Improving Object Recognition by Fusion of Multiple Views

Frank Deinzer*, Joachim Denzler, Heinrich Niemann
Chair for Pattern Recognition, Department of Computer Science
University Erlangen-Nürnberg, Martensstr.3, 91058 Erlangen, Germany
{deinzer,denzler}@informatik.uni-erlangen.de

Abstract

In the past decades most object recognition systems were based on passive approaches. But in the last few years a lot of research was done in the field of active approaches for object recognition. In this context there are several unique problems to be solved. One of them is how to fuse images from several viewpoints.

In this paper we present a well-founded approach for the fusion of multiple views based on a recursive density propagation problem. It uses the well-known CONDENSATION algorithm for solving the fusion in a continuous pose space.

Our experimental result we will show how the fusion can improve classification rates substantial especially for difficult conditions like heterogeneous background and not perfectly suited or weak classifiers.

1. Introduction

Object recognition has been tackled by passive approaches in the past. This means that based on one image a decision for a certain class and pose must be made or the image must be rejected. This neglects both facts that some other views might exist, which allow for a more reliable classification and that the fusion of multiple views might improve the recognition rates noticeable. This is one reason, why research has focused on active object recognition over the past years [6, 12, 2, 10, 3, 13].

One of the most important aspects in active object recognition is the fusion of a sequence of images taken from different viewpoints to obtain an overall classification and localization result. It is comprehensible that the fusion will not only be helpful for ambiguous objects (for which more than one view might be necessary to resolve the ambiguity) but will improve recognition rate in general. This is especially true for real world environments where one has to deal with, for example, heterogeneous background.

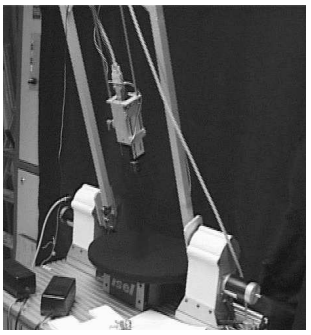
In this paper we present of a general fusion scheme based on the CONDENSATION algorithm [7]. There are three main reasons for applying the CONDENSATION algorithm. First, one has to deal inherently with multimodal distributions over the class and pose space of the objects. Second, moving the camera from one viewpoint to another will add uncertainty in the fusion process as the movement of the camera will always be disturbed by noise. Thus this uncertainty must be taken into account in the fusion process of the results acquired so far with the results computed from the current image. Third, it is not straight forward to model the involved probability distributions in closed form, especially if multiple hypothesis, i.e. multimodal distributions, shall be handled. These three aspects are strong criterions that the CONDENSATION algorithm is perfectly suited for the fusion of views in active object recognition. Especially, the ability to handle dynamic systems is advantageous because in viewpoint fusion the dynamics is given by the known but noisy camera motion between two viewpoints.

In Section 2, we will present the theoretical background of our approach base on the CONDENSATION algorithm and we will show how to apply our sensor data fusion to a real world object recognition system. The performed experiments and an introduction to the classifier used in the experiments are presented in Section 3. This section shows the practicability of our method in the context of classification of objects with heterogeneous background. Section 4 will close this paper with a conclusion and a short outlook to further investigations.

2 Fusion of Multiple Views

Active object recognition extends the classic passive approach in a manner that object classification and localization of an object is based on a sequence or series of images. These images are used to improve the robustness and reliability of the object classification and localization. In this active approach object recognition is not simply a task of repeated classification and localization for each image, but a well directed combination of a funded fusion of images.

*This work was partially funded by the German Science Foundation (DFG) under grant SFB 603/TP B2. Only the authors are responsible for the content.



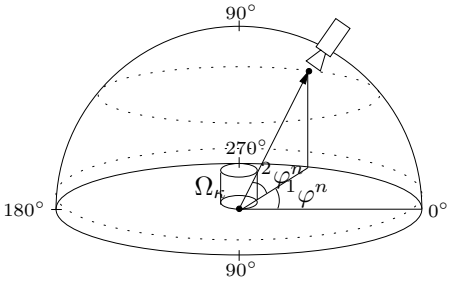


Figure 1: Experimental setup consists of turntable/arm combination. The possible pose space is defined by the azimuthal ${}^1\varphi^n$ and the latitude ${}^2\varphi^n$.

2.1 Density Propagation with the Condensation Algorithm

In active object recognition a series of observed images $\boldsymbol{f}_n, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{f}_0$ of an object are given together with the camera movements $\boldsymbol{a}_{n-1}, \ldots, \boldsymbol{a}_0$ between these images. Based on these observations of images and movements one wants to draw conclusions for a non-observable state $\boldsymbol{q}_n$ of the object. This state $\boldsymbol{q}_n$ must contain both the *discrete* class and the *continuous* pose of the object. This fact is important for the further proceeding.

