Image/Video Database Construction for:

1. Video Quality Assessment
2. Stereo Matching

Eddy Wu
Advisor: Prof. C.-C. Jay Kuo
Content

- Part I
  - Automatic Video Quality Assessment via Local Distortion Marking and Global Performance Ranking
    - Review of VQA Databases
    - VQA Database Construction

- Part II
  - Robust Stereo Matching via Uncalibrated Rectification and Multi-Method Fusion
    - Review of Stereo Matching
    - Considerations for Stereo Matching Database
MCL-V Database

- 12 References
- 2 Distortion Types
- 4 Distortion Levels
Total: 108 videos
Motivation

- Why Video database?
  - Design of automatic VQA metrics

- Why VQA?
  - Specifically for Netflix

Encoder Selection!
(R-D control, Inspection, etc)
Existing Databases

- Example: LIVE database\(^1\)
  - A commonly used VQA database

Source sequences (x10)
- With diversity

Distortion Types (x4) and levels (x4)
- MPEG-2
- H.264
- Transmission over IP network
- Transmission over Wireless networks

Subjective Test: ACR

Existing Databases

- More examples of available databases…

<table>
<thead>
<tr>
<th>Database</th>
<th>Year</th>
<th>SRC (# of reference videos)</th>
<th>HRC (# of test conditions)</th>
<th>Total # of Test Videos</th>
<th>Subjective Testing Method</th>
<th>Subjective Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQEG FR-TV-I</td>
<td>2000</td>
<td>20</td>
<td>16</td>
<td>320</td>
<td>DSCQS</td>
<td>DMOS (0~100)</td>
</tr>
<tr>
<td>IRCCyN/VC 1080i</td>
<td>2008</td>
<td>24</td>
<td>7</td>
<td>192</td>
<td>ACR</td>
<td>MOS (1~5)</td>
</tr>
<tr>
<td>IRCCyN/VC SD Rol</td>
<td>2009</td>
<td>6</td>
<td>14</td>
<td>84</td>
<td>ACR</td>
<td>MOS (1~5)</td>
</tr>
<tr>
<td>EPFL-PoliMI</td>
<td>2009</td>
<td>16</td>
<td>9</td>
<td>156</td>
<td>ACR</td>
<td>MOS (0~5)</td>
</tr>
<tr>
<td>LIVE Wireless</td>
<td>2009</td>
<td>10</td>
<td>16</td>
<td>160</td>
<td>SSCQE</td>
<td>DMOS (0~100)</td>
</tr>
<tr>
<td>MMSP SVD</td>
<td>2010</td>
<td>3</td>
<td>24</td>
<td>72</td>
<td>PC</td>
<td>MOS (0~100)</td>
</tr>
<tr>
<td>VQEG HDTV</td>
<td>2010</td>
<td>45</td>
<td>15</td>
<td>675</td>
<td>ACR</td>
<td>MOS (0<del>5) DMOS (1</del>5)</td>
</tr>
</tbody>
</table>

- Why existing databases do not fit our objective?

NOT REPRESENTATIVE!
Problems of Current Database

- Representativeness of a database

Universal Video Space

- Youtube Videos
- Netflix Videos
- Accessible videos
- Representative video

- Inclusive
- Diverse
Problems of Current Database

- Example of choices in LIVE
  - Not HD (1080P) quality
  - Unusual scenes
  - Lack of diversity
  - Dark?
  - Fast Motion?
Database Construction

- Steps of database construction

1. Video Collection

2. Source Video Selection
   - Inclusive
   - Diverse
   How to find features
   How to verify

3. Applying Distortions
   - Distortion type selection
   - Distortion level selection

4. Subjective test
# Video Collection

- Pool of public video sources

<table>
<thead>
<tr>
<th>Source</th>
<th>Website</th>
<th>Num. of Clips</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDVL</td>
<td><a href="http://www.cdvl.org/">http://www.cdvl.org/</a></td>
<td>&gt; 70</td>
</tr>
<tr>
<td>EBU</td>
<td><a href="http://tech.ebu.ch/testsequences">http://tech.ebu.ch/testsequences</a></td>
<td>5</td>
</tr>
<tr>
<td>Netflix</td>
<td>Private Source</td>
<td>A 7-min video</td>
</tr>
</tbody>
</table>
Source Video Selection

