Fusion of Depth Maps for Multi-view Reconstruction

Yu ZHAO1, Ningqing QIAN1

1Institute of Communications Engineering, RWTH Aachen University, 52056 Aachen, Germany
yu.zhao1@rwth-aachen.de, qian@ient.rwth-aachen.de

Abstract. Depth fusion has been considered as an effective approach to realize 3D reconstruction from multiple images captured at different points of view. However, the noises originating from redundant 3D points during fusion usually lead to over-complete representation and further deteriorate the 3D shape interpretation. The presented method models objects and scenes as 3D point clouds using multi-view images. By analyzing the depth consensus in multiple views a strict selection of points is made for precise reconstruction. The 3D point cloud is further processed by applying a bayesian-based model to analyze the photometric property of the images at sub-pixel accuracy. Finally, a set of experiments is carried out to study the influence of different parameters on the reconstruction quality.

Keywords
multi-view stereo, depth map fusion

1. Introduction

Accurate extraction of 3D structures from images taken with calibrated cameras from known positions has become a very important issue in various areas. The technique has been widely applied to restore and archive historic architectures. In driver assistance system, 3D perception of the vehicle surroundings helps to estimate and detect potential risks and actively take actions to avoid accidents in time.

According to [8] the algorithms are classified as voxel coloring, deformation-based methods, patch based methods and last but not least depth map fusion, which is addressed in this paper.

The depth fusion approach starts with determination of the 2D displacement between left and right image after selecting two views as a stereo pair. Given multiple depth maps the visibility based approach [3], [4] is limited to the detection and omission of uncertainties. This is a both fast and effective way whereas the probabilistic approach [5] (also bayesian fusion) corrects 3D coordinate errors via variation of the disparity and can be only solved iteratively. In this paper we present a new method that analyzes the origins of the uncertainties by fusion of the multiple depth maps and improve the reconstructed precision with the application of a bayesian model. We evaluate our method with the ground-truth data and show the quality of our reconstruction.

The paper is organized as follows. In Section 2, we describe the approaches that generate the depth map. The proposed method to fuse the multiple depth maps is explained in Section 3. In Section 4, experimental results will be presented and analyzed in detail. Finally we conclude our paper in Section 5.

2. Depth Map Generation

To compute the depth map, pixels are matched by DAISY[2] descriptor, which has been considered as an efficient tool to search point correspondence among images. The descriptors are based on directional gradients and computed from eight neighboring positions distributed on two rings around the pixel and also from the center position itself. The distance from the center to the outer most grid point is within eight pixels while the gradients are approximated by four angular quantizations. From the euclidian distance thereof the similarity degree between the two descriptors can be defined as

$$s(d) = \exp(-|D_x - D_x'(d)|^2)$$

(1)

The decision whether or not to assign a depth value is made after securing a minimum degree of similarity,

$$s(d_{best}) > 0.5.$$  

(2)

Also to ensure the uniqueness of the descriptor, we assume

$$\frac{s(d_{2ndbest})}{s(d_{best})} < 0.995.$$  

(3)

where $d_{2ndbest}$ is the second best match.

Unlike SIFT [1], features are not extracted from multi-scale resolutions, also no orientation is assigned which make DAISY whether scale- nor rotation-invariant. These properties are not relevant as long as zooming factors and viewing distances remain constant. Also for stereo matching cameras are usually not tilted around the optical axis. We rather want to profits from its efficiency.

In [2] DAISY descriptors are provided with occlusion masks for disparity calculation over large baselines. Bi-
nary masks exclude parts of the DAISY vector from stereo-
matching, since similarity of corresponding pixels is not as-
sumed under occlusion. This often leads to wrongly assigned
disparities. Due to the fact that all disturbances are elimi-
nated in the fusion step, implementation thereof is omitted.

Given a pixel in the reference view the search for its cor-
correspondence can be restricted on the epipolar lines.