In the context of a Bayesian approach, the knowledge on the object's state is given in form of the a posteriori density $p(\boldsymbol{q}_n|\boldsymbol{f}_n, \boldsymbol{a}_{n-1}, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0)$ and can be calculated from

$$p(\boldsymbol{q}_n|\boldsymbol{f}_n, \boldsymbol{a}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0) = \frac{1}{k_n} p(\boldsymbol{q}_n|\boldsymbol{a}_{n-1}, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0) p(\boldsymbol{f}_n|\boldsymbol{q}_n) \quad (1)$$

where

$$k_n = p(\boldsymbol{f}_n, \boldsymbol{a}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0)$$

denotes a normalizing constant that is left out in the following considerations. Under the Markov assumption

$$p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}, \ldots, \boldsymbol{q}_0, \boldsymbol{a}_0) = p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1})$$

for the state transition, equation (1) can be recursively rewritten as

$$p(\boldsymbol{q}_n|\boldsymbol{a}_{n-1}, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0) = \int_{\boldsymbol{q}_{n-1}} p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}) \cdot p(\boldsymbol{q}_{n-1}|\boldsymbol{a}_{n-1}, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0) d\boldsymbol{q}_{n-1} \quad (2)$$

It is obvious that this probability depends only on the camera movement $\boldsymbol{a}_{n-1}$. Its inaccuracy is modeled with a normally distributed noise component. Consequently the state transition probability can be written as

$$p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}) = \mathcal{N}(\boldsymbol{q}_{n-1} + \boldsymbol{a}_{n-1}, \Sigma)$$

with the covariance matrix Σ of the inaccuracy of the camera movement. This covariance matrix has to be estimated in advance by experiments or has to be set heuristically.

If continuous components in the state $\boldsymbol{q}_n$ can be avoided, the integral in equation (2) can be simplified to

$$p(\boldsymbol{q}_n|\boldsymbol{f}_{n-1}, \ldots, \boldsymbol{f}_0) = \sum_{\boldsymbol{q}_{n-1}} p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}) p(\boldsymbol{q}_{n-1}|\boldsymbol{f}_{n-1}, \ldots, \boldsymbol{f}_0) \quad (3)$$

and can easily be evaluated in an analytical way. For example, to classify an object Ω_κ in a sequence of images with

$$\boldsymbol{q}_n = \left(\ \Omega_\kappa \ \right),$$

$p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1})$ in equation 3 degrades to

$$p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}) = \begin{cases} 1 & \text{if } \boldsymbol{q}_n = \boldsymbol{q}_{n-1} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

since the object class does not change if the camera is moved, and consequently equation (3) must have an analytically solution.

But if one wants to use the fusion of multiple views in a general way with the possibility of continuous pose parameters in $\boldsymbol{q}_n$ it is no longer possible to simplify equation (2) to equation (3).

The classic approach for solving this recursive density propagation is the Kalman Filter [8, 1]. But in computer vision the necessary assumptions for the Kalman Filter ($p(\boldsymbol{f}_n|\boldsymbol{q}_n)$ being normally distributed) are often not valid. In real world applications this density $p(\boldsymbol{f}_n|\boldsymbol{q}_n)$ usually is not normally distributed due to object ambiguities, sensor noise, occlusion, etc. This is a problem since it leads to a distribution which is not analytically computable. An approach for the complicated handling of such multimodal densities are the so called particle filters. The basic idea is to approximate the a posteriori density by a set of weighted particles. In our approach we use the CONDENSATION algorithm (CONditional DENSity propaATION) [7]. It uses a

Figure 2: Objects of the data set within heterogenous background: Coke gray, Coke red, stapler green, stapler white, hole punch green, hole punch red, NaCl-bottle, pillbox, cup, cup with saucer.

sample set $C_n = \{\boldsymbol{c}_1^n, \ldots, \boldsymbol{c}_K^n\}$ to approximate the multimodal probability distribution in equation (1). Please note that we do not only have a continuous state space for $\boldsymbol{q}_n$ but a *mixed discrete/continuous state space* for object class and pose as mentioned at the beginning of this section. The practical procedure of applying the CONDENSATION to the fusion problem is illustrated in the next section.

2.2 Fusion of Multiple Views with the Condensation Algorithm

After the presentation of the theoretical background we will show how to use the CONDENSATION algorithm in a practical realization of sensor data fusion of multiple views.

