

Shape from Probability Maps with Image-Adapted Voxelization

Jordi Salvador and Josep R. Casas

Image Group – Technical University of Catalonia (UPC)

Abstract. This paper presents a Bayesian framework for Visual Hull reconstruction from multiple camera views with a 3D sampling scheme based on an irregular 3D grid, which becomes regular once projected onto the available views. The probabilistic framework consists in establishing a foreground probability for each pixel in each view rather than segmenting in order to obtain binary silhouettes of the foreground elements. Next, a Bayesian consistency test labels the occupancy of each image-adapted 3D sample. The proposed method, using image-adapted 3D sampling in the Bayesian framework, is compared to a shape-from-silhouette implementation with image-adapted voxelization, where the input data are binary silhouettes instead of probability maps; we also compare its performance to a state-of-the-art method based on regular 3D sampling with binary silhouettes and SPOT projection test.

1 Introduction

When trying to extract the Visual Hull [1] from images captured by a set of calibrated cameras, numerous approaches consist in, firstly, obtaining a silhouette image of the foreground object for which we want to obtain the Visual Hull and, secondly, applying a shape-from-silhouette algorithm on the segmented images (silhouettes), typically labelling elements of a regular 3D sampling grid (voxels) as being occupied or empty.

1.1 2D Foreground Segmentation

An extended method to extract foreground silhouettes from an image is to model the color properties of each background pixel as an independent Gaussian random variable and classify pixels in new images as foreground or background depending on their similarity to the generated pixel models [2]. For this, training images containing only background should be available.

The obtaining of silhouette images with this method is error-prone in certain scenarios, leading to silhouette images with mislabeled pixels where both false detections, in highlighted or shadowed areas, and misses, when foreground colors are highly similar to background ones, occur.

As a possible solution, we propose avoiding binary thresholding at pixel level when detecting foreground silhouettes. Instead, a foreground probability map can be obtained and used in a probabilistic Visual Hull computation.

1.2 Shape from Silhouette

Typically, shape-from-silhouette implementations rely, as many other multi-view analysis techniques, on a regular voxelization of 3D space, which is non-adapted to image data. In settings such as a *smart room* with a small number of evenly distributed cameras, voxels project to unbalanced sets of image data in each camera, from which analysis algorithms, including shape-from-silhouette, have to work out feature matches and consistency checks.

A common strategy to compute the voxelized Visual Hull from a set of silhouettes via shape-from-silhouette is to define a projection test where the center of each voxel is projected onto the images in order to be tested, assuming small or distant voxels. This method has the advantage of its low computational cost, but an important drawback is the undersampling of the actual voxel projections, which leads to wrong decisions, mainly in contours. An alternative method that shows better performance is the SPOT projection test [3], which takes samples of the actual voxel's projection at random positions and considers aspects such as the distance from a voxel to a camera or the accuracy of the binary silhouettes.

Image-Adapted Voxelization. An alternative strategy to regular voxelization is image-based scanning of the 3D scene. Matusik introduced the concept of image-based Visual Hulls to render an observed scene in real time from a virtual camera point of view without constructing an explicit auxiliary volumetric representation [4]. He claims that the advantage of performing geometric computations based on epipolar geometry in image space is the elimination of resampling and quantization artifacts in other volumetric approaches.

Image-adapted voxelization [5] follows the concept of image-based processing, but focusing on volumetric analysis applications. In particular, an image-based progressive scanning algorithm for multi-view analysis is used, and the corresponding geometry in 3D space is derived. This provides a volumetric representation for image data functionally equivalent to regular voxelization as volumetric data support for analysis algorithms, with the benefit that the 3D scanning procedure is better adapted to the image data than regular voxelization. The principles underlying this technique and its associated algorithm with detailed explanations can be found in [6] and [5].

Bayesian framework. The method proposed in this paper, *Shape from Probability Maps*, relies on the image-adapted 3D scanning procedure described above, extrapolating the shape-from-silhouette algorithm for Visual Hull reconstruction to a probabilistic framework with Bayesian formulation.

1.3 Document Organization

The document is organized as follows: in Section 2, we introduce the formulation and principles of shape-from-probability-maps. Then, in Section 3, the method is compared to two different implementations of standard shape-from-silhouette, one

based in image-adapted voxelization and the other in regular voxelization with SPOT projection test. To conclude, the main features of shape-from-probability-maps are summarized.

