Report on generation of video plus depth data base

Aljoscha Smolic  Yongzhe Wang  Marcus Müller  Peter Kauff
Report on generation of video plus depth data base

Aljoscha Smolic, Yongzhe Wang, Marcus Müller, Peter Kauff

Abstract: Availability of suitable test data is a crucial condition for research and development in video processing and coding. In order to lay the fundamentals for further research and development in the project, a database of video plus depth data was created. Depth was estimated using Hybrid Recursive Matching (HRM) previously developed at FHG-HHI for the left view of each stereo pair. Then, the corresponding right view was rendered from the left view and the estimated depth. For quality assessment the visual impression was evaluated inspecting synthesized right views and the stereo impression on small displays for mobile applications. Artefacts are visible in the synthesized right views. But due to masking effects with the original left view, they are almost not noticeable when watching stereo on a small display.

Keywords: test data, video plus depth, depth estimation, database
Executive Summary

Availability of suitable test data is a crucial condition for research and development in video processing and coding. In order to lay the fundamentals for further research and development in the project, a database of video plus depth was created. This was done for the data from the stereo video database described in D2.1. Depth was estimated using Hybrid Recursive Matching (HRM) previously developed at FHG-HHI for the left view of each stereo pair. Then, the corresponding right view was rendered from the left view and the estimated depth. For quality assessment the visual impression was evaluated inspecting synthesized right views and the stereo impression on small displays for mobile applications.

Artefacts are visible in the synthesized right views. This is an inherent problem of video plus depth approach for stereo. Depth estimation can never be perfect and can only be solved up to a residual error probability. The task is to minimize errors and to ensure that they have a minimum effect on the visual quality. Due to masking effects with the original left view, artefacts are almost not noticeable when watching stereo on a small display, which was found by visual inspection. Further improvement will be task of the project in cooperation with partners, i.e. post processing, filtering of depth maps, e.g. in cooperation with WP5.
Table of Contents

1. Introduction ........................................................................................................................................... 4
2. Depth Estimation Algorithms.................................................................................................................. 5
   2.1. Block-Recursion .................................................................................................................................. 6
   2.2. Pixel Recursion .................................................................................................................................. 7
3. Results ...................................................................................................................................................... 9
   3.1. Sequences with full copyright .......................................................................................................... 9
       3.1.1. FhG-HHI 3DV data, 4 sequences .............................................................................................. 9
       3.1.2. FhG-HHI Bullinger .................................................................................................................... 13
       3.1.3. KUK data, 9 sequences ............................................................................................................. 14
       3.1.4. TU Berlin, 2D video converted to 3D, 1 sequence .................................................................... 19
   3.2. Sequences without copyright ............................................................................................................ 20
       3.2.1. ETRI, 5 sequences ...................................................................................................................... 20
       3.2.2. Nagoya University MVC data, 2 sequences ............................................................................. 24
       3.2.3. Nagoya University FTV-3DV data, 3 sequences ....................................................................... 26
       3.2.4. ETRI 3DV data, 2 sequences .................................................................................................... 29
       3.2.5. GIST 3DV data, 1 sequence ...................................................................................................... 31
4. Conclusions .............................................................................................................................................. 32
1 Introduction

Availability of suitable test data is a crucial condition for research and development in video processing and coding. In order to lay the fundaments for further research and development in the project, a database of video plus depth was created, starting from the data in the stereo video database described in D2.1. Depth was estimated using Hybrid Recursive Matching (HRM) previously developed at FHG-HHI for the left view of each stereo pair. An overview of the HRM algorithms is given in section 2.

Section 3 presents the results. For each test sequence a left view and a right view are shown. The estimated depth data for the left view are shown as well. Then, the corresponding right view was rendered from the left view and the estimated depth using depth image based rendering. For quality assessment the visual impression was evaluated inspecting synthesized right views and the stereo impression on small displays for mobile applications. Examples are given in section 3.

