Multiple View Stereo by Reflectance Modeling

Kim, Sujung; Kim, Seong Dae; Dahl, Anders Lindbjerg; Conradsen, Knut; Jensen, Rasmus Ramsbøl; Aanæs, Henrik

Published in:
2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)

Link to article, DOI:
10.1109/3DIMPVT.2012.82

Publication date:
2012

Citation (APA):
Kim, S., Kim, S. D., Dahl, A. L., Conradsen, K., Jensen, R. R., & Aanæs, H. (2012). Multiple View Stereo by Reflectance Modeling. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)* (pp. 579-586). IEEE. https://doi.org/10.1109/3DIMPVT.2012.82

Multiple View Stereo by Reflectance Modeling

Sujung Kim, Seong Dae Kim
Korea Advanced Institute of Science and Technology
Daejeon, Repulic of Korea
sjkim@sdvision.kaist.ac.kr, sdkim@kaist.ac.kr

Anders Lindbjerg Dahl, Knut Conradsen, Rasmus Ramsbøl Jensen, Henrik Aanæs
DTU Informatics
Technical University of Denmark, Lyngby, Denmark
abd@imm.dtu.dk, kc@imm.dtu.dk, raje@imm.dtu.dk, haa@imm.dtu.dk

Abstract

Multiple view stereo is typically formulated as an optimization problem over a data term and a prior term. The data term is based on the consistency of images projected on a hypothesized surface. This consistency is based on a measure denoted a visual metric*, e.g. normalized cross correlation. Here we argue that a visual metric based on a surface reflectance model should be founded on more observations than the degrees of freedom (dof) of the reflectance model. If (partly) specular surfaces are to be handled, this implies a model with at least two dof. In this paper, we propose to construct visual metrics of more than one dof using the DAISY methodology, which compares favorably to the state of the art in the experiments carried out. These experiments are based on a novel data set of eight scenes with diffuse and specular surfaces and accompanying ground truth. The performance of six different visual metrics based on the DAISY framework is investigated experimentally, addressing whether a visual metric should be aggregated from a set of minimal images, which dof is best, or whether a combination of one and two dof should be used. Which metric performs best is dependent on the viewed scene, although there are clear tendencies for the two dof minimal metric to be the preferred one.*

1. Introduction

Multiple view stereo or the dense 3D reconstruction of the surface of an object from multiple calibrated images is one of the persistent central challenges of computer vision. This paper addresses this challenge by investigating image similarity measures – *the visual metrics* for surfaces with light reflectance properties that contain both specular and diffuse components.

A massive effort has recently been put in multiple view stereo, and advances have been achieved by recent benchmark datasets like the Middlebury multi-view stereo sets [17], the dense multi-view stereo of buildings from Strecha et al. [19], as well as works on large scale urban reconstruction of Furukawa et al. [8] and Gallup et al. [11]. Many recent landmark achievements [6, 9, 11, 12, 16, 22, 23, 24] have been obtained. These recent efforts have mainly focused on methods for optimization and regularization. The visual metrics used have been sums of squared differences (SSD) or normalized cross correlation (NCC) between image pairs. These visual metrics are well suited for diffuse reflecting surfaces, where the surface appearance is independent of the viewing direction, but not for more complex reflecting surfaces with specularities. Both the Middlebury datasets [17] and the buildings from [19] consist of diffuse objects, and therefore fit well with the simple visual metrics such as SSD and NCC.

Many real world objects are not well modeled as diffuse reflecting. Multiple view stereo algorithms can, however, handle a lot of these objects using NCC or SSD by robust statistics and an abundance of images. Such an abundance is, however, often not possible or practical, and in these cases, the SSD and NCC based frameworks brake down, and a more elaborate visual metric is needed.

In [14, 18], it is shown that visual metrics dealing with objects with more complex reflectance properties, e.g. specular, cannot be based on comparing image pairs. In this paper, we further this work and propose novel visual metrics based on modeling the reflectance. To do this the number of images should exceed the dof of the reflectance model, which is one in the diffuse case. Based on this realization, we investigate how to construct visual metrics dealing with diffuse *and* specular objects, and thus reflectance models with more than one dof. This results in a visual metric with

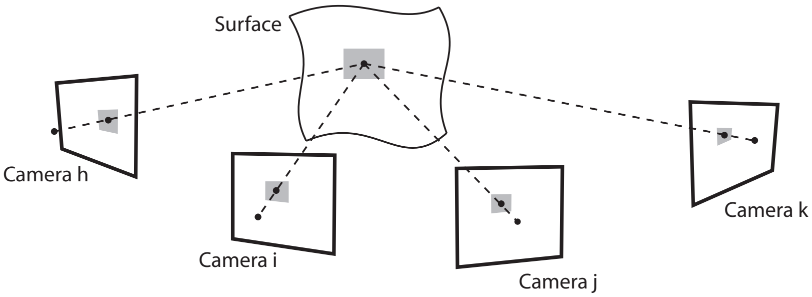

Figure 1. Illustration of the relationship between a surface patch and corresponding image patches. Visual metrics typically evaluate the support for the presence of a given surface patch in the data by comparing these image patches

better properties than the radiance tensor of Jin et al. [14].