Depth maps are equivalent to the representations of
point clouds, since the calculation of 3D positions is based
on disparities. The linear triangulation method [6] yields
good performances and can be easily implemented as well.
Given two image coordinates \( x = (u, v)^T \) and \( x' =
(u', v')^T \) in view 1 and view 2 for instance, the back
projected space coordinate \( X \) corresponds to the last column
of \( V \) after applying a singular value decomposition, \( A =
UDV^T \). With full calibration provided, \( A \) is composed of

\[
A = \begin{bmatrix}
    u_1^T p_1^T & p_1^T \\
    u_2^T p_2^T & p_2^T \\
    u_3^T p_3^T & p_3^T \\
    u_3^T p_{3, j} - p_2^T \\
\end{bmatrix},
\]

where \( p_i \) represents the rows of the the projection ma-
trices \( P \) and \( P' \). The method can be extended to more than
two views by adding additional cameras to the rows of \( A \).
By adding the disparity between view 2 and 3 on the pixel
\((u', v')^T\), thus acquiring \((u'', v'')^T\), the space coordinates
would be result of a trifocal triangulation,

\[
A = \begin{bmatrix}
    u_1^T p_1^T - p_1^T \\
    u_2^T p_3^T - p_3^T \\
    u_3^T p_3^T - p_3^T \\
    u_3^T p_{3, j} - p_2^T \\
\end{bmatrix},
\]

3. Depth Map Fusion

A set of depth maps can be acquired by repositioning a
camera multiple times, thus capturing a scene from different
angles. Multi-view stereo is preferred whenever reconstruc-
tion of both frontal and lateral sides is desired.

Due to overlapping fields of view redundancy may ex-
ist. Eliminating overrepresentation decreases data size while
all information is maintained. Noises arise from perform-
ing stereo matching on monochrome or occluded surfaces
as well as edges parallel to epipolar lines. These cannot
be completely avoided, only reduced by Eq. 2 and 3. Be-
cause wrongly placed 3D points create depth inconsistencies
as well as color inconsistencies at projections of neighboring
depth maps and color images they can be reliably detected.

To increase the quality of the reconstructed model, a
fast way is to exclude the noises and redundancy from being
triangulated into space. The further improvement can be re-
alized to test alternative positions and readjust the points by
maximizing a probability function.

3.1. Noise and Redundancy Removal

The depth stored in every pixel grid is reprojected back
into space at \( X_1 \). Let \( i \) denote the camera index of which a
point is triangulated and \( j \) the neighboring camera to which
the depth is given. \( D_{j,i} \) corresponds to the depth value \( X_2 \)
stored in depth map \( j \), on which \( X_1 \) is projected.

![Fig. 1.](image)

Noisy disparity values are cleared up before a redu-
dancy detection process reduces the data size. 3D points are
only retained with consistent depth evidences in neighboring
views. This required number is expressed by \( \text{req} \). A posi-
tion assumption is considered as valid if the deviation from
the neighboring perspective is below a threshold \( th_1 \).

\[
sup = sup + 1, \text{if } \frac{|D_{j,i} - D_{j,j}|}{D_{j,j}} < th_1.
\]

A decision whether to keep \( X_1 \) seen from camera \( j \) at the
depth \( D_{i,j} \) is made after reviewing all remaining depth maps
and counting the total number of consensus support. A re-
quired number has to be met, otherwise \( X_1 \) must be dis-
carded,

\[
X_1 = \begin{cases}
\text{noise, if } sup < \text{req} \\
\text{valid, if } sup > \text{req}
\end{cases}.
\]

If the depth of \( X_1 \) with respect to the neighboring camera
is smaller than the value computed at the projection of \( X_1 \)
in the neighboring depth map, no observation of \( X_2 \) is possible
due to occlusion. Therefore \( X_2 \) is not reconstructed,

\[
X_2 = \begin{cases}
\text{occluded, if } D_{i,j} <= D_{j,j} \\
\text{valid, if } D_{i,j} > D_{j,j}
\end{cases}.
\]

Furthermore overcomplete representations should be
detected. If the reconstructed points from the two depth val-
ues \( D_{i,j}, D_{j,j} \) are close,

\[
X_2 = \begin{cases}
\text{redundant, if } |D_{i,j} - D_{j,j}| < th_2 \\
\text{valid, if } |D_{i,j} - D_{j,j}| > th_2
\end{cases}.
\]
it follows that two pixels of different cameras represent the
exact same point. \( X_2 \) would be a duplication of \( X_1 \) and
should be removed.