Artefacts are visible in the synthesized right views. This is an inherent problem of video plus depth approach for stereo. Depth estimation can never be perfect and can only be solved up to a residual error probability. The task is to minimize errors and to ensure that they have a minimum effect on the visual quality. Due to masking effects with the original left view, artefacts are almost not noticeable when watching stereo on a small display, which was found by visual inspection.
2 Depth Estimation Algorithms

This section gives an overview of the Hybrid Recursive Matching (HRM) approach developed for high quality depth estimation at FHG-HHI in previous projects. It was proven to provide high quality, high reliable, time consistent depth video, while being capable for real-time performance. More details can be found in:


The main idea of the hybrid recursive stereo matching algorithm is to unite the advantages of block-recursive disparity matching and pixel-recursive optical flow estimation in one common scheme. The block-recursive part assumes that depth does not change significantly from one image to the next and that depth is nearly the same in the local neighbourhood. Obviously this assumption can not be fulfilled in all image areas - especially not in areas with high motion and at depth discontinuities. To update the results of the block-recursive stage in these areas, the pixel recursive stage calculates the optical flow by analysing gradients and grey value differences.

In more detail, the structure of the whole algorithm can be outlined in three subsequent processing steps (see Fig.1):

- three candidate vectors are evaluated for the current block position by recursive block matching;
- the candidate vector with the best result is chosen as the start vector for the pixel-recursive algorithm, which yields an update vector;
- the final vector is obtained by testing if the update vector from the pixel recursive stage is of higher quality than the start vector from the block-recursive one.

Compared to former approaches, this new hybrid recursive matching has two main advantages. The recursive structure speeds up the analysis dramatically. The combined choice of spatial and temporal candidates yields spatially and temporally consistent disparity fields due to an
efficient strategy of testing particular vector candidates. The latter aspect is very important to avoid temporal inconsistencies in disparity sequences, which may cause strongly visible and very annoying artefacts in virtual views synthesised on the basis of these disparities.

2.1 Block-Recursion

The idea of the block recursion is to use information of both the previous image and the spatial neighbourhood. This kind of recursion forces temporal and spatial consistency and it additionally reduces the local search range to a few pixels as known from many other matching algorithms. Usually, calculation of three matching scores per considered pixel is fully sufficient to achieve results comparable to a full search method. In our algorithm these three matching scores are calculated by using three candidates, which are defined by disparities from the previous and the current image. The following spatial and temporal candidates are tested for this purpose:

- A horizontal predecessor, taken from the left or right position in the actual frame.
- A vertical predecessor, taken from the bottom or top position in the actual frame.
- A temporal predecessor, taken from the same position in the previous frame.

It is obvious that the spatial candidates can only be determined from disparities, which have already been calculated. To determine the spatial candidate vectors in the most isotropic way, the video frames are scanned in two interleaved meander paths. At first, the odd lines are scanned from top-to-bottom and then the even lines are processed from bottom-to-top. This scanning scheme changes its order from frame to frame and the scan direction in consecutive lines switches between left-to-right and right-to-left. Moreover, as the HRM algorithm is able to cope with arbitrarily shaped video objects (e.g. the segmented portrayal of a confrere), the interleaved meander is adapted to the binary shape of the video object (see Fig. 2).

![Fig. 2. Meander scan for arbitrarily shaped video objects (left: even frames; right: odd frames)](image)

Intensive experiments have proven that this alternating and interleaved meander scan path leads to a faster convergence of disparity estimation, especially at moving edges where depth discontinuities occur. The interleaved meander leads to four different arrangements of the three candidates, which are shown in Fig. 3.

![Fig. 3. Candidates for the four different scan directions](image)
These candidates are tested to find the best match between the current pixel position in one image of a stereo rig (e.g. the left one) and the corresponding pixel position in the other image (e.g. the right one). The following shape-driven displaced block difference (DBD) is taken as criterion.

\[
DBD(d) = \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} s(x,y) \cdot \left| I_i(x,y) - I_i(x+d_i, y+d_i) \right|
\]

with

\[
s(x,y) = \begin{cases} 
1 & \text{if (x, y) inside object} \\
0 & \text{if (x, y) outside object}
\end{cases}
\]

This matching criterion is a shape adaptive version of the sum of absolute differences (SAD), which is used by many real-time algorithms. Other matching criteria known from literature, like normalized cross correlation (NCC), usually provide better results, but they are computationally more expensive than SAD.

