + xY_r$</td>
<td>$W = W_l$</td>
</tr>
<tr>
<td>8</td>
<td>$Y = Y_l$</td>
<td>$W = W_r$</td>
</tr>
<tr>
<td>9</td>
<td>$Y = Y_r$</td>
<td>$W = W_l$</td>
</tr>
</tbody>
</table>

- Cluster 1: Overlapping region
- Cluster 2: Situation 3, Disocclusion and boundary region
- Cluster 3: Situation 6, Disparity error region
- Cluster 4: Situation 8, Infrequent region
Approach

- PDS model – Overlapping region

  - Y\W contains information from left and right views, i.e., \( Y|Y = (1-x)Y_l + xY_r \)

  - Cluster 1 takes place with high probability, as in practical camera capture system, adjacent cameras usually capture scenes with a lot of contents being identical

\[ p_1(x) \approx constant \]
Approach

- PDS model – Disocclusion and boundary region (situation 3: $Y = Y_r$, $W = W_r$)

  - Cluster 2 contains merely one single reference view information, caused by disocclusion problem and limitation of camera capture range.

  \[ \frac{m'' - m'}{Kb_l} = \frac{m' - m'}{b_l} \]

  \[ \frac{m - m''}{Kb_l} = \frac{m - m'}{b_l} \]

  \[ p_3(x) = cx \]
Approach

- PDS model – Disparity Error Region (*Situation 6: Y = Y_r, W = (1 - x)W_l + xW_r*)
  
  It occurs due to that a pixel from left reference view shift to disocclusion region because of disparity error.

\[
P(m'_0 - m' < \Delta m' < m'_1 - m') = \int_{m'_0 - m'}^{m'_1 - m'} f(\Delta m')d\Delta m'
\]

\[
P(m''_0 - m'' < \Delta m'' < m''_1 - m'') = \int_{m''_0 - m''}^{m''_1 - m''} f(\Delta m'')d\Delta m''
\]

Pixels that might shift to disocclusion region from left side:

\[
N = \frac{L}{\Delta x} \times \int_0^L \frac{P(.)}{L} \, dl
\]

\[
N_{V2} \approx KN_{V1}
\]

\[
p_6(x) = c'x
\]
Approach

- PDS model – Infrequency region (situation 8: $Y = Y_l$, $W = W_r$)
  - It occurs when $Y_r$ coincidentally shifts to a new position, where the content originally belongs to $Y_l$, however a black hole appears due to unexpected shift of $Y_l$.
  - It has low probability to occur, and is abandoned by our model.

\[ p_8(x) \approx 0 \]
Conclusions

- We have proposed an analytical model, which is capable of estimating the depth-error-induced virtual view synthesis distortion (VVSD) in 3D video.

- We have decomposed VVSD into Distortion under Different Situations (DDS) weighted by and Probability under Different Situations (PDS).

- We have proved that DDS and PDS follows quadratic/biquadratic models and linear models, respectively.

- The proposed synthesis distortion model can fit empirical data under different scenes and different compression QP.
## DDS Modeling

\[
E[Z^2] = \sum_{k=1}^{9} E[Z^2 | Z = Z_k] \times p_k
\]

\[
= \sum_{k=1}^{9} E[(Y - W)^2 | (Y = Y_k, W = W_k)] \times p_k,
\]

\(Y_k\): warping using original depth  
\(W_k\): warping using reconstructed depth  