2 Shape from Probability Maps

In order to obtain a foreground segmentation of an image when background images are available, a simple yet functional approach consists in:

1. modelling each background pixel as a Gaussian random variable which describes its color, with mean and variance extracted from the background images
2. Classifying pixels from new images, testing whether they fit in the background model or not

Furthermore, the previous approach allows updating the background models with color information from pixels classified as belonging to the scene's background. This extension allows correctly segmenting images in scenarios where, for example, light conditions suffer slow variations.

An undesired characteristic of the method is that background pixels that are shadowed by the presence of a new element in the scene are also classified as belonging to the scene projection's foreground. Furthermore, foreground pixels which were highlighted because of surface reflections when the background model was generated, can also be classified as foreground pixels if the light ray that originated the highlight is blocked by some foreground element. For simplicity, we call these two phenomena *natural* and *blocking* shadows, respectively.

2.1 Foreground Detection with Natural and Blocking Shadow Modelling

A possible solution to this undesired behavior consists in modelling natural and blocking shadowed zones using a model similar to that used for background pixels. Thus, the proposed model for both of them is a scaling of the measured color when none of these two phenomena occurs. As a result, the probabilistic models for a background pixel's natural and blocking shadowed states consist in two gaussians with means and variances proportional to those of the corresponding *basic* background model. This approach for modelling shadows has already been used in [7]. Here, we generalize it for the blocking shadow case.

Let x_p be a $\{R, G, B\}$ pixel in an image and μ_p and σ_p^2 the mean and variance of its associated background model. We model natural background shadowing as a factor scale α_ν multiplying the original background color. Hence, the resulting model is a gaussian variable with mean $\alpha_\nu \times \mu_p$ and variance $\alpha_\nu^2 \times \sigma_p^2$. Equivalently, the blocking shadow background model becomes a gaussian variable with $\alpha_\beta \times \mu_p$ for the mean and $\alpha_\beta^2 \times \sigma_p^2$ for the variance. Intuitively, it is clear that α_β must be smaller than α_ν .

The resulting expression for the pixel's foreground probability depends on three priors (foreground, and natural and blocking shadowed background probabilities). A uniform random variable modelling foreground pixels, a normal or Gaussian model for each background pixel and two scaled versions of each background model for both natural and blocking shadowed background modelling are used to complete the pixel's foreground probability model:

$$p(f|x_p) = \frac{p(x_p|f)p(f)}{p(x_p|b)p(b) + p(x_p|\nu)p(\nu) + p(x_p|\beta)p(\beta) + p(x_p|f)p(f)}, \quad (1)$$

with $p(x_p|f)$ a uniform pdf and $p(x_p|b)$, $p(x_p|\nu)$ and $p(x_p|\beta)$ gaussian probability density functions for background, natural shadowed background and blocking shadowed background respectively. All the pdfs are in $\{R, G, B\}$ space.

This improvement of the 2D foreground detection procedure provides better background detection and, consequently, more reliable probability maps, although it does not constitute our main contribution.

2.2 Visual Hull Computation with A-priori Knowledge

The proposed analysis method relies on the progressive 3D scanning procedure previously introduced, an idea similar to that of octree-based analysis for voxels [8], but adapted to the available image settings and their associated calibration data.

The aim of the progressive scanning is to simplify the analysis in those areas where coarse resolutions are sufficient, either because of the emptiness of the scene in the feature space we are considering or because of the completeness of the observations of the features. On the opposite case, finer resolutions are used to scan space in those regions where the features measured from the available images do not show uniformity.

2.3 Consistency Test

The consistency test is responsible of marking a 3D region (like a voxel, in the case of regular voxelization) as belonging to the scene's foreground or background. In this case, the available data are groups of pixels in each camera view, with their associated pixel foreground probabilities.