Considerable evidence exists to support that the SIFT framework is superior to NCC when dealing with salient feature point matching [7, 15]. This has been exploited by Tola et al. [21] in the stereo case by changing the binning of the descriptors to the output of Gaussian filters, whereby the computations could be performed more efficiently resulting in the DAISY descriptor. To deal with more than two images, the DAISY framework combines the scores of image pair matchings as commonly done with NCC. As part of the investigation we also propose how to construct a visual metric with more than one dof by constructing a tensor of DAISY descriptors. We experimentally show that the DAISY tensor is superior.

The investigation of the proposed visual metric is based on a new data set of eight different scenes with diffuse and specular objects. This data is accompanied by ground truth obtained by a structured light scanner[1]. Firstly we demonstrate that the DAISY tensor is to be preferred to raw pixels because it is more robust and approximates the ground truth better. Following this, we investigate the difference between using one or two dof and a combination of the two. As for the latter, if the part of a scene is diffuse, then extra dof could lead to overfitting so that it might cause possible performance degradation. Lastly, we investigate if a visual metric should be aggregated from a minimal set of images, i.e. two in the diffuse case and three for the proposed visual metric, or directly based on all relevant images. The performed experiments are done in a 2.5D manner via the alpha expansion of [2]. This is a relatively simple reconstruction algorithm, which we deliberately have chosen over the state of the art algorithms because our focus is on the visual metric. If we chose an algorithm with stronger modeling capabilities, this could clutter the effect of the visual metric. Choosing an algorithm that does not use contextual information might, however, not reveal the potential of the visual metric in a realistic setup. We found the choice of the graph cut algorithm [2] a good tradeoff.

In this paper, we investigate the similarity metric similar to the work of [13], but we focus on multiple views opposed to stereo in their work. An in depth review of the multiple view stereo literature and introduction of this field can be found in [17, 5].

2. Visual Metrics

Multiple view stereo deals with estimating the 3D surface of an object or a scene from multiple images with the known camera calibrations. These known calibrations allow us to compute where a given 3D point is projected in the images, cf. Figure 1. Multiple view stereo is typically handled as an optimization problem, where we want to find the surface $\mathcal{S}$, which is most consistent with the images. Normally a prior is added. This prior is often formulated as a smoothing. The image consistency is formulated as a visual metric, $\mathrm{V}(\mathbf{x},\mathbf{n})$, evaluated at each point $\mathbf{x}$ on the surface with normal $\mathbf{n}$. The optimization problem thus becomes[2]

$$\min_{\mathcal{S}} \sum_{\mathbf{x}\in\mathcal{S}} \mathrm{V}(\mathbf{x},\mathbf{n}(\mathbf{x})) \; + \mathrm{Prior}(\mathcal{S}) \;\; , \qquad (1)$$

where $\mathbf{n}(\mathbf{x})$ is the surface normal at $\mathbf{x}$. The visual metric is based on a planar patch at $\mathbf{x}$ with normal $\mathbf{n}$, whereupon the relevant images are projected as illustrated in Figure 1. Different visual metrics then employ different measures to quantify the consistency. A typical example is the use of NCC between pairs of projected patches, e.g. projections from cameras i and j in Figure 1. It is the construction of these consistency measures we investigate further in this paper.

2.1. The Radiance Tensor

In Jin et al. [14], a visual metric is constructed via a radiance tensor. For a given surface patch described by $\mathbf{x}$ and $\mathbf{n}$, this radiance tensor is constructed by firstly enumerating the relevant, visible, images by $i \in \{1,\ldots,n\}$. Denote the m pixel intensities of the associated projected patches as

[1] This data set is available at *http://roboimagedata.imm.dtu.dk/reflectance/*.

[2] This is typically formulated as an integral over $\mathcal{S}$, which is then later discretized.

$\mathbf{r}_i(\mathbf{x},\mathbf{n})$, where $\mathbf{r}_i(\mathbf{x},\mathbf{n})$ is an m dimensional vector. These vectors are then combined into the $m \times n$ radiance tensor

$$\mathbf{R}(\mathbf{x},\mathbf{n}) = \left[\begin{array}{cccc} \mathbf{r}_1(\mathbf{x},\mathbf{n}) & \mathbf{r}_2(\mathbf{x},\mathbf{n}) & \cdots & \mathbf{r}_n(\mathbf{x},\mathbf{n}) \end{array} \right] \ . \tag{2}$$

In the ideal case where the patch coincides with the surface and no other noise is present either, a patch should look the same from all directions, up to scale, in the diffuse case. In this ideal case, all $\mathbf{r}_i(\mathbf{x},\mathbf{n})$ should thus be scaled versions of each other, and the rank of $\mathbf{R}(\mathbf{x},\mathbf{n})$ becomes one. A main result of [14] is that if the reflectance model of a surface is described by the diffuse plus specular Phong model, then the rank of $\mathbf{R}(\mathbf{x},\mathbf{n})$ should be two in the ideal case.

