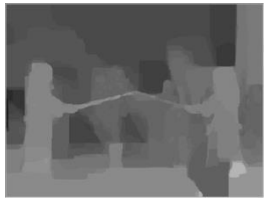
Depth Error Induced Virtual View Synthesis Distortion Estimation for 3D Video Coding

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

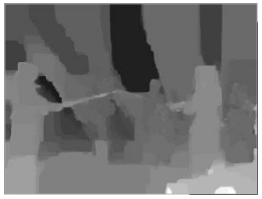
Ngai-Man (Man) Cheung
Singapore University of Technology and Design

Multiview Video Plus Depth

Depth map



...

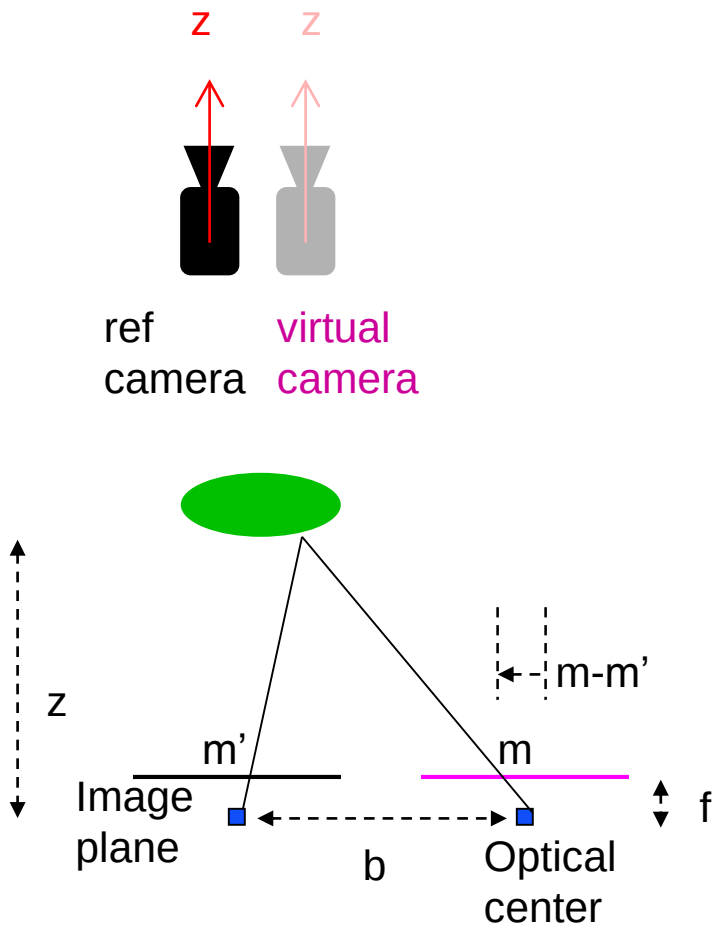


Texture image

- 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

Camera array:

Depth-Image-Based Rendering



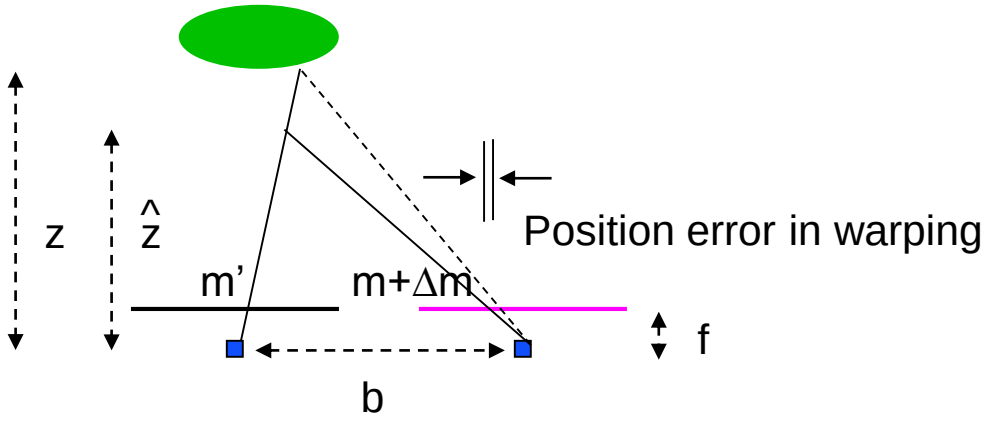
- With 1-D parallel arrangement of cameras, there is only horizontal disparity m-m'
- Based on the disparity, location in the virtual image can be identified to warp the pixel into