As an extension to the 2D foreground segmentation procedure, we also propose a Bayesian formulation in order to describe the occupancy of a 3D region. Let X_c be the group of pixels containing the projection of an analyzed 3D region onto a camera c , then:

$$p(f_{3D}|X_1, \dots, X_{N_C}) = \frac{p(X_1, \dots, X_{N_C}|f_{3D})p(f_{3D})}{p(X_1, \dots, X_{N_C})}, \quad (2)$$

with

$$p(X_1, \dots, X_{N_C}) = p(X_1, \dots, X_{N_C}|f_{3D})p(f_{3D}) + p(X_1, \dots, X_{N_C}|b_{3D})p(b_{3D}), \quad (3)$$

where $p(f_{3D})$ and $p(b_{3D})$ stand for the 3D foreground probability and the 3D background probability, respectively, and N_C stands for the number of cameras. As a result, those 3D regions where Eq. 2 lies above 0.5 are marked as belonging to the Visual Hull of the scene.

3D Probability Model. The 3D foreground conditional probability is modelled as a function of the actual measured data: the cameras' foreground probabilities. We assume independence between the assigned probabilities in each view despite the foreground objects are the same in all views. This is reasonable since the foreground detection step in a view is not influenced the results of the foreground detection in other views. As a result, we model the foreground conditional probability of a 3D region as the product of each of the cameras' projection models:

$$p(X_1, \dots, X_{N_C} | f_{3D}) = \prod_{c=1}^{N_C} p(X_c | f_{3D}). \quad (4)$$

Equivalently, for the 3D background conditional probability:

$$p(X_1, \dots, X_{N_C} | b_{3D}) = \prod_{c=1}^{N_C} p(X_c | b_{3D}). \quad (5)$$

Camera Projection Model. We define the probabilistic model of the projection of a scanned 3D region on a camera image as the average foreground probability of the pixels lying inside of the projection. The 3D probabilities, conditioned to 3D background and foreground, result in

$$p(X_c | f_{3D}) = \frac{1}{N_p} \sum_{p=1}^{N_p} \max(p(f|x_p), \epsilon), \quad \epsilon \ll 1 \quad (6)$$

for 3D foreground, and

$$p(X_c | b_{3D}) = \frac{1}{N_p} \sum_{p=1}^{N_p} \max((1 - p(f|x_p)), \epsilon), \quad \epsilon \ll 1 \quad (7)$$

for 3D background, with N_p as the number of pixels lying inside of the projection of the 3D region onto the probability map, in general different for each camera c . The term ϵ is introduced in order to balance the effect of erroneous 2D segmentations as background when computing 3D foreground conditional probabilities.

3 Experiments and Results

In this section we test shape-from-probability-maps and compare it to standard shape-from-silhouette methods in two different configurations. In the first one, we use regular voxelization with SPOT as the projection test, whereas in the second one image-adapted 3D scanning is used.

3.1 Performance Metrics

In order to estimate the actual performance from indirect measurements in 2D projections, we propose to evaluate the system employing verification measures commonly used in information retrieval and also curves derived from these measures, obtained from the projection of the reconstructed Visual Hulls onto the original points of view of the captured images.

- *Recall* measures the ability of the method to detect a high number of foreground regions:

$$\text{Recall} = \frac{\#\text{correct foreground}}{\#\text{correct foreground} + \#\text{wrong background}} \quad (8)$$

- *Precision* measures the ability of the system to detect accurately the actual foreground regions:

$$\text{Precision} = \frac{\#\text{correct foreground}}{\#\text{correct foreground} + \#\text{wrong foreground}} \quad (9)$$

- *F-measure*, also known as the harmonic mean of the two previous quantities, measures the combination of Precision and Recall and gives a more reliable measure of the performance of the system:

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (10)$$

3.2 Data Preparation

The data set was chosen to allow quantitatively measuring the performance by using the proposed metric.

Testing Data. The testing data is composed by ten sets of five images captured in a *smart room*. Four of the cameras are installed at the corners of the room and the fifth is configured as a zenithal camera, with a focal length shorter than that in the corner cameras. For all the cameras the calibration parameters were obtained as in [9], and the distortion parameters were used to obtain undistorted versions of the captured images. In Figure 1, both one of the captured images and its undistorted version are shown.

Groundtruth. The groundtruth is composed by ten sets of five binary silhouettes obtained by manual segmentation of the undistorted images. In Figure 5 (d), the groundtruth silhouette for one of the undistorted images is shown.