In the rank two case of [14], the singular values[3] of $\mathbf{R}(\mathbf{x},\mathbf{n})$, $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$, form the basis of a visual metric. Given a patch on the true surface in the ideal case, only the first two singular values σ_1 and σ_2 should be non-zero. This corresponds to $\mathbf{R}(\mathbf{x},\mathbf{n})$ having rank two. The visual metric, $J(\mathbf{x},\mathbf{n})$, from [14] is thus

$$J(\mathbf{x},\mathbf{n}) = \sum_{i=3}^{n} \sigma_i^2 \ , \tag{3}$$

which is equal to the total variation of the noise for the patch being on the true surface. A similar visual metric corresponding to a diffuse model would similarly be [4]

$$\sum_{i=2}^{n} \sigma_i^2 \ . \tag{4}$$

2.2. Visual Metric as Model Fitting Residual

An interpretation of the visual metric in (3) is that a linear subspace is fitted to the data, i.e. the $\mathbf{r}_i(\mathbf{x},\mathbf{n})$, and that the visual metric is put equal to the squared residual error. This linear subspace has dimension two, corresponding the the models dof. The same interpretation can be made of (4) except that a 1D subspace is fitted. Similarly, the cross-correlation, ρ_{ij}, between $\mathbf{r}_i(\mathbf{x},\mathbf{n})$ and $\mathbf{r}_j(\mathbf{x},\mathbf{n})$ is the best fit of the model

$$\left\| \frac{\mathbf{r}_i(\mathbf{x},\mathbf{n}) - \mu_i}{\|\mathbf{r}_i(\mathbf{x},\mathbf{n}) - \mu_i\|} - \alpha \frac{\mathbf{r}_j(\mathbf{x},\mathbf{n}) - \mu_j}{\|\mathbf{r}_j(\mathbf{x},\mathbf{n}) - \mu_j\|} \right\|_2^2 \ , \tag{5}$$

where μ_i is the mean of $\mathbf{r}_i(\mathbf{x},\mathbf{n})$ – i.e. $\alpha^* = \rho_{ij}$. The residual error is $1 - \rho_{ij}^2$. The NCC can thus also be interpreted as residual error after fitting a one parameter model.

An implication of viewing a visual metric as a model fitting residual is that we need more observations, i.e. $|n|$, than the dof of the underlying reflectance model. If not, the residual, and thus the visual metric, will always be zero. Thus, the diffuse model works well with only two observations, ($n = 2$), since it has one dof.

The model fitting residual interpretation does not need to be possible for all conceivable visual metrics. However, given a reflectance model, then its dof is equal to the dimension of the possible ways in which a surface patch can change appearance between image views in general. Thus, at least one more image observation, $\mathbf{r}_i(\mathbf{x},\mathbf{n})$, than the dof is needed. This again implies that if a visual metric is based on a reflectance model, then it needs to be based on at least one plus the dof observations.

The conclusions reached here are generalizations of [18], which is based on more formal arguments. Note also that visual metrics are often made invariant under different actions, e.g. rotation in the SIFT descriptor [15]. Such an invariance removes an effect instead of modeling it, and as such it does not increase the dof.

3. Visual Metrics for Specular Surfaces

Specular surfaces are best described by a two or larger dof reflectance model. So based on the above reasoning, we wish to investigate how we may best construct visual metrics of more than the usual one dof. First of all, we propose an extension of the SIFT methodology to the two or larger dof case via a DAISY tensor.

3.1. DAISY Tensor

A DAISY descriptor [21] of a gray scale image is computed from orientated image derivatives. These derivatives are convolved by Gaussian kernels and the filter output form the entries of a DAISY descriptor vector $\mathbf{d}_i(\mathbf{x},\mathbf{n})$. We propose forming a tensor of the relevant DAISY descriptors, described by 3D point $\mathbf{x}$ and normal $\mathbf{n}$ as in line with (2)

$$\mathbf{D}(\mathbf{x},\mathbf{n}) = \left[\begin{array}{cccc} \mathbf{d}_1(\mathbf{x},\mathbf{n}) & \mathbf{d}_2(\mathbf{x},\mathbf{n}) & \cdots & \mathbf{d}_n(\mathbf{x},\mathbf{n}) \end{array} \right] \ . \tag{6}$$

Let the singular values of $\mathbf{D}(\mathbf{x},\mathbf{n})$ be given by $\{\varsigma_1, \varsigma_2, \ldots, \varsigma_n\}$. Then we can form visual metrics as[5]

$$D_1(\mathbf{x},\mathbf{n}) = \sum_{i=2}^{n} \varsigma_i^2 \tag{7}$$

$$D_2(\mathbf{x},\mathbf{n}) = \sum_{i=3}^{n} \varsigma_i^2 \ . \tag{8}$$

3.2. Further Lines of Investigation

In line with findings for two view stereo [20] and salient features [7], our experiments show that the DAISY tensor outperforms the radiance tensor as a basis for a visual metric. Likewise, we only consider linear subspaces of a given degree as representatives of models of a given dof.