A further feature, which is very important for the real-time capability, is that the HRM algorithm does not use any local search around the candidate vector in the block-recursive stage. However, this feature also requires an update mechanism. Obviously, without any update the block recursive stage is not able to follow spatio-temporal deviations in the disparity field. It is therefore the task of the pixel recursion to act as a permanent update process. For this purpose it injects one new candidate vector per pixel position into the block recursion.

### 2.2 Pixel Recursion

Pixel-recursive disparity estimation is a low-complexity method, which calculates dense displacement fields using a simplified optical flow approach. Following the principal of the optical flow, an update vector \( d \) is calculated on the basis of spatial gradients and gradients between the two frames. The gradient between the frames is approximated by the displaced pixel difference (DPD) given by corresponding points in the left and right images.

\[
d(x,y) = d_i - DPD(d_i, x, y) \cdot \frac{\nabla I(x,y)}{\| \nabla I(x,y) \|^2}
\]

with

\[
DPD(d_i, x, y) = \left[ I_L(x, y) - I_R(x+d_i, y+d_i) \right] \cdot \left[ \frac{\nabla I(x,y)}{\| \nabla I(x,y) \|^2} \right]
\]

To meet real-time requirements, the gradient calculation is approximated as follows:

\[
d(x,y) = d_i - DPD(d_i, x, y) \cdot \left[ \frac{\nabla I(x,y)}{\| \nabla I(x,y) \|^2} \right]
\]

\[
u_x = \begin{cases} 
0 & \text{if } \frac{\partial I(x,y)}{\partial x} < \Theta \\
\left[ \frac{\partial I(x,y)}{\partial x} \right] & \text{else}
\end{cases}
\]

\[
u_y = \begin{cases} 
0 & \text{if } \frac{\partial I(x,y)}{\partial y} < \Theta \\
\left[ \frac{\partial I(x,y)}{\partial y} \right] & \text{else}
\end{cases}
\]

\[
\frac{\partial I(x,y)}{\partial x} = \frac{I(x+1,y) - I(x-1,y)}{2}, \quad \frac{\partial I(x,y)}{\partial y} = \frac{I(x,y+1) - I(x,y-1)}{2}
\]
The threshold value in equation (4) is usually set to a value of 2 or 3. It decreases the sensitivity of pixel-recursion to noise in unstructured image regions. With rectified images, pixel recursion is only used for the x-component of the disparity vector in equations (3) while setting \( u_y \) always to zero in (4).

Multiple pixel-recursive processes are started at every first pixel position of the odd lines in a block around the considered pixel (see Fig. 4 with a 4x4 block). Usually, in optical flow applications the disparity vector is improved iteratively at every pixel position and the iteration is finalized by a threshold criterion. However, in our approach we only apply one iteration step to each pixel position to guarantee real-time processing. As a result, each “iteration process” ends up with one incremental update vector per pixel, which is added to the initial vector to obtain the local update vector. The local update vector of the previous pixel is then taken as initial vector for the next pixel. The very first position of every pixel recursive processing path is initialised with the start vector from the block recursive stage.

![Figure 4: Outline of the pixel-recursion scheme](image)

Finally, after processing all paths of multiple pixel-recursions, the vector with the smallest DPD among all pixel-recursive processes is taken as the final update vector. This output of pixel recursion is not necessarily the last pixel position at the end of one of the scanning paths. Often, the pixel-recursion converges over some pixels and than diverges again. Thus, the vector with the smallest DPD can be taken from an intermediate position somewhere in the middle of a scanning path. After selecting the final update vector, its DBD is calculated and compared to the DBD of the start vector from block recursion. If the DBD of the update vector is smaller than the one of the start vector, the update vector is chosen as the final output vector, otherwise the start vector is retained.

Pixel recursion plays an important role in areas with depth discontinuities and fast motion because in this case the probability of completely wrong candidates is extremely high in the block-recursive stage. This special conditions hold for boundaries of fast moving foreground objects, but also for the beginning of the sequence or for scene cuts. In this sense, pixel recursion considerably improves the spatial and the temporal transient behaviour of the whole algorithm, whereas block recursion is mainly responsible for spatial and temporal smoothing of disparities in areas of homogenous depth and moderate motion.
3 Results

In this section we present the results for depth estimation and view synthesis. Depth was estimated using HRM from original left and right view of a stereo pair as described in previous section for the left view. Then, a virtual right view was rendered from the original left view and the associated, estimated depth map. In the following examples in this section we see original left and right view of a stereo pair, estimated depth for the left view and the synthesized right view.