Fig. 1. A captured image as seen from one of the cameras (a) and its corresponding undistorted version (b)

Camera's Background Models Generation. In addition, 200 frames from each camera were captured with the room completely empty. These training frames can be used to model each camera's background. The parameters that we used for the model generation were a learning rate $\rho = 0.01$ and an initial variance $\sigma^2 = 20$. This initial variance equals the minimum variance accepted for modelling each pixel. This way we ensure that the resulting Gaussian model will have a slightly smoother slope than the one we would obtain by not setting a minimum variance when background pixels show little fluctuations between frames.

The prior probabilities have been arbitrarily assigned in order to be able to obtain reasonably accurate silhouettes, while keeping a realistic scenario including typical artifacts (misses and false detections) that appear when segmenting images with a method based on each pixels' background models. The chosen default values for these priors are $p(f) = 0.2$ for the foreground probability, $p(\nu) = 0.15$ for the natural shadow probability and $p(\beta) = 0.15$ for the blocking shadow probability. The resulting background prior is $1 - p(f) - p(\nu) - p(\beta) = 0.5$. The chosen scaling factors for modelling natural and blocking shadowed background are $\alpha_\nu = 0.8$ and $\alpha_\beta = 0.65$, respectively.

3.3 Shape from Silhouette

In order to compute the Visual Hull with this method, either with regular or image-adapted voxelization, it is necessary to provide binary images (silhouettes) marking each pixel of the captured projections of the 3D scene as belonging to foreground or background. We decide for thresholding the foreground probability maps with a confidence threshold equal to 0.5. After this step, we obtain 2D silhouettes showing both misses and false detections that can be used for computing the foreground scene's Visual Hull. A 2D silhouette obtained by this method is shown in Figure 2.

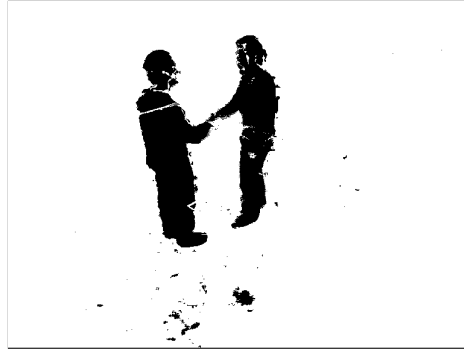


Fig. 2. Binary segmented silhouette for one of the captured images

For a fair comparison (actually slightly favorable to regular voxelization), an intra-voxel distance of 1 cm is chosen as equivalent to the maximal resolution (3×3 pixels) used in the image-adapted voxelization algorithm.

Regular Voxelization with SPOT Projection Test. The best performance for this method with the given configuration is obtained with a number of tests per projection equal to 20, with a false alarm rate equal to 0.1, a miss rate also equal to 0.1 and a foreground probability equal to 0.01. From these values, SPOT's optimal threshold was obtained as specified by its author. The results of this best configuration, averaged over 10 scenes with 5 images each, are shown in Table 1. The reconstruction of the Visual Hull for one of the scenes from

Table 1. SfS with SPOT projection test for regular voxelization's average performance

Precision	Recall	F-measure
87.34%	81.01%	83.61%

a virtual camera position is shown in Figure 3. Note that severe errors in 2D foreground segmentation cause splits in the reconstructed volumes.

Image-Adapted Voxelization. In this case, the projection test consists in checking if any of the pixels lying in the projection of the analyzed 3D region belongs to the silhouette's foreground, once the finest resolution of 3×3 pixels has been set. As a consistency test, when in each of the available images the projection test is positive, the 3D region is marked as belonging to the Visual Hull. The results of this configuration, again averaged over the same 10 scenes with 5 images each, are shown in Table 2. As it can be seen, this method slightly improves the performance of the previous one, thanks to the more balanced usage of the available data.



Fig. 3. Visual Hull reconstructed by shape-from-silhouette with regular voxelization and SPOT projection test

Table 2. SfS for image-adapted voxelization's average performance

Precision	Recall	F-measure
84.87%	84.37%	84.25%

3.4 Shape from Foreground Probability Maps

We have performed several runs, with different 3D foreground probability priors, in order to characterize in a higher level of detail the new method with its Precision-Recall curve. In order to better depict the obtained results, they are grouped in Figure 4.