[3] In general $n < m$, and there is thus n singular values of $\mathbf{R}(\mathbf{x},\mathbf{n})$.

[4] Note the starting index of the summation.

[5] Note the starting indices of the summations

3.2.1 Minimal vs. All

For salient features, matching performance is increased for smaller differences in viewing angle between images. It is partly because the approximation of the planar patch assumption becomes less profound. It is thus relevant to ponder whether visual metrics should be based on aggregations of *minimal* sets of images, as done with NCC in [22], or if *all* relevant images should be used at once as done in [14], cf. (3). Using all relevant images at once increases the redundancy in the data giving bigger noise reduction. Also the larger difference in viewing angle will generally give a better baseline to depth ratio, and thus better depth estimation, cf. [10]. To shed light on this matter, we compare the two alternatives experimentally.

The visual metrics directly using all relevant images are given by (7) and (8). The size of the minimal sets is one plus the dof of the model since there needs to be a residual. In the two dof case, we denote these sets $\{i,j,k\} \in \mathcal{C}_3$. The visual metric is then aggregated from the squared third singular value of

$$\left[\begin{array}{ccc} \mathbf{d}_i(\mathbf{x},\mathbf{n}) & \mathbf{d}_j(\mathbf{x},\mathbf{n}) & \mathbf{d}_k(\mathbf{x},\mathbf{n}) \end{array}\right] \ , \tag{9}$$

which we denote $\Gamma^3_{ijk}(\mathbf{x},\mathbf{n})$, i.e.

$$\begin{gathered} \Gamma^3_{ijk}(\mathbf{x},\mathbf{n}) = \varsigma_3^2 = \\ \min_{\mathbf{v}_1,\mathbf{v}_2} \sum_{m\in\{i,j,k\}} \left\| \mathbf{d}_m(\mathbf{x},\mathbf{n}) - [\mathbf{v}_1\mathbf{v}_2][\mathbf{v}_1\mathbf{v}_2]^T \mathbf{d}_m(\mathbf{x},\mathbf{n}) \right\|_2^2 , \end{gathered} \tag{10}$$

where $\mathbf{v}_1, \mathbf{v}_2$ is an orthonormal basis of a 2D linear subspace. The two dof minimal visual metric considered here is then given by

$$M_2(\mathbf{x},\mathbf{n}) = \sum_{\{i,j,k\}\in\mathcal{C}_3} \Gamma^3_{ijk}(\mathbf{x},\mathbf{n}) \ , \tag{11}$$

which is the sum of ς_3^2 for all relevant image triplets. In an analog fashion, the one dof minimal visual metric is given by

$$M_1(\mathbf{x},\mathbf{n}) = \sum_{\{i,j\}\in\mathcal{C}_2} \Gamma^2_{ij}(\mathbf{x},\mathbf{n}) \ . \tag{12}$$

3.2.2 Model Averaging

Although two dof visual metrics are superior when dealing with specular surfaces, one dof visual metrics suffice for diffuse surfaces. In the latter case, a two dof visual metric would possibly overfit leading to performance loss. A visual metric averaging the one and two dof models is also investigated. We propose an additional pair of visual metrics

$$\begin{aligned} D_{1.5}(\mathbf{x},\mathbf{n}) &= \frac{1}{2}D_2(\mathbf{x},\mathbf{n}) + \frac{1}{2}D_1(\mathbf{x},\mathbf{n}) \\ &= \frac{1}{2}\varsigma_2^2 + \sum_{i=3}^{n}\varsigma_i^2 && (13) \\ M_{1.5}(\mathbf{x},\mathbf{n}) &= \sum_{\{i,j,k\}\in\mathcal{C}_3} \Gamma^{2.5}_{ijk}(\mathbf{x},\mathbf{n}) && (14) \\ \text{where} \quad & \Gamma^{2.5}_{ijk}(\mathbf{x},\mathbf{n}) = \frac{1}{2}\varsigma_2^2 + \varsigma_3^2 \ . \end{aligned}$$

3.2.3 Investigated Visual Metrics

In summary, our investigation is based on eight visual metrics. Two are based on the raw pixel intensities J, (3), with two different patch sizes. Six are based on the DAISY tensor, i.e. D_1, $D_{1.5}$, D_2, M_1, $M_{1.5}$, and M_2, investigating the combined possibilities of

- If one dof, two dof or an averaged alternative should be used.
- If the visual metric should be based directly on all relevant images or on a combination of minimal subsets.