$$m - m' = \frac{fb}{z}$$

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

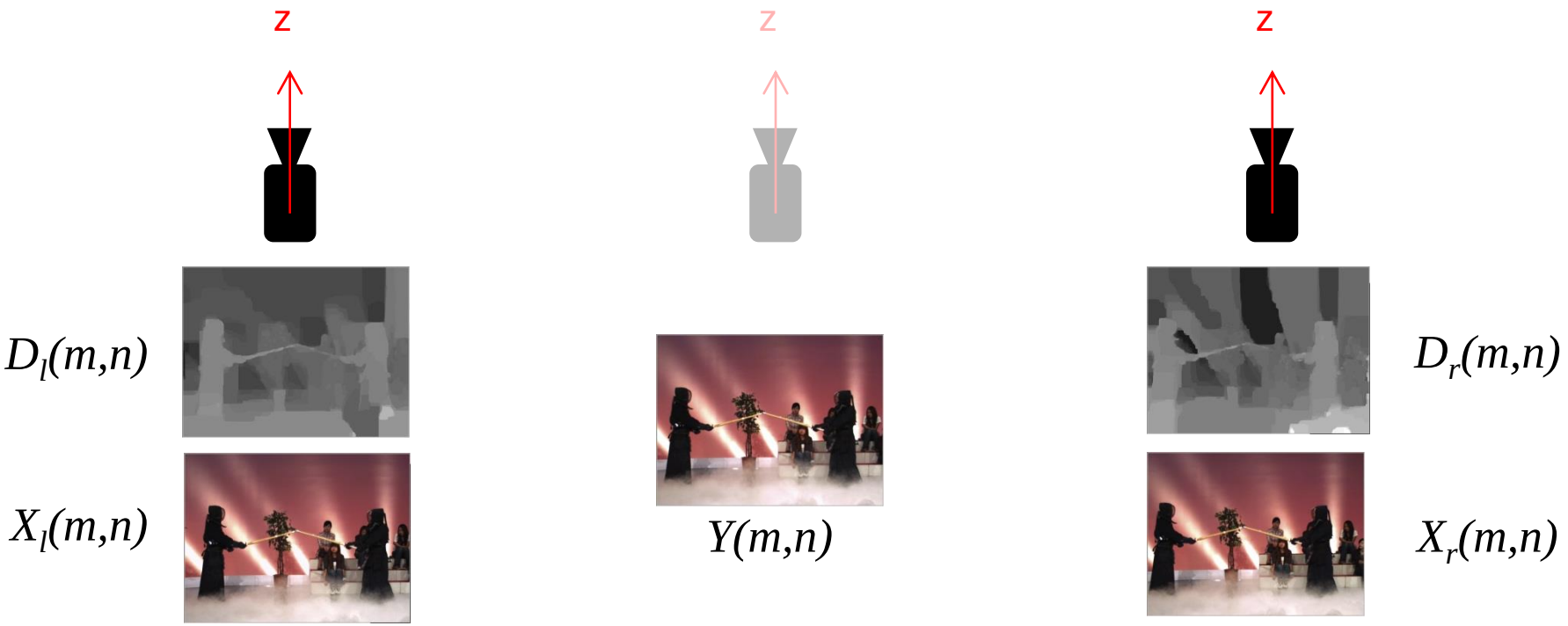
Contributions

- 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

Related Work

- 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]

Synthesis Model



-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 positon
-Two steps: (i) warping and (ii) merging

Synthesis Model: Warping

$D_l(m,n)$

m' m

$X_l(m,n)$ $Y_l(m,n)$ $X_r(m,n)$

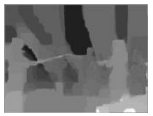
Pixel copying
(Warping)

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:

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

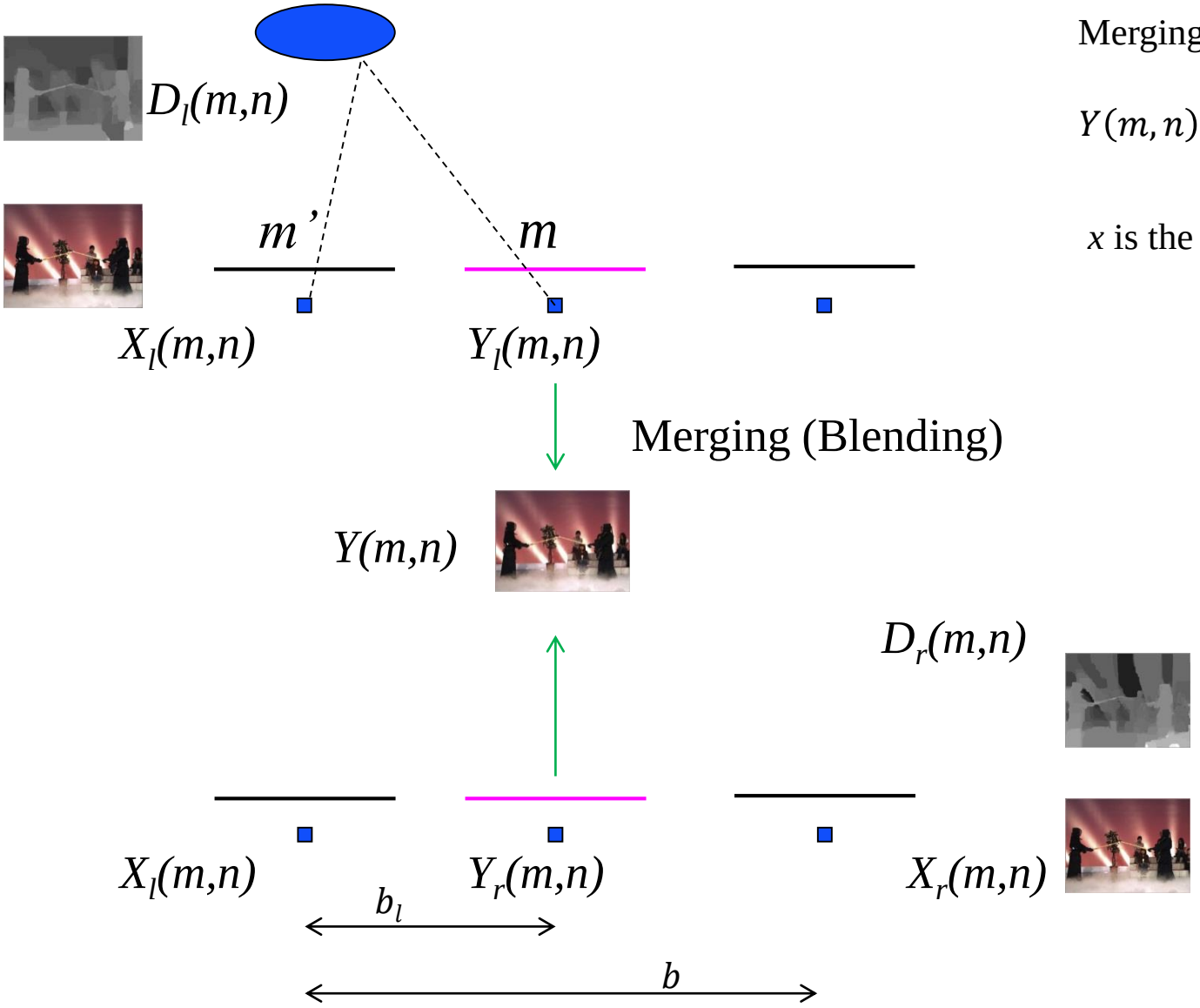
$D_r(m,n)$



m m''

$X_l(m,n)$ $Y_r(m,n)$ $X_r(m,n)$

Synthesis Model: Merging



Merging by linear combination:

$$Y(m,n) = (1-x)Y_l(m,n) + xYr(m,n)$$

x is the normalized distance:

$$x = \frac{b_l}{b}$$

Synthesis Distortion


$D_l(m,n)$

m' m

$X_l(m,n)$ $Y_l(m,n)$

$Y(m,n)$

$X_l(m,n)$ $Y_r(m,n)$ $X_r(m,n)$

Original depth

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

$\hat{D}_l(m,n)$ $\hat{D}_r(m,n)$

m' $m+\Delta m$

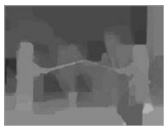
$X_l(m,n)$ $W_l(m,n)$ $X_r(m,n)$

$W(m,n)$

$X_l(m,n)$ $W_r(m,n)$ $X_r(m,n)$

Reconstructed depth

Synthesis Distortion




$D_l(m,n)$

m' m

$X_l(m,n)$ $Y_l(m,n)$

$Y(m,n)$

$X_l(m,n)$ $Y_r(m,n)$ $X_r(m,n)$

$\hat{D}_l(m,n)$ $\hat{D}_r(m,n)$

m' $m+\Delta m$

$X_l(m,n)$ $W_l(m,n)$ $X_r(m,n)$

$W(m,n)$

$X_l(m,n)$ $W_r(m,n)$ $X_r(m,n)$

$Z(m,n)$

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

Z is the synthesis distortion signal

Original depth **Reconstructed depth**

Synthesis Distortion

$D_l(m,n)$

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

$\hat{D}_r(m,n)$

m' m $m+\Delta m$

$X_l(m,n)$ $Y_l(m,n)$ $X_l(m,n)$ $W_l(m,n)$ $X_r(m,n)$

$Y(m,n)$

$W(m,n)$

$Z(m,n)$

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

Z is the synthesis distortion signal

$X_l(m,n)$ $Y_r(m,n)$ $X_r(m,n)$ $X_l(m,n)$ $W_r(m,n)$ $X_r(m,n)$

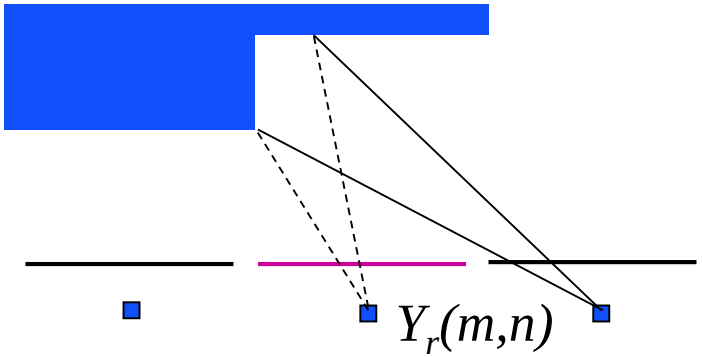
Original depth **Reconstructed depth**

Refined Synthesis Model

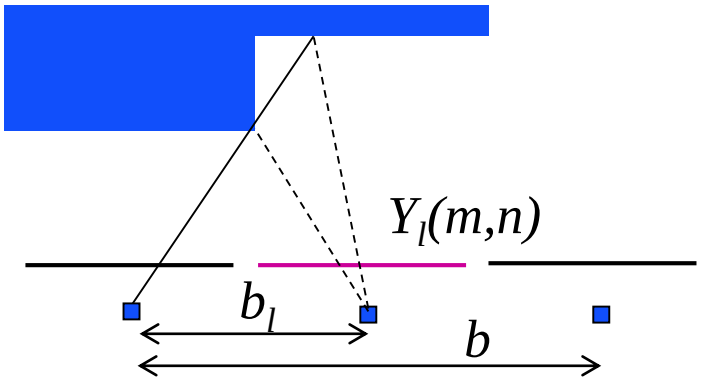
$$Y(m,n) = \begin{cases} (1-x)Y_l(m,n) + xYr(m,n) & \text{when both } Y_l(m,n),\ Y_r(m,n) \text{ are available} \\ Y_l(m,n) & \text{only } Y_l(m,n) \text{ is available} \\ Y_r(m,n) & \text{only } Y_r(m,n) \text{ is available} \end{cases}$$

$$x = \frac{b_l}{b}$$

Situation with disocclusion:



$X_r(m,n)$



$X_l(m,n)$

Refined Synthesis Model

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

$$x = \frac{b_l}{b}$$

Similar for $W(m,n)$

(warping using reconstructed depth)

$$W(m,n) = \begin{cases} (1-x)W_l(m,n) + xWr(m,n) & \text{when both } W_l(m,n),\ W_r(m,n) \text{ are available} \\ W_l(m,n) & \text{only } W_l(m,n) \text{ is available} \\ W_r(m,n) & \text{only } W_r(m,n) \text{ is available} \end{cases}$$

Refined Synthesis Model

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

$$x = \frac{b_l}{b}$$

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)

$$W(m,n) = \begin{cases} (1-x)W_l(m,n) + xWr(m,n) & \text{when both } W_l(m,n),\ W_r(m,n) \text{ are available} \\ W_l(m,n) & \text{only } W_l(m,n) \text{ is available} \\ W_r(m,n) & \text{only } W_r(m,n) \text{ is available} \end{cases}$$

Synthesis Distortion Estimation

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

9 different situations that Z should be computed:

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

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

Probability under Different Situations (PDS)

Distortion under Different Situations (DDS)

Synthesis Distortion Estimation: Main Results

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

- Distortion under Different Situations (DDS): $E[(Y\text{-}W)^2|Y=Y_k, W=W_k]$ are quadratic / biquadratic functions of x

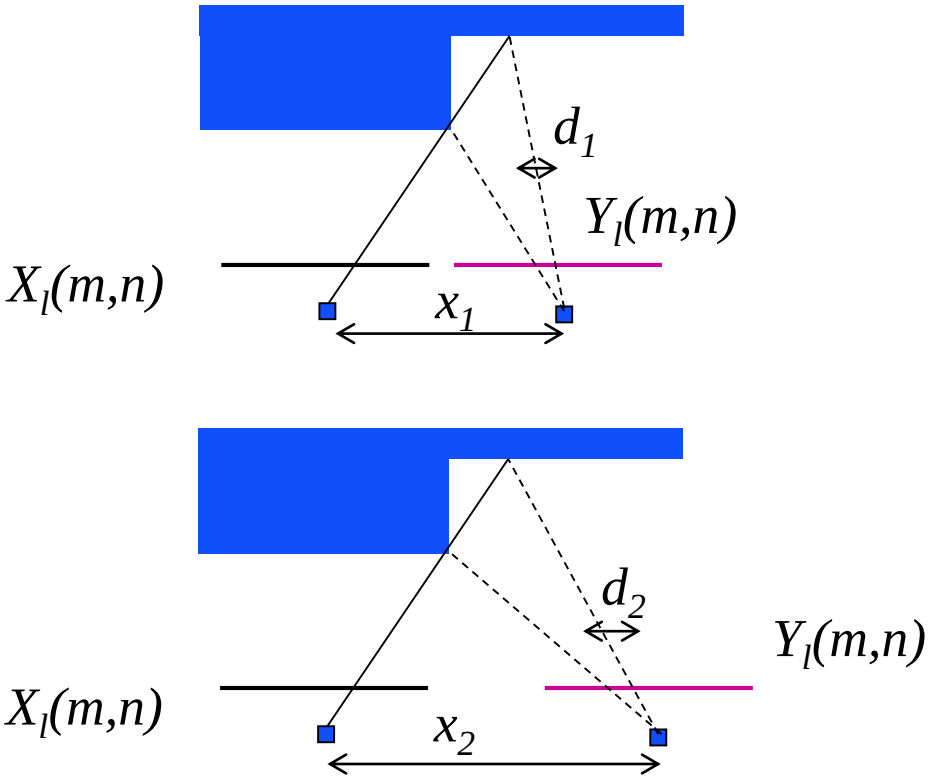
$$
\begin{aligned}
E[Z^2] &= \sum_{k=1}^{9} E[Z^2|Z = Z_k] \times p_k \\
&= \sum_{k=1}^{9} E[(Y - W)^2|(Y = Y_k, W = W_k)] \times p_k,
\end{aligned}
$$