We also include the averaged results for the several runs of the probabilistic method parameterized with different priors. They are listed in Table 3. As it can

Table 3. Shape from Foreground Probability Maps for image-adapted voxelization's average performance

Prior $p(f_{3D})$	Precision	Recall	F-measure
0.0005	75.32%	97.23%	84.71%
0.0004	75.51%	97.17%	84.81%
0.0003	75.83%	97.06%	84.97%
0.0002	81.45%	88.33%	84.32%
0.0001	85.20%	85.15%	84.79%
0.00005	87.51%	83.51%	85.08%
0.00003	89.08%	82.06%	85.04%
0.00001	91.16%	78.68%	84.05%

be seen, the performance of the new method is consistently above the results obtained with any of the two shape from silhouette methods included in the comparison. An exception appears with the last prior included, 0.00001, which

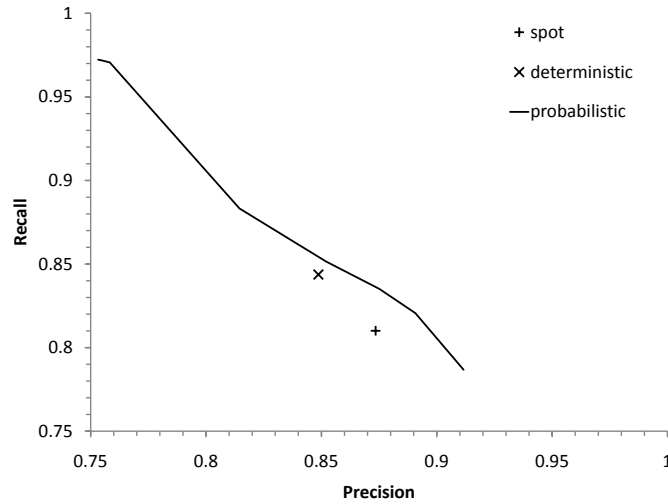


Fig. 4. Precision-Recall curve for different values of the prior 3D foreground probability $p(f_{3D})$. Results for SfS with image-adapted voxelization (*deterministic*) and SfS with regular voxelization and SPOT projection test (*spot*) are also used for comparison

is obviously too small to describe correctly the scene's actual 3D foreground probability.

In Figure 5, the projections of one of the reconstructed scenes onto one of the available cameras are shown for shape-from-silhouette with regular voxelization and SPOT projection test, shape-from-silhouette with image-adapted voxelization and shape-from-probability-maps with 3D foreground prior equal to 0.00005 (our best run). Please note that we have not applied contour smoothing in our reconstructions. As it can be seen, shape-from-probability-maps is able to reconstruct the right arm of the men at the left side of the images more accurately than the other methods.

Limits. As it can be seen, the maximum Recall for shape-from-probability-maps, obtained when 3D foreground priors are high, is clearly above the maximum achievable Precision, obtained when priors are low. This is consistent with the fact that the finest level of detail of the image-adapted voxelization scanning procedure is set to 3×3 pixels. This condition introduces some additional foreground pixels when projecting a 3D region which is in contact with the Visual Hull's surface onto the original cameras' points of view. Nevertheless, as the results with image-adapted and regular voxelization show, the performance of this 3D sampling method is superior to that of regular voxelization.