4. Experimental Results

To perform multiple view stereo experiments on objects with specular and diffuse surface reflectance models, we compiled a new data set consisting of eight different scenes as shown in Figure 2. The scenes show specular reflectances and have planar to non-planar surfaces. We chose to vary the baseline of the different data sets to ensure significant specularities to challenge the visual metric. This is done by visual inspection. The number of images was kept constant at five and the maximum angles between images of the eight scenes were: #1 – 20°, #2 – 40°, #3 – 20°, #4 – 20°, #5 – 20°, #6 – 30°, #7 – 40°, #8 – 40°. These angles are an indication of the baselines used, and are listed in Figure 2.

The recorded images have a spatial resolution of 1200 × 1600 pixels recorded as 8 bit RGB converted to gray scale. The data set was recorded with an industrial robot arm using a setup similar to [1, 17]. We have, however, mounted the structured light scanner on the robot arm holding the camera. Hereby the ground truth 3D point-set was perfectly aligned with the camera position and provides a good coverage of the scenes. This enabled us to evaluate multiple view stereo algorithms by measuring the distance from the ground truth points of the structured light scan to the multi view reconstruction.

The average reconstruction errors and standard deviations are shown in Table 1 and the graph of averge reconstruction errors is illustrated in Figure 3. One reconstruction

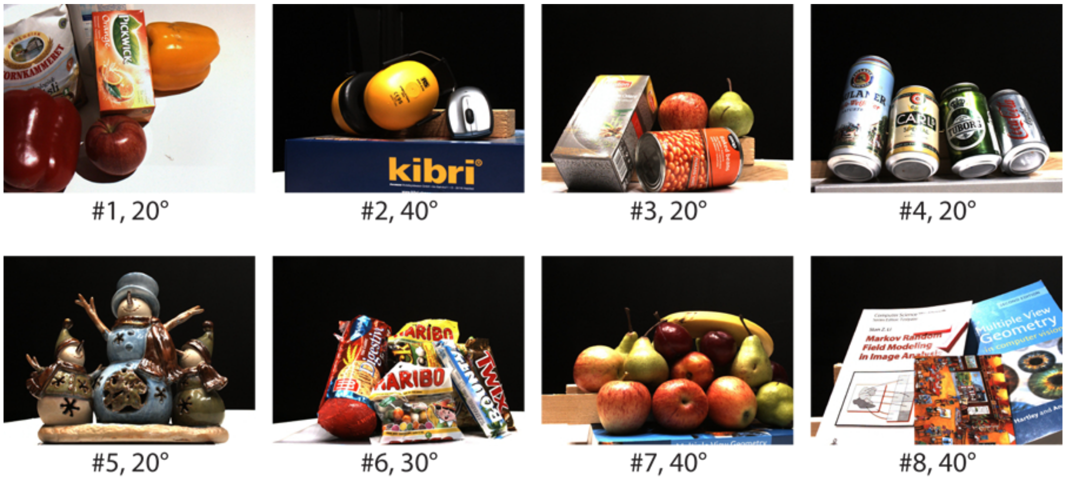

Figure 2. The scenes of our investigation numbered #1 - #8. The numbers after the comma indicate the baseline in degrees.

example is shown in Figure 4. The reconstruction errors were computed by taking the absolute difference between the estimated depths and ground truth for each pixel, but only where there were ground truth measurements.

To get a clearer picture of the performance of the visual metrics, we have solved the multiple view stereo reconstruction optimization problem (1) via the alpha-expansion algorithm of Boykov et al. [2], which is a very well understood optimization algorithm. For the same reason we have also avoided iterating over a visibility mask as done in [21]. In this way, we avoid complicating factors that impair the evaluation of the visual metrics.

The algorithm of [2] works by finding an optimal depth for each pixel in a reference image, where the depth is taken from a discrete set of ordered depth values. The depth resolutions used for the different scenes are determined by range of the ground truth data points and divided into equal sized steps of approximately 1 mm. This resulted in between 110 and 180 discrete steps in the different scenes.

We evaluate six different DAISY based visual metrics. The $\mathbf{d}_i(\mathbf{x}, \mathbf{n})$ is computed similarly to the DAISY descriptor in [21]. We compute the descriptor on a 31×31 pixels[6] image patch with three spatial sampling rings of six positions resulting in 19 spatial sampling positions. At each position the eight smoothed signed derivatives are sampled resulting in a 152 dimensional descriptor. The smoothing factor of the center point and first ring is $\sigma = 3$, for the second ring $\sigma = 5.5$, and for the third ring $\sigma = 8$. The raw based visual metrics evaluated are $J(\mathbf{x}, \mathbf{n})$ from (3) with a patch size of 11×11 and 31×31. The first is chosen because it is the recommendation by Jin et al. [14], the second is chosen in order to have the same terms as the DAISY based visual metrics. In the following we denote these two visual metrics as $J^{11}(\mathbf{x}, \mathbf{n})$ and $J^{31}(\mathbf{x}, \mathbf{n})$ respectively.