- Diversity check
  - Why feature selection?
    - Diversity is defined in multiple feature spaces

- Feature #1 (e.g. brightness)
- Feature #2 (e.g. Color richness)
- Feature #N
Source Video Selection

- Feature selection (18 Features in 3 groups)
  - Mainly by empirical choices from Netflix and MCL

High-level video genres
- Cartoon / Anime
- Sports
- Indoor
- Scene change
- Camera motion

Mid-level semantics
- Face
- People
- Water
- # of objects
- Obvious salience

Low-level features
- Brightness
- Contrast
- Texture
- Motion
- Color Variance
- Color Richness
- Sharpness
- Film Grain

Source Video Selection

- **Special video genres**

  - **Cartoon**
    - Simple colors, clear edges

  - **Sports**
    - Fast-moving object
    - Simple background

- **Scene change**
  - Affect human perception
  - Only consistent scene changes
Source Video Selection

- **Semantics Matters**
  - People focus on meaningful objects

- **Mid-level semantics**
  - Face
  - People
  - Water
  - # of objects
  - Obvious salience

- **Face close-up**
  - Visual Salience
  - Example: drama

- **Multi-objects**
  - Another typical scenario
Source Video Selection

- **Low-level features:**
  - Directly affect compression distortions
  - Visual Masking effects

  - **Low Brightness**
    - Example: Horror movies

  - **High texture**
    - Masking effects
Source Video Selection

- Proof of Diversity – Quantitative measures
  - Texture:
    \[ SI = \max_{time} \{ \text{std}_\text{space} [\text{Sobel}(F_n)] \} \]
  - Motions:
    \[ TI = \max_{time} \{ \text{std}_\text{space} [M_n(i,j)] \} \]

Source Video Selection

- **Proof of diversity - Subjective Review of Diversity**

<table>
<thead>
<tr>
<th>Video Properties</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cartoon/CG Animation</td>
<td></td>
</tr>
<tr>
<td>Sports</td>
<td></td>
</tr>
<tr>
<td>Indoor</td>
<td></td>
</tr>
<tr>
<td>Scene change</td>
<td></td>
</tr>
<tr>
<td>Camera motion [Still, Pan, Zoom, Misc.]</td>
<td>PS</td>
</tr>
<tr>
<td>Face close-up</td>
<td></td>
</tr>
<tr>
<td>People</td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td></td>
</tr>
<tr>
<td>Obvious salience</td>
<td></td>
</tr>
<tr>
<td>Film Grain Noise</td>
<td></td>
</tr>
<tr>
<td>Object number [many(3) ~ few(1)]</td>
<td></td>
</tr>
<tr>
<td>Brightness [high(3) ~ low(1)]</td>
<td></td>
</tr>
<tr>
<td>Contrast [high(3) ~ low(1)]</td>
<td></td>
</tr>
<tr>
<td>Texture [high(3) ~ low(1)]</td>
<td></td>
</tr>
<tr>
<td>Motion [high(3) ~ low(1)]</td>
<td></td>
</tr>
<tr>
<td>Color variance [high(3) ~ low(1)]</td>
<td></td>
</tr>
<tr>
<td>Color richness [high(3) ~ low(1)]</td>
<td></td>
</tr>
<tr>
<td>Sharpness [high(3) ~ low(1)]</td>
<td></td>
</tr>
</tbody>
</table>

- **BigBuckBunny [BB]**
- No. of subjects: 4
- Consistency: high (coarse quantization)
**Source Video Selection**