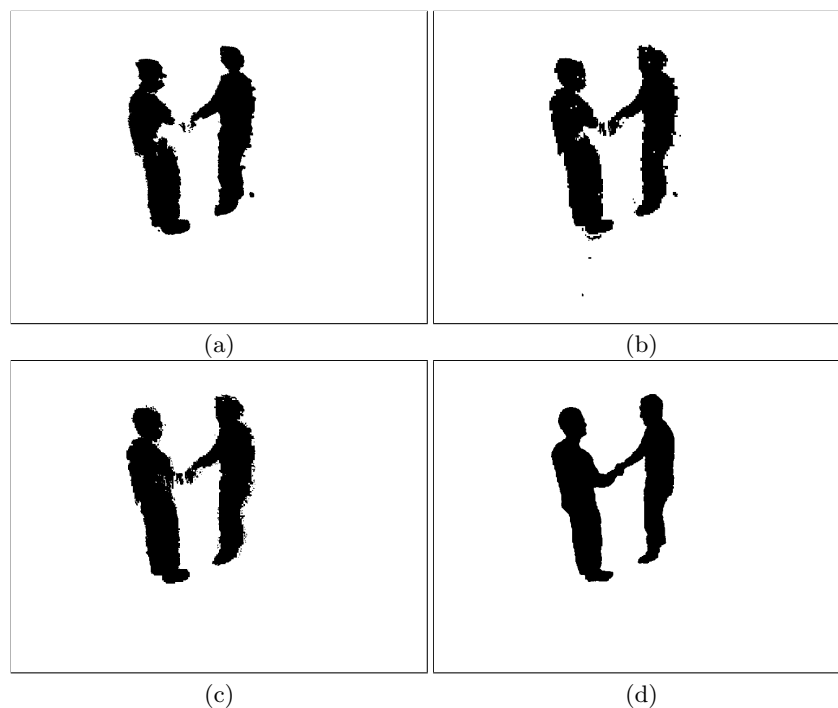


Fig. 5. Projections of one of the reconstructed scenes onto one of the available cameras for shape-from-silhouette with regular voxelization and SPOT projection test (a), shape-from-silhouette with image-adapted voxelization (b), and shape-from-probability-maps with $p(f_{3D}) = 0.00005$ (c). The corresponding groundtruth silhouette (d) is also shown

4 Conclusions

An alternative method to compute the Visual Hull of a 3D scene captured by several cameras, shape-from-probability-maps with image-adapted voxelization, has been presented. Its main features are:

1. the use of a 3D scanning algorithm better adapted to the available information than the commonly used regular voxelization, and
2. the introduction of a Bayesian formulation that avoids early image segmentation in favor of considering pixels as a foreground probability values

Compared to the standard shape-from-silhouette algorithm with voxelization adapted to images, the input data for the proposed method offer more information by not thresholding foreground detection results, delaying this decision to allow for contributions from the other views before the foreground-background classification. We have to point out that the probability maps obtained from 2D foreground detection are rather binary, allowing for just a marginal gain when comparing our new method with shape-from-silhouette. Thus, the next steps in

order to exploit the proposed Bayesian framework should be in the direction of enhancing the foreground detection. Comparison with regular voxelization shows that the performance of image-adapted 3D analysis also goes beyond that of regular voxelization, due to the more balanced usage of available image data, which does not introduce an additional arbitrary sampling in 3D space. To sum up, as seen in the results provided, shape-from-probability-maps delivers a *Recall-Precision* curve above the performance of any of the other state-of-the-art methods included in the comparison.

This work has been partially supported by the Spanish Ministerio de Educación y Ciencia, under project TEC2007-66858/TCM, and by the Spanish Administration agency CDTI, under project CENIT-VISION 2007-1007. The authors would also like to thank Marcel Alcoverro for his help with parameters optimization in the tests with SPOT projection test for regular voxelization.

References

1. Laurentini, A.: The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16** (1994) 150–162
2. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2** (1999) 2246
3. Cheung, K.: Visual Hull Construction, Alignment and Refinement for Human Kinematic Modeling, Motion Tracking and Rendering. PhD thesis, Carnegie Mellon University (2003)
4. Matusik, W., Buehler, C., Raskar, R., Gortler, S., McMillan, L.: Image-based visual hulls. In: *Proc. 27th conference on Computer Graphics and interactive techniques.* (2000) 369–374
5. Casas, J., Salvador, J.: Image-based multi-view scene analysis using 'conexels'. In: *Proceedings of the HCSNet workshop on Use of vision in human-computer interaction.* Volume 56. (2006) 19–28
6. Salvador, J., Casas, J.R.: Image-adapted voxelization in multicamera settings. In: *Multimedia Signal Processing, 2006 IEEE 8th Workshop on.* (2006) 161–165
7. Wang, Y., Tan, T., Loe, K., Wu, J.: A probabilistic approach for foreground and shadow segmentation in monocular image sequences. *Journal of the Pattern Recognition Society* **38** (2005) 1937–1946
8. Meagher, D.: Geometric modeling using octree encoding. *Computer Graphics Image Process* **19** (1982) 129–147
9. Bouguet, J.: Camera calibration toolbox for matlab, http://www.vision.caltech.edu/bouguetj/calib_doc (2000)