A summary of experimental results is shown in Table 1, Figure 3 and Figure 4. From the quantified errors in Table 1 several things can be concluded. Firstly, the DAISY based visual metrics outperform the raw based visual metrics, $J^{11}(\mathbf{x}, \mathbf{n})$ and $J^{31}(\mathbf{x}, \mathbf{n})$, by a large margin. This is clear evidence that a DAISY based visual metric should be preferred supporting the findings of [21]. Also $J^{31}(\mathbf{x}, \mathbf{n})$ consistently outperforms $J^{11}(\mathbf{x}, \mathbf{n})$.

We also note that the best performing descriptor varies between the two 2-dof DAISY descriptors, D_2 and M_2, and the D_2 favors the data sets with the small baselines. This indicates that the minimal cases are better at dealing with perspective distortion, and this is more important than a good depth to baseline ratio.

5. Perspective and Conclusion

In this paper, we have linked surface reflectance models with the visual metrics used for multiple view stereo. We conclude that we need more observations for a visual metric than the dof of an underlying reflectance model. Thus, more than two observations are needed to handle (partly) specular objects. We proceeded by proposing a method for including more than two images or observations into a visual metric, while incorporating the DAISY framework. This proved superior to directly using raw pixels regarding the ability to approximate the ground truth of our data. This is consistent with findings for salient feature matching [7] and two view stereo [21].

We have also put forth a new multiple view data set with ground truth, which spans different reflectance models better than any available data set we are aware of. This data set is the basis of our experimental evaluations. The evaluations, first of all, consider the dof of the underlying reflectance model. Our experiments also address whether the visual metric should be aggregated from a minimal set of

[6] In this case $m = 31 \times 31 = 961$.

$V(\mathbf{x},\mathbf{n})$	Scene #1		Scene #2		Scene #3		Scene #4	
	mean	std.	mean	std.	mean	std.	mean	std.
$D_1(\mathbf{x},\mathbf{n})$	10.67	22.55	10.79	17.43	4.99	13.59	4.18	4.69
$D_2(\mathbf{x},\mathbf{n})$	8.34	19.28	10.77	19.14	**3.82**	10.64	**3.59**	3.90
$D_{1.5}(\mathbf{x},\mathbf{n})$	9.70	21.68	11.07	18.83	4.25	12.75	3.91	4.23
$M_1(\mathbf{x},\mathbf{n})$	7.00	15.93	8.69	15.36	5.56	13.86	3.80	4.31
$M_2(\mathbf{x},\mathbf{n})$	**6.24**	14.08	**6.46**	12.47	5.35	13.52	4.01	4.78
$M_{1.5}(\mathbf{x},\mathbf{n})$	7.12	15.84	9.01	15.98	5.00	12.41	3.74	3.79
$J^{11}(\mathbf{x},\mathbf{n})$	29.45	42.91	32.67	33.93	18.79	31.11	10.16	21.78
$J^{31}(\mathbf{x},\mathbf{n})$	20.84	38.05	23.57	30.80	15.39	29.19	4.40	12.40

$V(\mathbf{x},\mathbf{n})$	Scene #5		Scene #6		Scene #7		Scene #8	
	mean	std.	mean	std.	mean	std.	mean	std.
$D_1(\mathbf{x},\mathbf{n})$	2.32	4.51	4.06	12.56	6.22	13.31	12.36	29.94
$D_2(\mathbf{x},\mathbf{n})$	**1.77**	3.28	2.68	6.45	4.53	10.99	11.94	29.87
$D_{1.5}(\mathbf{x},\mathbf{n})$	2.16	4.44	3.44	10.42	5.62	12.56	12.23	30.02
$M_1(\mathbf{x},\mathbf{n})$	1.90	2.61	2.76	6.15	4.22	9.16	12.30	29.87
$M_2(\mathbf{x},\mathbf{n})$	1.83	2.38	**2.65**	5.78	**3.90**	9.24	**10.59**	27.77
$M_{1.5}(\mathbf{x},\mathbf{n})$	1.91	2.90	2.70	6.36	4.41	9.44	10.76	27.77
$J^{11}(\mathbf{x},\mathbf{n})$	15.31	26.34	29.22	44.75	17.62	31.70	33.51	46.00
$J^{31}(\mathbf{x},\mathbf{n})$	10.10	22.09	17.65	36.90	13.62	26.66	20.51	38.79

Table 1. Average reconstruction errors and standard deviation (in mm) for the eight visual metrics and eight scenes. Note that the reported standard deviation is for the errors and not for the mean. If we assume a few hundred *independent* observations, the main differences between the means are significant. The ground truth consists of about 300.000 correlated observations, so a few hundred independent observations seems a reasonable assumption. The fact that the stdandard deviation is larger than the mean is a consequence of the reconstruction errors following a very skew distribution with a very fat tail in the direction of large errors. For each scene, the best mean value is denoted by **bold face**.