- **Proof of Diversity – Comparing with LIVE**

<table>
<thead>
<tr>
<th>Video Properties</th>
<th>MCL-V</th>
<th>LIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cartoon/CG Animation</td>
<td><img src="chart" alt="MCL-V Cartoon/CG Animation" /></td>
<td><img src="chart" alt="LIVE Cartoon/CG Animation" /></td>
</tr>
<tr>
<td>Sports</td>
<td><img src="chart" alt="MCL-V Sports" /></td>
<td><img src="chart" alt="LIVE Sports" /></td>
</tr>
<tr>
<td>Indoor</td>
<td><img src="chart" alt="MCL-V Indoor" /></td>
<td><img src="chart" alt="LIVE Indoor" /></td>
</tr>
<tr>
<td>Scene change</td>
<td><img src="chart" alt="MCL-V Scene change" /></td>
<td><img src="chart" alt="LIVE Scene change" /></td>
</tr>
<tr>
<td>Camera motion [Still, Pan, Zoom, Misc.]</td>
<td><img src="chart" alt="MCL-V Camera motion" /></td>
<td><img src="chart" alt="LIVE Camera motion" /></td>
</tr>
<tr>
<td>Face close-up</td>
<td><img src="chart" alt="MCL-V Face close-up" /></td>
<td><img src="chart" alt="LIVE Face close-up" /></td>
</tr>
<tr>
<td>People</td>
<td><img src="chart" alt="MCL-V People" /></td>
<td><img src="chart" alt="LIVE People" /></td>
</tr>
<tr>
<td>Water</td>
<td><img src="chart" alt="MCL-V Water" /></td>
<td><img src="chart" alt="LIVE Water" /></td>
</tr>
<tr>
<td>Obvious salience</td>
<td><img src="chart" alt="MCL-V Obvious salience" /></td>
<td><img src="chart" alt="LIVE Obvious salience" /></td>
</tr>
<tr>
<td>Film Grain Noise</td>
<td><img src="chart" alt="MCL-V Film Grain Noise" /></td>
<td><img src="chart" alt="LIVE Film Grain Noise" /></td>
</tr>
<tr>
<td>Object number [many(3) ~ few(1)]</td>
<td><img src="chart" alt="MCL-V Object number" /></td>
<td><img src="chart" alt="LIVE Object number" /></td>
</tr>
<tr>
<td>Brightness [high(3) ~ low(1)]</td>
<td><img src="chart" alt="MCL-V Brightness" /></td>
<td><img src="chart" alt="LIVE Brightness" /></td>
</tr>
<tr>
<td>Contrast [high(3) ~ low(1)]</td>
<td><img src="chart" alt="MCL-V Contrast" /></td>
<td><img src="chart" alt="LIVE Contrast" /></td>
</tr>
<tr>
<td>Texture [high(3) ~ low(1)]</td>
<td><img src="chart" alt="MCL-V Texture" /></td>
<td><img src="chart" alt="LIVE Texture" /></td>
</tr>
<tr>
<td>Motion [high(3) ~ low(1)]</td>
<td><img src="chart" alt="MCL-V Motion" /></td>
<td><img src="chart" alt="LIVE Motion" /></td>
</tr>
<tr>
<td>Color variance [high(3) ~ low(1)]</td>
<td><img src="chart" alt="MCL-V Color variance" /></td>
<td><img src="chart" alt="LIVE Color variance" /></td>
</tr>
<tr>
<td>Color richness [high(3) ~ low(1)]</td>
<td><img src="chart" alt="MCL-V Color richness" /></td>
<td><img src="chart" alt="LIVE Color richness" /></td>
</tr>
<tr>
<td>Sharpness [high(3) ~ low(1)]</td>
<td><img src="chart" alt="MCL-V Sharpness" /></td>
<td><img src="chart" alt="LIVE Sharpness" /></td>
</tr>
</tbody>
</table>
Distortion Design

- Two chosen distortion types
  1. H.264 VBR compression only
  2. Rescaling to 720P with H.264 VBR compression
Distortion Design

- Artifacts from Compression
  - Blocky
  - Staircase
  - Ghost

- Artifacts from scaling + compression
  - Original
  - Blurred
Distortion Level Selection

- A video database contains
  - Multiple sources with **distinguishable** distortion levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Quality</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5</td>
<td>Excellent</td>
<td>Uncompressed references</td>
</tr>
<tr>
<td>L4</td>
<td>Good</td>
<td>Distortion not perceivable by most subjects</td>
</tr>
<tr>
<td>L3</td>
<td>Fair</td>
<td>Distortion perceivable, but not annoying</td>
</tr>
<tr>
<td>L2</td>
<td>Poor</td>
<td>Preserved most structure and texture</td>
</tr>
<tr>
<td>L1</td>
<td>Bad</td>
<td>Lowest acceptable quality</td>
</tr>
</tbody>
</table>

- Available Bitrate range 300 Kbps ~ 10 Mbps
  How to select the levels from a wide range of bitrates?
Distortion Level Selection

- Perceptual vs Quantitative?
  - Mainly by Visual Check
  - Assisted with Quantitative measurements