A point that is likely to be discarded should not pre-
vent other points from being placed in the same spot, which
 corresponds to the first case of Eq. 8. Prior to redundancy
removal \( X_1 \) itself is considered to be at risk to be wiped out
as redundancy if its territory measured by \( t h_3 \) is occupied by
other points. Otherwise if an area within \( t h_3 \) was represented
solely by \( X_1 \), in other words high amount of information is
concentrated there, the remaining thereof is guaranteed, so
no other point should take its position in the environment
defined by \( t h_2 \).

\[
X_1 = \begin{cases} 
irrelevant \text{ for } (8), \text{ if } \exists j \quad \frac{|D_{i,j} - D_{i,j}|}{D_{j,j}} < t h_3 \\
relevant \text{ for } (8), \text{ if } \forall j \quad \frac{|D_{i,j} - D_{i,j}|}{D_{j,j}} > t h_3
\end{cases}
\tag{9}
\]

\( t h_3 \) is typically with few magnitudes smaller than \( t h_2 \).

Finally using the described methods the merged depth-
maps produce a point cloud that is both precise and compact.

3.2. Outlier Correction

Bayesian Fusion evaluates particle positions from the
viewpoint of conditional probabilities. By maximizing the
probability

\[
P(I|V, D_{sub}, \sigma, \Sigma) = A + B
\tag{10}
\]

the positions of \( X \) is readjusted.

\( I \) denotes the \( n \) given color images with \( I(x) \) \( i = 1, \ldots, n \)
being the color vector of pixel \( x \). \( \sigma \) and \( \Sigma \) corresponds to
user defined parameters. \( D_{sub} \) stands for the set of depth
maps that influences the spatial positions \( X \). These are up-
sampled by factor 10, i.e the disparities are decreased or in-
creased by 0.5 at maximum in steps of 0.1 pixel from an ini-
tial integer value. Instead of the best integer pixel among all
candidates, the optimum sub pixel disparity is chosen which
maximizes Eq. 10. Also a boolean visibility map \( V \) is in-
troduced to indicate for each point in space, whether that is
occluded in the image \( I_i \).

In our approach bayesian fusion is carried out prior to the
visibility-based fusion introduced in Section 3.1. Com-
bining both approaches the quantization error originating
from discrete pixel coordinates is reduced.

Eq. 10 is basically composed of a normal distribution
\( A \) and an uniform distribution \( B \). The color distribution of
the image at the projection \( I_j(x) \) is modeled as a multivariate
gaussian distribution

\[
A = \prod_X \prod_j P(V_j | C(X), \Sigma) \prod_j N(I_j(X))
\tag{11}
\]

\( B = \prod X \prod_j P(V_j, x = 0 | D_{sub}, \tau) U(I_j(X)) \).
\tag{12}

The visibility probability is estimated comparing \( D_{i,j} \)
with \( D_{j,j} \)

\[
P(V_j | x = 1 | D_{sub}, \sigma) = 0.9 \cdot \exp(\frac{(D_{i,j} - D_{j,j})^2}{2\sigma^2})
\tag{13}
\]

A small deviation of the distances usually means that the
same point \( X \) is also seen from camera \( j \). Symmetrically oc-
cclusions are usually accompanied by greater depth discrep-
ancies which happens if the view on \( X \) from camera \( j \) is
blocked by another object. The probability for the later case
corresponds to the complementary

\[
P(0 | D_{sub}, \sigma) = 1 - P(1 | D_{sub}, \sigma)
\tag{14}
\]

4. Results

\[\text{Fig. 3. Ground truth polygons of the scenes "Fountain" and "Herzjesu".}\]

We reconstructed two scenes named "Fountain" and
"Herzjesu" using 11, respectively 8 images at 3072x2048
resolution. The result $R$ is evaluated by comparing it with a reference model $G$ generated by a LIDAR scanner. Because $G$ is given as a triangle mesh we also performed an identical surface approximation [7] on our model. To penalize sparse and noisy reconstructions, precision and completeness are simultaneously considered.