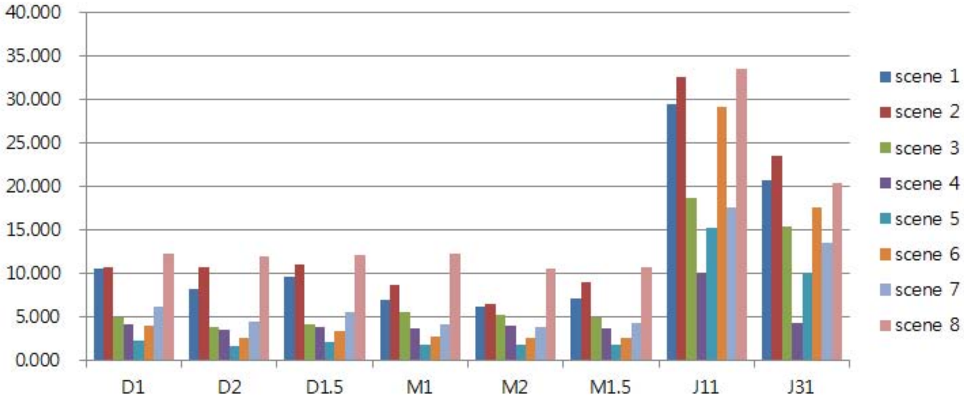

Figure 3. Graph of mean errors in Table 1. It shows the average reconstruction errors (in mm) on the vertical line and eight metrics on the horizontal line with eight scenes with different color. Note that DAISY based visual metrics are superior to the raw based visual metrics by a large margin, and M_2 is slightly better than other DAISY based metrics. The effect of subtle differences among DAISY based visual metrics can be seen in Figure 4.

images, as done with NCC in [3], or if all relevant images should be used directly as in [14]. Our experimental results show that the use of two dof is favorable. The choice between all or the minimal case seems to depend on the baseline – with a small baseline favoring using all images. As argued in the introduction, elaborate visual metrics are mostly needed for limited image budget, and thus large baselines, favoring M_2.

Since the state of the art in visual metrics [21, 14] is also represented in the visual metrics we investigated, the M_2 proposed here looks like a strong choice for a visual metric in relation to multiple view stereo. To further argue the

