Automatic Stereoscopic Video Synthesis from a Casual Monocular video

Lin Zhong, Sen Wang, Dimitris Metaxas

Contributions:

- An automatic stereoscopic videos synthesis framework is introduced, which can convert casually captured 2D videos of complex scenes into 3D videos.

- Proposed a novel Bayesian based view synthesis (BVS) approach, which could handle these challenges in synthesizing high quality virtual view.
  - geometry information estimated from monocular videos are not so accurate, which can't meet the requirement for current image-based rendering (IBR) methods.
  - Image quality of synthesis virtual views degrade greatly from the occlusion/disocclusion problems.

System framework

Experiments

(Two public video data, Two self-taken video data, One clip from famous movie)

Bayesian based view synthesis

Objective function (maximization):

\[
p(S_F^c, \{M_F\}, \{f_C\}) \frac{p(f_C)}{p(S_F^c, M_F)}\prod_{i=1}^{N} p(f_C_i | S_F^c, M_F, D_I)
\]

Rewrote by Bayesian (maximization):

\[
\prod_{i=1}^{N} p(M_F | S_F^c, f_C_i) \cdot p(S_F^c)\prod_{i=1}^{N} p(f_C_i | S_F^c, M_F, D_I)
\]

1) \(p(M_F | S_F^c, f_C_i)\) is the color-consistency prior.
2) \(p(S_F^c)\) is the prior on the synthesized virtual view.
3) \(p(f_C_i | S_F^c, M_F, D_I)\) is the confidence on the correspondences.

Convert to energy lost function (Minimization):

\[
\sum_{i=1}^{N} \beta_i \min\left( |M_F_i(f_C_i, x, y) - S_F^c(x, y)| + \alpha_j \right)
\]

\[
+ \lambda \sum_{(x, y)} \left( |S_F^c(x, y) - AegN(S_F^c(x, y))| \right)
\]

Table 1: Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M_F)</td>
<td>Input original monocular frame sequence, (i = 1, ... , N)</td>
</tr>
<tr>
<td>(M_P)</td>
<td>Estimated camera parameters for (M_F)</td>
</tr>
<tr>
<td>(D_I)</td>
<td>Recovered dense depth map for (M_F)</td>
</tr>
<tr>
<td>(LSF, RSP)</td>
<td>Estimated camera parameters for stereo left/right views of (M_F)</td>
</tr>
<tr>
<td>(f_C, BSP, M_F)</td>
<td>Synthesized left/right views for frame (M_F) with image-based rendering (IBR) methods</td>
</tr>
</tbody>
</table>

Structure from motion:

- estimate (1) camera parameters, (2) sparse point clouds, from the 2D video.

Depth map recovery:

- (1) Stereo pair matching.
- (2) Rough disparity map by match points searching in stereo pairs.
- (3) Refine disparity/depth map by building graph model and information inference (belief propagation).

Bayesian based view synthesis:

- synthesis virtual views of left/right eye frame sequence (shown in Red/Yellow) with the estimated 3D geometry information (camera parameters, depth maps).

Synthesized virtual images

Synthesized 3D frames/videos