To measure the precision of a reconstruction, according to [8] the distance between vertices in $R$ and the nearest vertices on $G$ are given. Having a statistics about these the value $d_{\text{prec}}$ is computed such that $85\%$ of the points on $R$ are within this distance away from $G$.

To measure completeness, the opposite to what is done so far, the calculation of distances to $R$ starting from $G$, is performed. Vertices on $G$ without any vertex on $R$ at distance below 0.1 will be considered as not covered.

### 4.1. Three-View Triangulation

![Surfaces reconstructed by stereo (left) and trifocal triangulation (right).](image)

We started with a comparison of triangulation methods. Methods from Section 3.1 were applied to free our point cloud from noises and redundancies. For this purpose $\text{req}=2$, $\text{th}_1=0.005$, $\text{th}_2=0.01$ and $\text{th}_3=1e^{-8}$ were chosen. The recovered model using 3-view is closer to the ground truth than using 2-View. However the completeness deteriorates, since the existence of the second correspondence pair could not always be guaranteed. In the following sections only 2-view triangulation is considered.

<table>
<thead>
<tr>
<th>scene</th>
<th>triangulation</th>
<th>$d_{\text{prec}}$</th>
<th>completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fountain</td>
<td>2-view</td>
<td>0.097261</td>
<td>69.3837%</td>
</tr>
<tr>
<td>Fountain</td>
<td>3-view</td>
<td>0.037023</td>
<td>64.4742%</td>
</tr>
<tr>
<td>Herzjesu</td>
<td>2-view</td>
<td>0.041337</td>
<td>59.6526%</td>
</tr>
<tr>
<td>Herzjesu</td>
<td>3-view</td>
<td>0.036443</td>
<td>55.7928%</td>
</tr>
</tbody>
</table>

Tab. 1. Influence of the triangulation method on the reconstruction quality.

<table>
<thead>
<tr>
<th>scene</th>
<th>req</th>
<th>$\text{th}_1$</th>
<th>$d_{\text{prec}}$</th>
<th>completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fountain</td>
<td>1</td>
<td>0.01</td>
<td>8.06695</td>
<td>1.2546%</td>
</tr>
<tr>
<td>Fountain</td>
<td>2</td>
<td>0.01</td>
<td>0.107808</td>
<td>69.9384%</td>
</tr>
<tr>
<td>Fountain</td>
<td>3</td>
<td>0.01</td>
<td>0.095817</td>
<td>68.5784%</td>
</tr>
<tr>
<td>Herzjesu</td>
<td>1</td>
<td>0.01</td>
<td>19.3622</td>
<td>2.3252%</td>
</tr>
<tr>
<td>Herzjesu</td>
<td>2</td>
<td>0.01</td>
<td>0.047875</td>
<td>60.8376%</td>
</tr>
<tr>
<td>Herzjesu</td>
<td>3</td>
<td>0.01</td>
<td>0.042161</td>
<td>57.9048%</td>
</tr>
</tbody>
</table>

Tab. 2. Influence of the parameters $\text{th}_1$ and $\text{req}$ on the reconstruction quality.