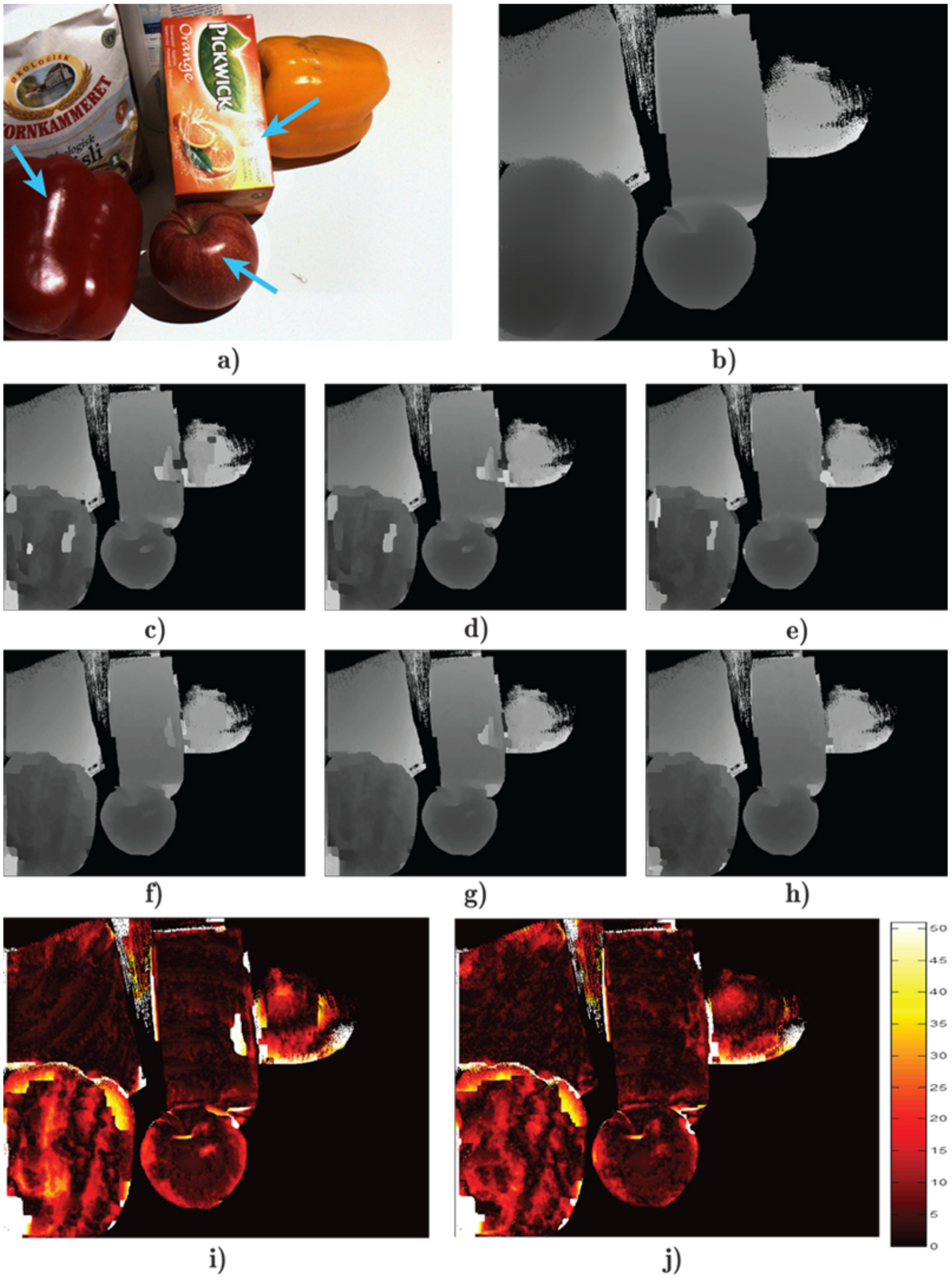

Figure 4. The reconstruction results of Scene #1, wrt. the DAISY based visual metrics. The figures illustrate **a)** The sample input image with blue arrows marking the distinct specularities to notice in the results. **b)** The ground truth. **c) - h)** reconstructed depth maps by the following visual metrics D_1, $D_{1.5}$, D_2, M_1, $M_{1.5}$, and M_2 respectively. **i)** and **j)** the reconstruction errors of M_1 and M_2 respectively, i.e. f) and h) minus b). Note the differences around the specularities.

matter in relation to robustness, e.g. occlusions, some of the current good choices of addressing this [3, 22] use minimal cases, and thus our M_2 visual metric should be usable in these robust frameworks.

Our findings favor basing visual metrics on underlying surface reflectance models. This opens the new interesting question of how these models should be formulated. In this work, we have limited these reflectance models to be linear subspaces to avoid a combinatorial explosion since we already compared eight visual metrics. It is, however, likely that other models, e.g. more physical based models comprising nonlinear manifolds, would perform better. In this regard the work of [4] is inspirational. Also it is likely that probabilistic models of the reflectance should be formulated, but this would require much more than eight scenes.

References

[1] H. Aanæs, A. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *International Journal of Computer Vision*, pages 1–18, 2011. 4

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. 2, 5

[3] N. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779, 2008. 6, 7

[4] M. Chandraker and R. Ramamoorthi. What an image reveals about material reflectance. In *IEEE International Conference on Computer Vision*, 2011. 8

[5] R. Cipolla, S. Battiato, and G. Farinella. *Computer Vision: Detection, Recognition and Reconstruction*. Studies in Computational Intelligence. Springer, 2010. 2

[6] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1161–1174, 2011. 1

[7] A. Dahl, H. Aanæs, and K. Pedersen. Finding the best feature detector-descriptor combination. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 318–325, 2011. 2, 3, 5

[8] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *IEEE Computer Vision and Pattern Recognition*, pages 1434–1441, 2010. 1

[9] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8), 2010. 1

[10] D. Gallup, J. Frahm, P. Mordohai, and M. Pollefeys. Variable baseline/resolution stereo. In *IEEE Computer Vision and Pattern Recognition*, 2008. 4

[11] D. Gallup, J. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2007. 1

[12] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz. Multi-view stereo for community photo collections. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007. 1

[13] H. Hirschmuller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2009. 2

[14] H. Jin, S. Soatto, and A. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *International Journal of Computer Vision*, 63(3):175–189, 2005. 1, 2, 3, 4, 5, 6

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2, 3

[16] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007. 1

[17] S. Seitz, B. Curless, J. Deiabel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Computer Vision and Pattern Recognition*, pages 519–528, 2006. 1, 2, 4

[18] S. Stefano, A. Anthony, and J. Hailin. Tales of shape and radiance in multi-view stereo. In *IEEE International Conference on Computer Vision*, 2003. 1, 3

[19] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1

[20] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *IEEE Computer Vision and Pattern Recognition*, 2008. 3

[21] E. Tola, V. Lepetit, and P. Fua. Daisy: an efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, 2010. 2, 3, 5, 6

[22] G. Vogiatzis, C. Hernandez, P. Torr, and R. Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12), 2007. 1, 4, 7

[23] H. Vu, R. Keriven, P. Labatut, and J. P. Pons. Towards high-resolution large-scale multi-view stereo. In *IEEE Computer Vision and Pattern Recognition*, pages 1430–1437, 2009. 1

[24] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007. 1