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

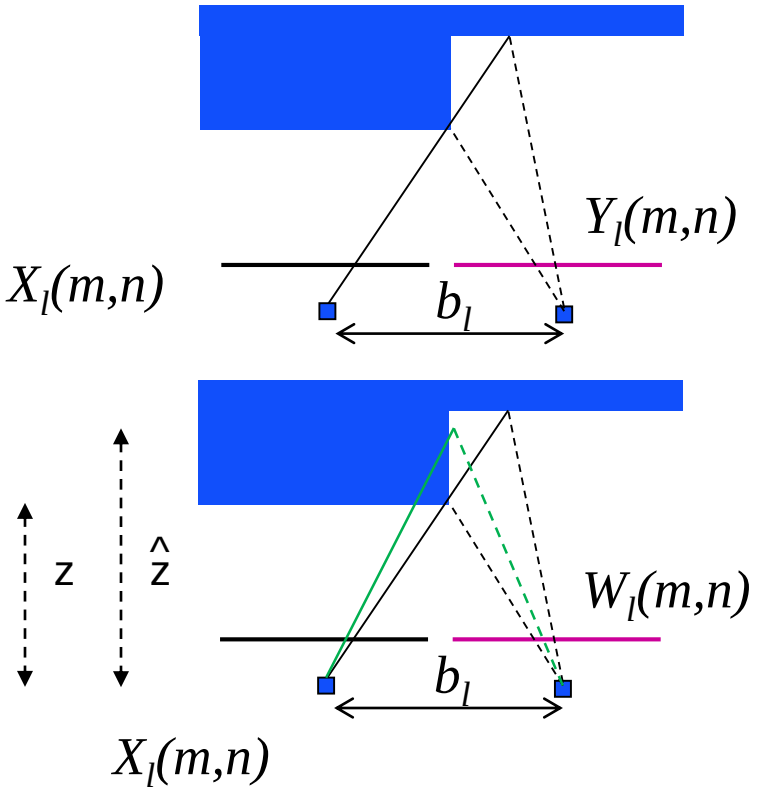


$$\frac{d_2}{d_1} = \frac{x_2}{x_1}$$

$$p_3(x) = cx$$

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



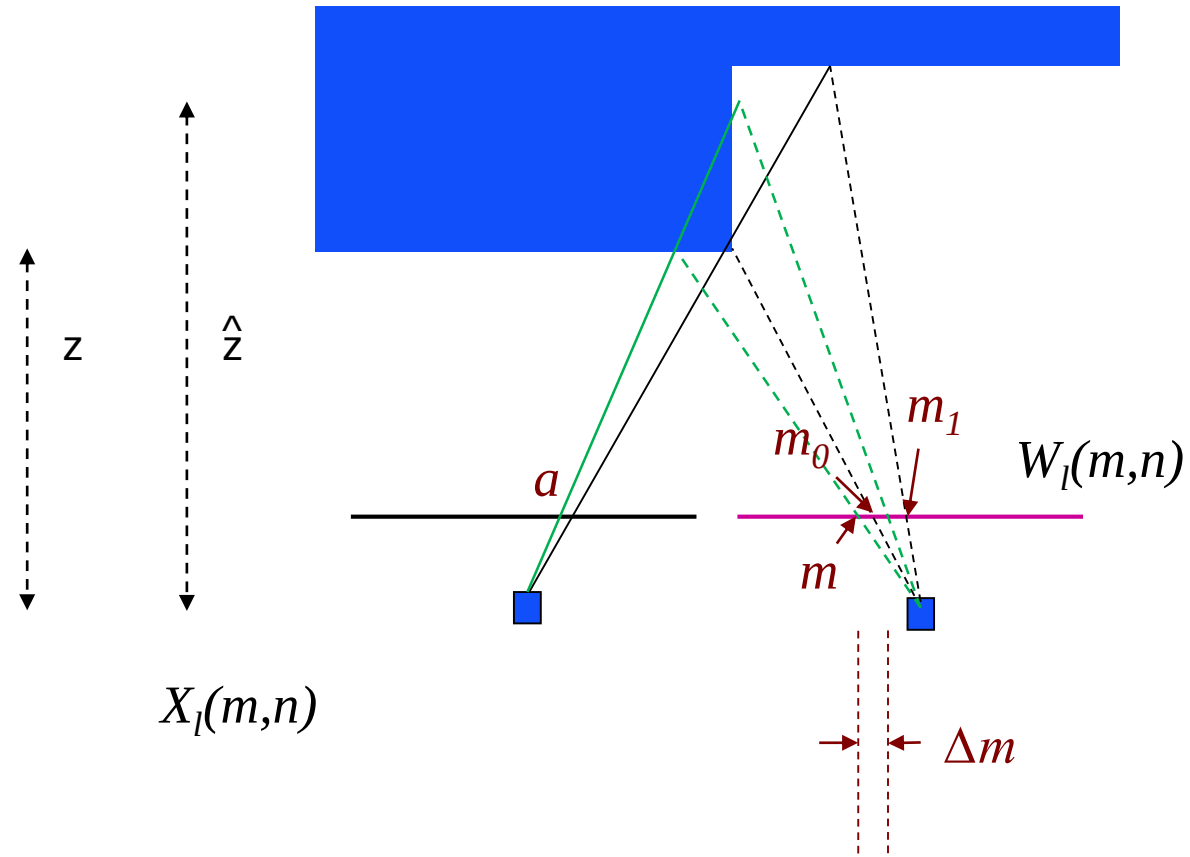
With original depth

With reconstructed depth

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 Δm.

-Expected number of pixels shifted into the disocclusion region:

$$E[N] = \sum_a \gamma(a)$$

-Approximately, $E[N]$ is directly proportional to $x = b_l / b$

$$\Rightarrow \quad p_6(x) = c'x$$

Synthesis Distortion Estimation

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

Probability under Different Situations (PDS)

Distortion under Different Situations (DDS)