Internal Test:
- # of subjects: 5
- “Preset level”
  - Assist Pair-wise comparison algorithm
  - Final scores are still decided by subjective test

Level 4
5032 K bps

Level 4
1369 K bps

- Assisted with Quantitative measurements
Conclusion to Part I

- Database construction steps
  - Representative Source Selection
  - Diversity of Source Video
  - Distortion Type and Level Selection

- Future work (done, on-going)
  - Subjective test (pair-wise comparison)
  - VQA metrics design
Part II

- Robust Stereo Matching via Uncalibrated Rectification and Multi-Method Fusion

  - Introduction to Stereo Matching
  - Stereo Matching Databases
  - Objective
  - Considerations For Database Construction
Stereo Vision

- How human perceive 3D from stereo images
What is Stereo Matching

- Generate DENSE disparity map from stereo images

Left view

Right view

P(Lx, Ly)

P'(Rx, Ry)

Rectified (Ly = Ry)

Dense disparity map
Disparity = Lx - Rx
Stereo Matching

Applications of depth map

- Surveillance
- Car, pedestrian detection
- Blind guide
- Distance measurement
- Interactive conference
- 3D face reconstruction
- Augmented reality
- Virtual try-on

- Gesture control
- Gaming
- Medical application
- Shopping, public facilities

3D Vision Applications

Human-computer interface

3D-aided / interactive systems

Environmental detection

- Free-Viewpoint TV
- Auto-stereoscopic display
- Depth adaptation
- Retargeting

- Surveillance
- Car, pedestrian detection
- Blind guide

University of Southern California
### Stereo Matching

- **Comparison between depth sensing technologies**

<table>
<thead>
<tr>
<th></th>
<th>ToF camera</th>
<th>Stereo vision</th>
<th>Structured light</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correspondence problem</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Extrinsic calibration</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Auto-Illumination</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Untextured surfaces</td>
<td>Good</td>
<td>Bad</td>
<td>Good</td>
</tr>
<tr>
<td>Depth range</td>
<td>Light power dependent</td>
<td>Baseline dependent</td>
<td>Light power dependent</td>
</tr>
<tr>
<td>Image resolution</td>
<td>QVGA</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Strong light environment</td>
<td>Bad</td>
<td>Good</td>
<td>Bad</td>
</tr>
<tr>
<td>Low light environment</td>
<td>Good</td>
<td>Bad</td>
<td>Good</td>
</tr>
<tr>
<td>Process 3D content</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

**Time-of-flight**

**Structured light**
### Development of Stereo Matching

- **Middlebury Database, 2002**

![Stereo Matching Table](image)

| Algorithm         | Avg. | Rank |  
|-------------------|------|------|---
| TSGO [143]        | 9.3  | 1.3  | 1  |
| ADCensus [82]     | 12.7 | 1.7  | 4  |
| AdaptingBP [15]   | 15.9 | 1.25 | 1  |
| CoopRegion [39]   | 16.5 | 1.6  | 4  |
| RDP [87]          | 21.4 | 1.39 | 1  |
| MultiREF [123]    | 21.6 | 1.55 | 1  |
| DoubleBP [34]     | 22.3 | 1.12 | 6  |
| MDPM [140]        | 22.3 | 1.59 | 21 |
| OutlierConf [16]  | 22.8 | 1.43 | 1  |
| SegAggr [146]     | 24.2 | 1.39 | 3  |
| AdaptiveGF [127]  | 26.5 | 1.53 | 15 |

![Table Image](image)

Development of Stereo Matching

- Stereo Matching Works

Publication compared in Middlebury
(Total: 147 after 2002)

Publication #

Year

Matching cost computation
Cost aggregation
Energy Minimization
Disparity refinement

Local Global
Development of Stereo Matching

- Matching Cost Computation

**Sum of Absolute Differences**

\[ SAD = \iint_{W} |I'(x, y) - I(x, y)| \, dx \, dy \]

**Sum of Square Differences**

\[ SSD = \iint_{W} [I'(x, y) - I(x, y)]^2 \, dx \, dy \]

**Census**

\[ C_1(i, j) = (I(x + i, y + j) > I(x, y)) \]

| 125 | 126 | 125 | 0 | 0 | 0 |
| 127 | 128 | 130 | 0 | 1 |   |
| 129 | 132 | 135 | 1 | 1 | 1 |