### Distortion under Different Situations (DDS)

| Situation \(k\) | \(Y_k\) | \(W_k\) | \(E[(Y - W)^2 | (Y = Y_k, W = W_k)]\) |
|-----------------|---------|---------|----------------------------------|
| 1               | \(Y = (1-x)Y_l + xY_r\) | \(W = (1-x)W_l + xW_r\) | \([1-x]^2E(Y_l - W_l)^2 + x^2E(Y_r - W_r)^2 + 2x(1-x)E(Y_l - W_l)(Y_r - W_r)\) |
| 2               | \(Y = Y_l\) | \(W = W_l\) | \([E(Y_l - W_l)^2]\) |
| 3               | \(Y = Y_r\) | \(W = W_r\) | \([E(Y_r - W_r)^2]\) |
| 4               | \(Y = (1-x)Y_l + xY_r\) | \(W = W_l\) | \([E(Y_l - W_l)^2 + x^2E(Y_r - Y_l)^2 + 2xE(Y_l - W_l)(Y_r - Y_l)\] |
| 5               | \(Y = Y_l\) | \(W = (1-x)W_l + xW_r\) | \([E(Y_l - W_l)^2 + x^2E(W_l - W_r)^2 + 2xE(Y_l - W_l)(W_l - W_r)\] |
| 6               | \(Y = Y_r\) | \(W = (1-x)W_l + xW_r\) | \([E(Y_r - W_r)^2 + (1-x)^2E(W_l - W_l)^2 + 2(1-x)E(Y_r - W_r)(W_l - W_l)\] |
| 7               | \(Y = (1-x)Y_l + xY_r\) | \(W = W_r\) | \([E(Y_r - W_r)^2 + (1-x)^2E(Y_l - Y_r)^2 + 2(1-x)E(Y_l - W_l)(Y_l - Y_r)\] |
| 8               | \(Y = Y_l\) | \(W = W_r\) | \([E(Y_l - W_r)^2]\) |
| 9               | \(Y = Y_r\) | \(W = W_l\) | \([E(Y_r - W_l)^2]\) |
DDS Modeling

DDS are functions of these basic terms:
- Virtual view distortion caused by a single reference view: $E(Y_l-W_l)^2$, $E(Y_r-W_r)^2$
- Virtual view distortion caused by two reference views: $E(Y_r-Y_l)^2$, $E(W_r-W_l)^2$
- Cross-correlation terms: $E(Y_l-W_l)(W_l-W_r)$ etc.

| Situation k | $Y_k$ | $W_k$ | $E[(Y-W)^2| (Y = Y_k, W = W_k)]$ |
|-------------|-------|-------|---------------------------------|
| 1           | $Y = (1-x)Y_l + xY_r$ | $W = (1-x)W_l + xW_r$ | $[1-(1-x)^2] E(Y_l - W_l)^2 + x^2 E(Y_r - W_r)^2 + 2x(1-x) E(Y_l - W_l)(Y_r - W_r)$ |
| 2           | $Y = Y_l$       | $W = W_l$       | $E(Y_l - W_l)^2$ |
| 3           | $Y = Y_r$       | $W = W_r$       | $E(Y_r - W_r)^2$ |
| 4           | $Y = (1-x)Y_l + xY_r$ | $W = W_l$       | $E(Y_l - W_l)^2 + x^2 E(Y_r - Y_l)^2 + 2x E(Y_l - W_l)(Y_r - Y_l)$ |
| 5           | $Y = Y_l$       | $W = (1-x)W_l + xW_r$ | $E(Y_l - W_l)^2 + x^2 E(W_l - W_r)^2 + 2x E(Y_l - W_l)(W_l - W_r)$ |
| 6           | $Y = Y_r$       | $W = (1-x)W_l + xW_r$ | $E(Y_r - W_r)^2 + (1-x)^2 E(W_r - W_l)^2 + 2(1-x) E(Y_r - W_r)(W_r - W_l)$ |
| 7           | $Y = (1-x)Y_l + xY_r$ | $W = W_r$       | $E(Y_r - W_r)^2 + (1-x)^2 E(Y_l - Y_r)^2 + 2(1-x) E(Y_r - W_r)(Y_l - Y_r)$ |
| 8           | $Y = Y_l$       | $W = W_r$       | $E(Y_l - W_r)^2$ |
| 9           | $Y = Y_r$       | $W = W_l$       | $E(Y_r - W_l)^2$ |

It can be shown that the basic terms are quadratic functions of $x$
Thus, DDS are biquadratic functions of $x$
Synthesis Distortion Estimation

\[ E[Z^2] = c_5 x^5 + c_4 x^4 + c_3 x^3 + c_2 x^2 + c_1 x + c_0 \]

\(c_i\) depends on:
- Image characteristic \(E[X^2(m,n)], E[X(m,n) X(m-1,n)]\)
- Variance of depth error (depth QP)
- Distance between left and right reference views \(b\), focal length \(f\)
3D Video

- 3D video (3DV) datasets usually consist of:
  - Multiple video sequences captured by cameras at different positions
  - Associated depth maps
- Per-pixel depth information allows synthesis of virtual views at user-chosen viewpoints via depth-image-based rendering (DIBR)
With per pixel depth information transmitted, we can derive the disparity at each position and determine where to copy the pixel to the virtual view during rendering.

\[ m - m' = \frac{fb}{z} \]