<table>
<thead>
<tr>
<th>scene</th>
<th>bayesian subpixel</th>
<th>$d_{\text{prec}}$</th>
<th>completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fountain</td>
<td>without</td>
<td>0.097261</td>
<td>69.3837%</td>
</tr>
<tr>
<td>Fountain</td>
<td>with</td>
<td>0.094451</td>
<td>69.4913%</td>
</tr>
<tr>
<td>Herzjesu</td>
<td>without</td>
<td>0.041337</td>
<td>59.6526%</td>
</tr>
<tr>
<td>Herzjesu</td>
<td>with</td>
<td>0.040773</td>
<td>60.1716%</td>
</tr>
</tbody>
</table>

Tab. 3. Effect of bayesian sub-pixel correction on the reconstruction quality.

### 4.2. Noise and Redundancy Detection

![Influence of denoising parameters on the reconstruction quality. From left to right $\text{req}=1$, $\text{req}=2$, $\text{req}=3$. $\text{th}_1=0.001$ is used in all three cases.](image)

The denoising process is controlled by $\text{req}$, the number of depth consensuses from neighboring views which is required at minimum and $\text{th}_1$, below which consensus is defined. A strict selection of high-quality points increases on the one hand accuracy, completeness on the other hand may decrease. $\text{req}=2$ and $\text{th}_1=0.001$ are considered as good choices (fig. 5). Keeping $\text{req}$ low, correctly reconstructed point are less vulnerable to be erased, but with $\text{req}$ below 2 the scene becomes noisy, so that geometries and shapes cannot be perceived properly.

The parameters for redundancy detection $\text{th}_2$ and $\text{th}_3$ are tested in the range $[1, 0.001]$ and $[1e^{-7}, 1e^{-10}]$ respectively. In fact the algorithm is not steered by these two, in all cases approximately the same amount of redundancy (10\%) is removed.

### 4.3. Bayesian Sub-pixel Correction

So far all tests purely relied on the fusion described in Section 3.1. In this section a preliminary optimization took place which maximizes Eq. 10. The color channels are assumed to be uncorrelated, so $\Sigma$ has the shape of an identity
matrix with value 10 on the diagonal components, whereas $\sigma^2$ is chosen as 2.5. Afterwards noise/redundancy removal was carried out using $th_1=0.005$, $req=2$, $th_2=0.01$, $th_3=1e^{-8}$ as usual. The only difference are the finer depth values. Table 3 shows that adding the probabilistic approach points in space are placed more accurately, shown by a decrease in $d_{prec}$. An increased completeness is explained by the positive influence on the noise detection, by which more points are below the inlier threshold $th_1$, so lesser invalid points have to be removed.

5. Conclusion

In this paper a powerful algorithm for refining point clouds is presented. For the given datasets all noise is removed without exception. Also a reduction at about 10% of the number of cloud points was achieved by eliminating redundancies. Improvement by probabilistic optimization is measured. We find out that it is possible to increase the resolution of disparity even beyond image resolutions. A further optimized bayesian model should contain a smoothness term. Its implementation for high-resolution photos is however challenging, because the number of neighboring points would be extremely large, which have to be updated after each repositioning of a point. The disadvantage of 3-view triangulation compared to 2-view is the lower completeness, however it improves accuracy of the 3D point positions. As a future extension it could be combined with an if-query whether a third correspondence exists, in which case one would switch to 3-view mode.

Acknowledgements

The research presented in this paper was supervised by Prof. J. R. Ohm, IENT, RWTH Aachen University. The authors would like to thank all the members of IENT for kindly assistance and help.

References


About Authors...

Yu ZHAO

was born in Jilin, China in 1987. He is currently pursuing the Master’s degree in Electrical Engineering, Information Technology and Computer Engineering at RWTH Aachen University. He received his bachelor degree in Electrical Engineering, Information Technology and Computer Engineering from RWTH Aachen University, in 2011.

Ningqing QIAN

was born in Fuzhou, China in 1981. She received the Master’s degree in Electrical Engineering, Information Technology and Computer Engineering from RWTH Aachen University, Aachen, Germany in 2008 and is currently working as a Ph.D. student at the Institute of Communications Engineering, RWTH Aachen University. Her focus is on algorithms and applications in multi-view 3D reconstruction.