In general the depth maps are not perfect. This is an inherent problem of the video plus depth approach. However, we found that the estimated depth maps are temporally and spatially smooth and stable. This is due to the recursive structure of the HRM algorithm. Artefacts are visible in synthesized right views. But when watching stereo video including the original left view and the synthesized right view on a small display for mobile terminals, artefacts are almost not noticeable. Stereo masking effects known as binocular suppression theory lead to a good overall impression. Thus the overall quality of estimated depth maps is sufficient.

3.1 Sequences with full copyright

This section presents the results for sequences with full copyright available to the project.

1.1.1. FhG-HHI 3DV data, 4 sequences

<table>
<thead>
<tr>
<th>Resolution</th>
<th>XGA, 1024x768</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame rate</td>
<td>16.67 Hz</td>
</tr>
<tr>
<td>Length</td>
<td>100 frames, 6 s</td>
</tr>
<tr>
<td>Video format</td>
<td>YUV4:2:0, uncoded</td>
</tr>
<tr>
<td>Stereo format</td>
<td>multiview, 16 cameras with stereo distance</td>
</tr>
<tr>
<td>Copyright</td>
<td>full</td>
</tr>
</tbody>
</table>

Content, characteristics, quality

**Book_Arrival, Leaving_Laptop, Door_Flowers**

Indoor, studio/soap-type, 2 persons interacting in a room with various requisites, moderate object motion, no camera motion, high detail, complex depth structure, studio light

**Book_Arrival:**

![Book_Arrival Image]
Leaving_Laptop:
Door_Flowers:

Alt_Moabit
Outdoor, street scenery, traffic, significant object motion, no camera motion, medium detail, medium depth structure, natural light
1.1.2. *FhG-HHI Bullinger*

Resolution 960x720
Frame rate 25 Hz
Length 1:04 min
Video format mp4, H.264 coded, left/right stereo in one image
Stereo format left, right
Copyright full

Content, characteristics, quality

*Bullinger*

Indoor, news/portrait-type, persons talking to camera, moderate object motion, no camera motion, moderate detail, simple depth structure, studio light, professional production, with audio
1.1.3. KUK data, 9 sequences

These sequences were kindly provided by KUK Filmproduktion, a professional production company specialized in 3D movie production based in Munich. The material is in professional quality in HD format. 

http://www.kuk-film.de/

Resolution HD, 1920x1080
Frame rate 30 Hz
Length 3-8 s
Video format TIF, avi uncoded, avi DivX coded, separate files
Stereo format left, right
Copyright full by permission of KUK Filmproduktion http://www.kuk-film.de/

Content, characteristics, quality

Hands
Indoor, close-up with black background, complex object motion (water fountain), no camera motion, complex detail (transparency, reflections), complex depth structure, studio light, professional production, very challenging for any type of video processing

Flower1, Flower2, Flower3, Grasshopper, Caterpillar, Snail
Indoor, close-up with black background, simple to complex object motion, no or simple camera motion, medium detail, medium depth structure, studio light, professional production
**Flower1:**

![Flower1 Image](image1)

**Flower2:**

![Flower2 Image](image2)
**Flower3:**

![Flower3](image1)

**Grasshopper:**

![Grasshopper](image2)
Caterpillar:

Snail:

Horse
Outdoor, cottage with horse, simple object motion, simple camera motion, high detail, medium complex depth structure, natural light, professional production

Car
Outdoor, street/traffic scene shot from driving car, complex object motion, complex camera motion, high detail, complex depth structure, natural light, professional production
1.1.4. TU Berlin, 2D video converted to 3D, 1 sequence

This sequence is kindly provided by the Technical University of Berlin, Communication Systems Group. It was generated from a normal 2D video by 2D/3D conversion.

http://www.nue.tu-berlin.de/people/knorr/stereoscopic.html

Resolution 720x576
Frame rate 25Hz
Length 2:13 min
Video format avi, DivX coded, separate files
Stereo format left, right
Copyright full by permission of TU Berlin, http://www.nue.tu-berlin.de

Content, characteristics, quality

Charlotte_Potsdam

Outdoor, landscape/sightseeing consumer captured type, no object motion, difficult camera motion (handheld, shaking, consumer captured), high detail, complex depth structure, natural light, 2D/3D conversion
3.2 **Sequences without copyright**

This section presents the results for sequences available to the project, but without copyright. Most of these data are official MPEG testdata to be used for development of the new 3D video coding standards.