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

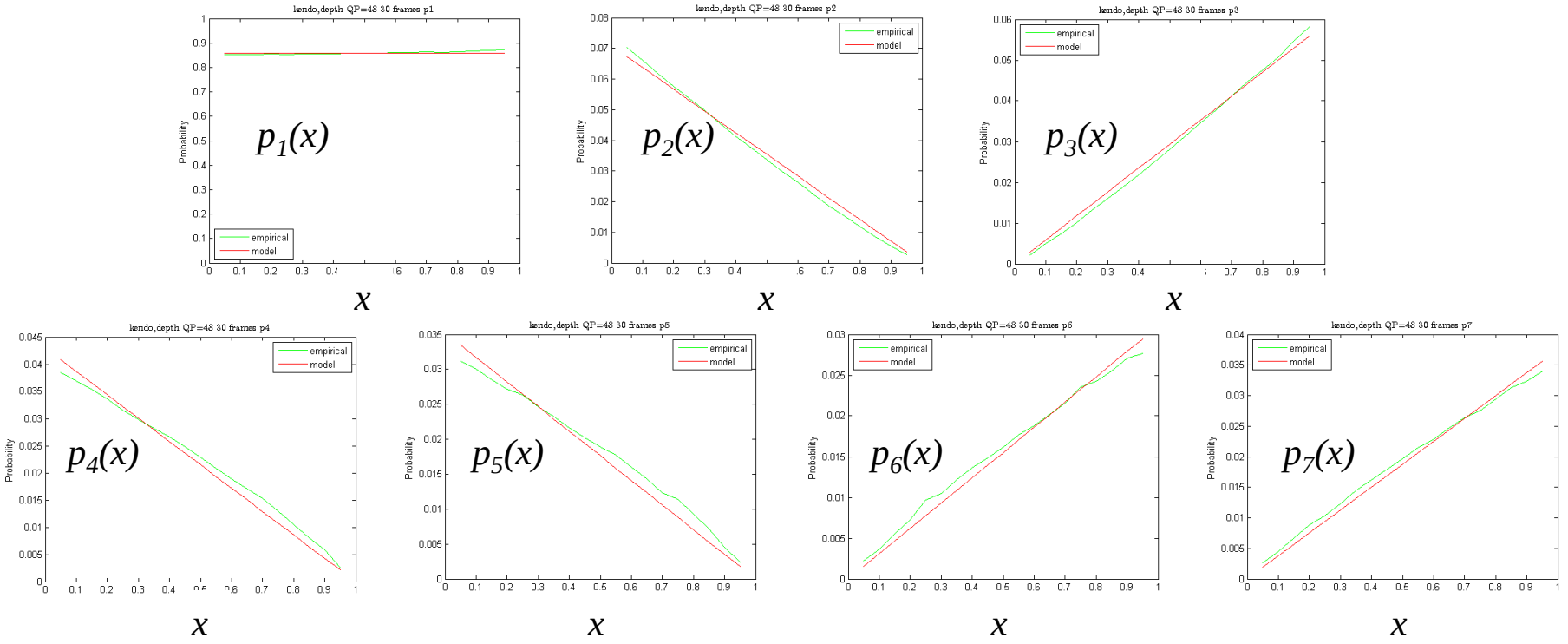
Linear combination of DDS model (quadratic/biquadratic) and PDS model (linear)

$$E[Z^2] = c_5x^5 + c_4x^4 + c_3x^3 + c_2x^2 + c_1x + c_0$$

Experiment Results

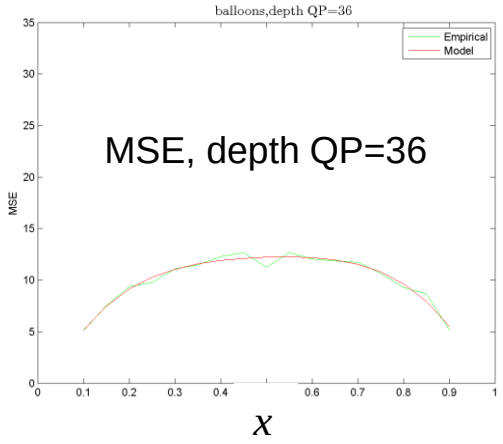
- Test video sequences: kendo, balloons, champagne, etc.
- Depth images in reference views are quantized with several QP: 36, 40, 44, 48

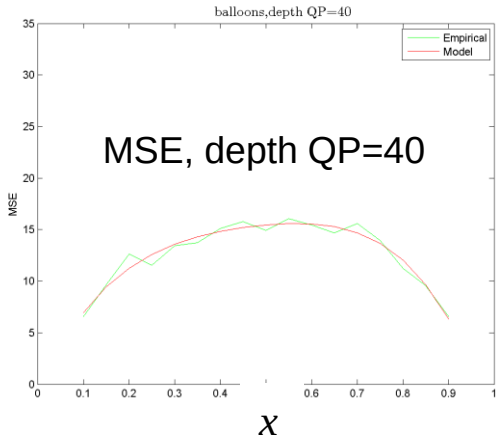
Verification of linear model for PDS with kendo sequence under QP=48:

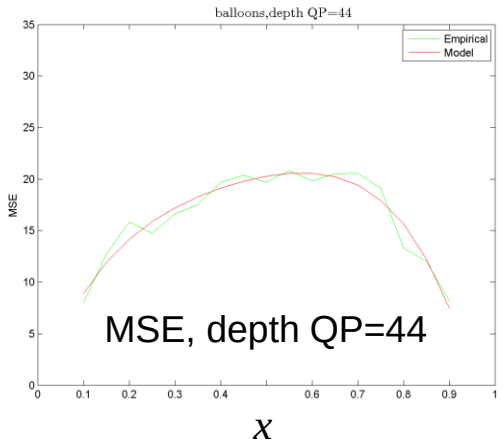


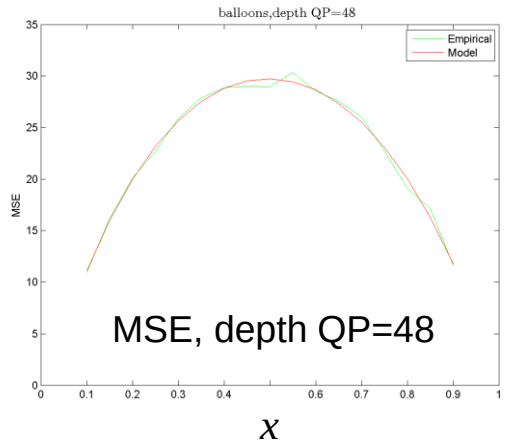
Experiment Results

-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









Empirical

Model

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

Backup

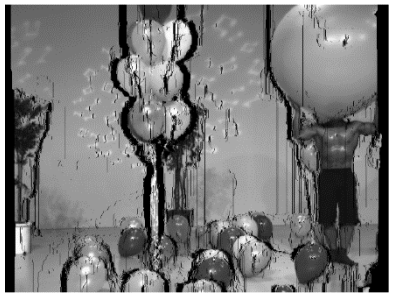
Contributions

- 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

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



Cluster 1
Overlapping region

Cluster 2, situation 3
Disocclusion and boundary region

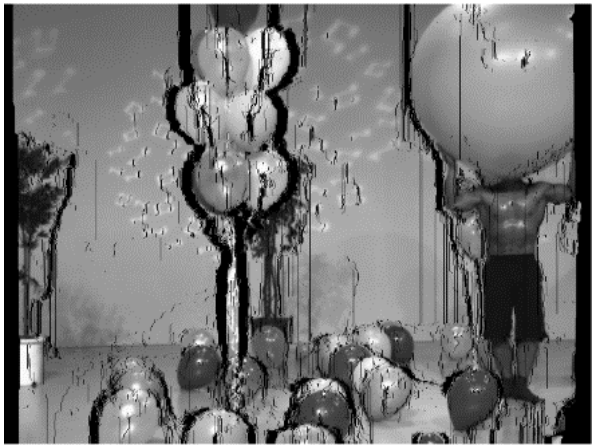
Cluster 3, situation 6
Disparity error region

Cluster 4,situation 8
Infrequent region

Approach

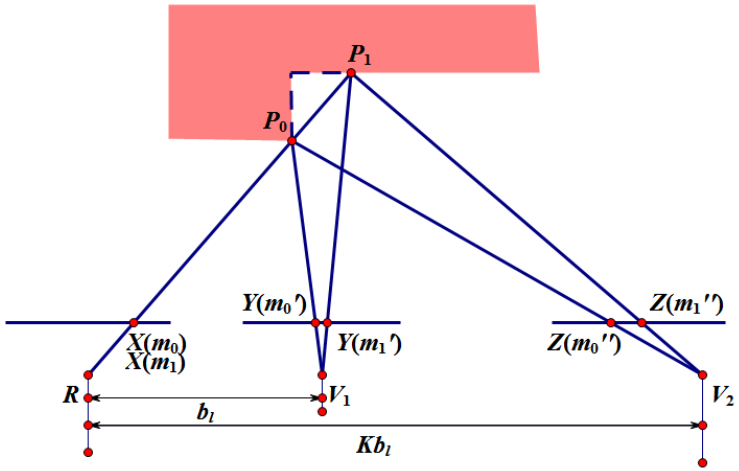
- PDS model – Overlapping region
  - Y\W contains information from left and right views, i.e., Y|Y=(1-x)Yl+ xYr
  - Cluster 1 takes place with high probability, as in practical camera capture system, adjacent cameras usually capture scenes with a lot of contents being identical

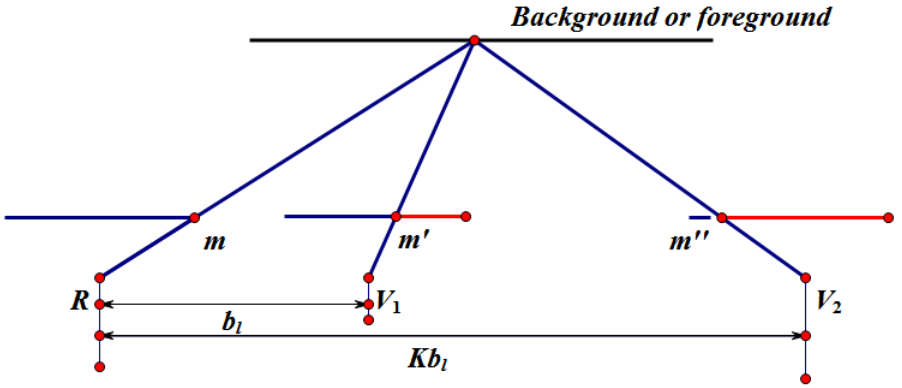
$$p_1(x) \approx constant$$



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.





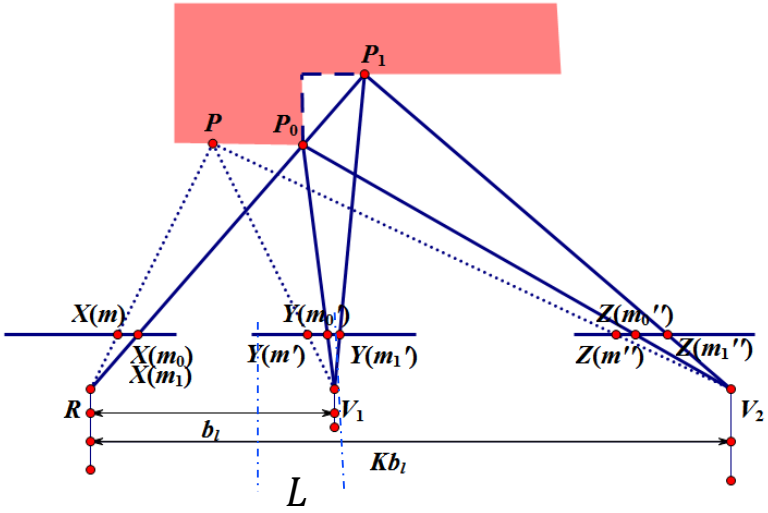
$$\frac{m_1'' - m_0''}{Kb_l} = \frac{m_1' - m_0'}{b_l}$$

$$\frac{m - m''}{Kb_l} = \frac{m - m'}{b_l}$$

$$p_3(x) = cx$$

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:

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

$$N_{V2} \approx K N_{V1}$$

$$p_6(x) = c'x$$

Approach

- PDS model – Infrequency region $(situation\ 8: Y = Y_l,\ \ W = W_r)$
  - It occurs when Yr coincidentally shifts to a new position, where the content originally belongs to Yl, however a black hole appears due to unexpected shift of Yl.
  - 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

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

Probability under Different Situations (PDS)

Distortion under Different Situations (DDS)

Y_k: warping using original depth

W_k: warping using reconstructed depth

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	$Y=Y_r$	$W=(1-x)W_l+xW_r$	$[E(Y_r-W_r)^2+(1-x)^2E(W_r-W_l)^2+2(1-x)E(Y_r-W_r)(W_r-W_l)]$
7	$Y=(1-x)Y_l+xY_r$	$W=W_r$	$[E(Y_r-W_r)^2+(1-x)^2E(Y_l-Y_r)^2+2(1-x)E(Y_r-W_r)(Y_l-Y_r)]$
8	$Y=Y_l$	$W=W_r$	$[E(Y_l-W_r)^2]$
9	$Y=Y_r$	$W=W_l$	$[E(Y_r-W_l)^2]$

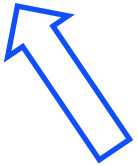
DDS Modeling

DDS are functions of these basic terms:

-Virtual view distortion caused by a single reference view: $E(Y_l\text{-}W_l)^2$, $E(Y_r\text{-}W_r)^2$

-Virtual view distortion caused by two reference views: $E(Y_r\text{-}Y_l)^2$, $E(W_r\text{-}W_l)^2$

-Cross-correlation terms: $E(Y_l\text{-}W_l)(W_l\text{-}W_r)$ etc.



Situation k	Y_k	W_k	$E[(Y-W)^2 \mid (Y=Y_k, W=W_k)]$
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	$Y=Y_r$	$W=(1-x)W_l+xW_r$	$[E(Y_r-W_r)^2+(1-x)^2E(W_r-W_l)^2+2(1-x)E(Y_r-W_r)(W_r-W_l)]$
7	$Y=(1-x)Y_l+xY_r$	$W=W_r$	$[E(Y_r-W_r)^2+(1-x)^2E(Y_l-Y_r)^2+2(1-x)E(Y_r-W_r)(Y_l-Y_r)]$
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

$$E[Z^2] = c_5x^5 + c_4x^4 + c_3x^3 + c_2x^2 + c_1x + c_0$$

c_i depends on:

- Image characteristic $E[X^2(m,n)]$, $E[X(m,n)\,X(m\text{-}1,n)]$
- Variance of depth error (depth QP)
- Distance between left and right reference views b, focal length f

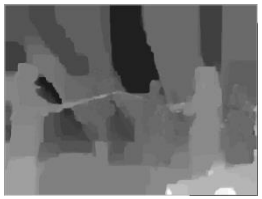
3D Video

- 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)

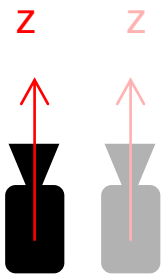
Camera array:

Depth map

Texture image

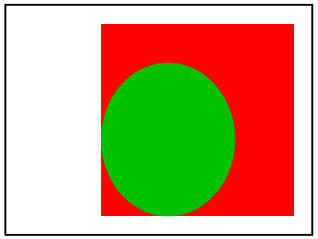


Depth-image-based Rendering

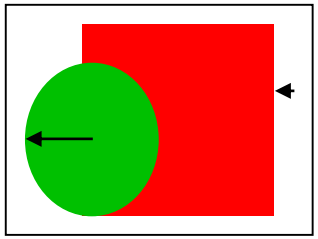


ref camera

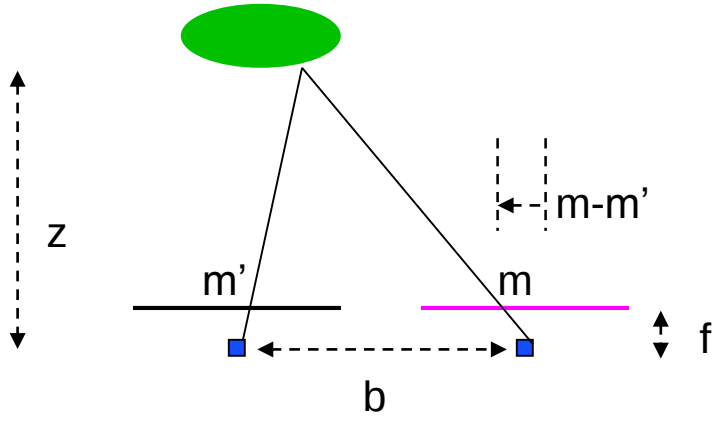
virtual camera



captured view



synthesized view



$$m - m' = \frac{fb}{z}$$

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