→ [00001111]

only compare bit signature
Development of Stereo Matching

- Cost Aggregation
  - Window, Adaptive window, etc
- Energy minimization

![Diagram showing Forward cost updating, Backtracking, and Dynamic Programming](image)

- Winner-Take-All (Local)
- Dynamic Programming (Global)
Development of Stereo Matching

- Energy minimization

Graph Cut (Global)
Development of Stereo Matching

- **Available databases (with depth)**

<table>
<thead>
<tr>
<th>Database</th>
<th># of data</th>
<th>Resolution</th>
<th>Calibration data</th>
<th>Depth generation</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middlebury</td>
<td>4 (+35)</td>
<td>Varies (low)</td>
<td>Rectified</td>
<td>CG / Manually</td>
<td>Some with M-views</td>
</tr>
<tr>
<td>MPEG-3DVC</td>
<td>~20</td>
<td>Varies (high)</td>
<td>Some provided</td>
<td>S.M. / Manually</td>
<td>Some with M-views</td>
</tr>
<tr>
<td>KITTI</td>
<td>194</td>
<td>1382 x 512</td>
<td>Provided</td>
<td>Device</td>
<td>Vehicle app. only</td>
</tr>
</tbody>
</table>

- **Stereo databases**
  - Syntim, HCI, EISATS, CMU CIL, CMU VASC, JISCT, 6D-Visioin, etc

**Images:**
- 3DVC
- KITTI
Problems of Existing Databases

- Problems in Middlebury database
  - Only 4 images for benchmarking
  - Limited Resolution
  - No Challenging (Content too perfect, no distortions)

Cones (450x375)  Teddy (450x375)  Tsukuba (384x288)  Venus (433x348)
Problems of Existing Databases

- Examples of what happens in real situations

Depending on Applications
Problems of Existing Databases

- Examples of depth quality degradation
  - Simulating camera rotation of 2 degrees

original

Camera rotation
Problems of Existing Databases

- Contents that affect depth quality

<table>
<thead>
<tr>
<th>Content factors:</th>
<th>Camera factors:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object amount &amp; allocation</td>
<td>Image resolution</td>
</tr>
<tr>
<td>Object shape (sticks)</td>
<td>Camera noise</td>
</tr>
<tr>
<td>Texture-ness</td>
<td>Camera focus</td>
</tr>
<tr>
<td>Texture repeated-ness (x/y)</td>
<td>Shutter speed</td>
</tr>
<tr>
<td>Object movement (blurring)</td>
<td>Calibration mis-alignment (focal, shift, rotation, lens distortion)</td>
</tr>
<tr>
<td>Background complexity</td>
<td></td>
</tr>
<tr>
<td>Scene illumination</td>
<td></td>
</tr>
</tbody>
</table>

Depth Estimation method:
- Local-based Stereo Matching (WTA, DP, ...)
- Global-based Stereo Matching (SGBM, GC, BP, ...)

38/46
Objective of The Research

Challenging database

Uncalibrated Rectification

Stereo Matching

Disparity Map

With Camera Distortions, etc
Considerations for Database

- Database consist of
  - Stereo Image Images (Distorted, with diversity)
  - Evaluation metrics (With Ground-Truth disparity map)

- Ways of generating ground-truth
  - Synthetic database
  - Current database + distortion
  - Acquisition with stereo cameras
    - Low-cost depth sensor
    - High-end devices
    - Manually editing
    - Subjective test
Considerations for Database

- **Synthetic database (e.g. CG)**
  - **Pros**
    - Easy to control the environment
    - Easy to generate depth map
  - **Cons**
    - Environment is not real enough

- **Current database + distortions**

**Artificial vs. Realistic**

3D tree in frontal view  Tilted (fake)
Considerations for Database

- Stereo Camera + Low cost depth sensor

Difficulties:
1. Alignment
2. Low depth quality and resolution
3. Disparity vs depth conversion
Considerations for Database

- Stereo cameras + High-end devices
  - Problems: cost, etc
Considerations for Database

- Probable strategy
  - Computer graphics
  - Stereo Camera + Subjective Test
    - Evaluate the quality on rendered view instead of depth map
Conclusion of Part II

- Review of the Stereo Matching
- The problem of current database
- Considerations for database construction

- On-going:
  - Databases build-up
  - Develop of uncalibrated rectification algorithm
The End

Thank You