1.1.5. **ETRI, 5 sequences**

These excellent stereo video sequences were kindly provided to the Mobile3DTV project by the Korean research institute “Electronics and Telecommunications Research Institute (ETRI)” [http://www.etri.re.kr/](http://www.etri.re.kr/). The Mobile3DTV project keeps an informative collaboration with ETRI, since they are active in the same research area.

- **Resolution**: 320x240
- **Frame rate**: 30 Hz
- **Length**: 240-450 frames, 8-15 s
- **Video format**: YUV4:2:0, uncoded
- **Stereo format**: left, right
- **Copyright**: can be used in Mobile3DTV project

**Content, characteristics, quality**

**Diving**

Indoor, sports, complex object motion, complex camera motion, slow motion effect, high detail, complex depth structure, natural light
**Mountain**

Outdoor, camera move over landscape, partially no object motion, partially simple/medium object motion (water, reflections), simple camera motion, high detail, medium complex depth structure, natural light

![Mountain images](image)

**Performance**

Outdoor, large scale show, extremely complex object motion, no camera motion, extremely high detail, complex depth structure, show light including extreme effects, very challenging for any type of video processing
Soccer 1

Outdoor, sports (football), complex object motion, complex camera motion, scene cut (close up to total view), high detail, complex depth structure, natural stadium light
Soccer 2
Outdoor, sports (football), complex object motion, complex camera motion, high detail, complex depth structure, natural stadium light
1.1.6. Nagoya University MVC data, 2 sequences

These multiview video sequences with stereo spacing were provided to MPEG for development of the MVC standard by Nagoya University, Tanimoto Lab.

http://www.tanimoto.nuee.nagoya-u.ac.jp/mpeg/mpeg_ftv.html

Resolution	VGA, 640x480
Frame rate	30 Hz
Length	300 frames, 10 s
Video format	YUV4:2:0, uncoded
Stereo format

- Rena: linear multiview, 16 cameras with stereo distance
- Akko&Kayo: multiview in 5x3 star arrangement, stereo distance

Copyright available to MPEG

Content, characteristics, quality

**Rena**

Indoor, studio type, medium complex object motion, no camera motion, medium detail, medium complex depth structure, studio light

**Akko& Kayo**
Indoor, studio type, medium complex object motion, no camera motion, medium detail, medium complex depth structure, studio light
1.1.7. Nagoya University FTV-3DV data, 3 sequences

These multiview video sequences with stereo spacing were provided to MPEG for development of the FTV/3D Video standard by Nagoya University, Tanimoto Lab.

http://www.tanimoto.nuee.nagoya-u.ac.jp/mpeg/mpeg_ftv.html

Resolution 1280x960
Frame rate 30 Hz
Length 300 frames, 10 s
Video format YUV4:2:0, uncoded
Stereo format multiview, 80 cameras with stereo distance
Copyright available to MPEG

Content, characteristics, quality

**Champagne_tower**

Indoor, studio type, complex object motion (fluids, reflections, transparency), no camera motion, high detail, medium complex depth structure, studio light

**Dog**
Indoor, studio type, medium complex object motion, no camera motion, medium detail, medium complex depth structure, studio light

Pantomime
Indoor, studio type, medium complex object motion, no camera motion, medium detail, medium complex depth structure, studio light
1.1.8. ETRI 3DV data, 2 sequences

These multiview video sequences with stereo spacing were provided to MPEG for development of the FTV/3D Video standard by ETRI & MPEG-Korea Forum.

ftp://203.253.130.48

Resolution  
XGA, 1024x768

Frame rate  
30 Hz

Length  
300 frames, 10 s

Video format  
YUV4:2:0, uncoded

Stereo format  
multiview, 12 cameras with stereo distance

Copyright  
available to MPEG

Content, characteristics, quality

**Lovebird_1, Lovebird_2**

Outdoor, quiet dialogue in movie type of scene, simple object motion, no camera motion, high detail, complex depth structure, natural light

**Lovebird1:**
Lovebird2:
### 1.1.9. GIST 3DV data, 1 sequence

These multiview video sequences with stereo spacing were provided to MPEG for development of the FTV/3D Video standard by Gwangju Institute of Science and Technology (GIST), Korea.

http://vclab.gist.ac.kr/index.html

- **Resolution**: XGA, 1024x768
- **Frame rate**: 30 Hz
- **Length**: 300 frames, 10 s
- **Video format**: YUV4:2:0, uncoded
- **Stereo format**: multiview, 9 cameras with stereo distance
- **Copyright**: available to MPEG
- **Content, characteristics, quality**

**Newspaper**

Indoor, studio, soap opera type of scene, simple object motion, no camera motion, high detail, medium complex depth structure, studio light
4 Conclusions

Generation of the video plus depth database was reported starting from the stereo video database described in D2.1. Depth was estimated using Hybrid Recursive Matching (HRM) previously developed at FHG-HHI for the left view of each stereo pair. Then, the corresponding right view was rendered from the left view and the estimated depth. For quality assessment the visual impression was evaluated inspecting synthesized right views and the stereo impression on small displays for mobile applications.

In general the depth maps are not perfect. This is an inherent problem of the video plus depth approach. However, we found that the estimated depth maps are temporally and spatially smooth and stable. This is due to the recursive structure of the HRM algorithm. Artefacts are visible in synthesized right views. But when watching stereo video including the original left view and the synthesized right view on a small display for mobile terminals, artefacts are almost not noticeable. Stereo masking effects known as binocular suppression theory lead to a good overall impression. Thus the overall quality of estimated depth maps is sufficient.

Further improvement will be task of the project in cooperation with partners, i.e. post processing, filtering of depth maps, e.g. in cooperation with WP5.
Confidential Part: List of available stereo video data without copyright or permission of use

This section lists results for sequences that are available but with unclear copyright. Use of these data cannot be reported in a public deliverable or in any publication. They can not be distributed via the project’s public repository. If some of these data are useful for the project, it will be intended to get permission to use these data officially from owners. If successful, the data will be added to public part.

1. NHK, 4 sequences

These data originate from NHK, Japan Broadcasting Corporation. Unfortunately, the copyright is still unclear and permission of use is not available.

Resolution 320x240
Frame rate 30 Hz
Length 450 frames, 15 s
Video format YUV4:2:0, uncoded
Stereo format left, right
Copyright unclear
Content, characteristics, quality

Balloons

Outdoor, children running with balloons, complex object motion, partially simple camera motion, partially no camera motion, scene cut, high detail, complex depth structure, natural light
**Flower Pot**
Outdoors, street/square scenery, simple object motion, simple camera motion, high detail, complex depth structure, natural light

**Street Organ**
Outdoors, close up scenery with moving persons and requisites, complex object motion, medium complex camera motion, high detail, complex depth structure, natural light
Trapeze

Outdoor, acrobats and landscapes, scene cuts, complex object motion, complex camera motion, high detail, complex depth structure, natural light
MOBILE3DTV - Mobile 3DTV Content Delivery Optimization over DVB-H System - is a three-year project which started in January 2008. The project is partly funded by the European Union 7th RTD Framework Programme in the context of the Information & Communication Technology (ICT) Cooperation Theme.

The main objective of MOBILE3DTV is to demonstrate the viability of the new technology of mobile 3DTV. The project develops a technology demonstration system for the creation and coding of 3D video content, its delivery over DVB-H and display on a mobile device, equipped with an auto-stereoscopic display.

The MOBILE3DTV consortium is formed by three universities, a public research institute and two SMEs from Finland, Germany, Turkey, and Bulgaria. Partners span diverse yet complementary expertise in the areas of 3D content creation and coding, error resilient transmission, user studies, visual quality enhancement and project management.

For further information about the project, please visit www.mobile3dtv.eu.