As noted in section 2.1 we need to include the class and pose of the object into our state $\boldsymbol{q}_n$ to classify and localize objects. This leads to the following definitions of the state

$$\boldsymbol{q}_n = \begin{pmatrix} \Omega_\kappa \\ {}^1\varphi^n \\ \vdots \\ {}^J\varphi^n \end{pmatrix} \tag{5}$$

and the samples

$$\boldsymbol{c}_i^n = \begin{pmatrix} \Omega_\kappa \\ {}^1\varphi_i^n \\ \vdots \\ {}^J\varphi_i^n \end{pmatrix} \tag{6}$$

where ${}^j\varphi^n$ denotes the pose of the j-th degree of freedom for the camera position. The camera movements are defined accordingly as

$$\boldsymbol{a}_n = \begin{pmatrix} \Delta^1\varphi^n \\ \vdots \\ \Delta^J\varphi^n \end{pmatrix} \tag{7}$$

with $\Delta^j\varphi^n$ denoting the relative changes of the viewing position of the camera.

In our experimental setup (see Figure 1) we have only two degrees of freedom. The camera can move on a half-sphere around the object with the azimuthal ${}^1\varphi^n \in [0°; 360°)$ and the latitude ${}^2\varphi^n \in [0°; 90°]$.

In the practical realization of the CONDENSATION, one starts with an initial sample set $C^0 = \{\boldsymbol{c}_1^0, \ldots, \boldsymbol{c}_K^0\}$ with samples distributed uniformly over the state space. For the generation of a new sample set C^n, samples $\boldsymbol{c}_i^n$ are

1. drawn from C^{n-1} with probability

$$\frac{p(\boldsymbol{f}_{n-1}|\boldsymbol{c}_i^{n-1})}{\sum_{j=1}^{K} p(\boldsymbol{f}_{n-1}|\boldsymbol{c}_j^{n-1})} \tag{8}$$

2. propagated with the necessarily predetermined sample transition model

$$\boldsymbol{c}_i^n = \boldsymbol{c}_i^{n-1} + \begin{pmatrix} 0 \\ r_1 \\ \vdots \\ r_J \end{pmatrix} \tag{9}$$

   with $r_j \sim \mathcal{N}(\Delta^j\varphi^n, \sigma_j)$ and the variance parameters of the Gaussian transition noise σ_j. They model the inaccuracy of the camera movement under the assumption that the errors of the camera movements are independent between the degrees of freedom. These variance parameters have to be defined in advance.

3. evaluated in the image by $p(\boldsymbol{f}_n|\boldsymbol{c}_i^n)$. This evaluation is performed by the classifier. The only need to the classifier that shall be used together with our fusion approach must be its ability to evaluate this density.

For a more detailed explanation on the theoretical background of the approximation of equation (1) by the sample set C^N we would like to refer to [7].

At this point we want to note that it is important to include the class Ω_κ into the object state $\boldsymbol{q}_n$ and the samples $\boldsymbol{c}_i^n$. An alternative would be to omit this by setting up several sample sets – one for each object class – and perform the CONDENSATION separately on each set. But this would not result in an integrated classification/localization, but in separated localizations on each set under the assumption of observing the corresponding object class. No fusion of the object class over the sequence of images would be done in that case.

3. Experimental Evaluation

In this section we will present our experimental results. We will also introduce our used classifier to evaluate $p(\boldsymbol{f}_n|\boldsymbol{c}_i^n)$ from Section 2.2 and will shortly describe our data set.

3.1 Statistical Eigenspace

Currently we are evaluating our approach for viewpoint selection with a statistical variation of a classifier based on the eigenspace approach introduced by [9] which is similar to [2].

Object recognition is dealing with assigning a class number Ω_κ to an object found in an image of size $N \times M$ which is represented by the column vector $\boldsymbol{f} \in \mathbb{R}^{(N \cdot M)}$. This image vector $\boldsymbol{f}$ is transformed into a feature vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_D)^T \in \mathbb{R}^D$ by a linear transformation

$$\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{f} \in \mathbb{R}^{D \times (N \cdot M)} \tag{10}$$

with

$$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_D)^T. \tag{11}$$

The $\boldsymbol{\phi}_i$ correspond to the D largest eigenvectors of the covariance matrix of all the training images $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$.

During the training each training image $\boldsymbol{f}_i$ is overlaid by random noise many times and each resulting picture is projected into the eigenspace. The obtained feature vectors are used to estimate parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ of a normal distribution in the eigenspace. These parameters are stored together with the know class κ_i and pose $\boldsymbol{\theta}$ of the object in image $\boldsymbol{f}_i$ and form a model

$$\boldsymbol{B}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \Omega_{\kappa_i}, \boldsymbol{\theta}_i\} \tag{12}$$

for this view.

A necessary enhancement to enable this statistical eigenspace approach to work with heterogeneous background were the introduction of binary masks for each training image. These masks were used to eliminate the background. We are aware that this method is weak compared to other current object recognition techniques but we focus on the fusion of multiple views in this paper and want to show that fusion can improve recognition results even under difficult conditions.

Object	Number of fusioned images. 1	2	5	10	15	20	25
Coke gray	0	5	40	75	90	100	100
Coke red	0	45	95	95	100	100	100
stapler green	5	10	30	80	100	100	100
stapler white	60	70	85	100	100	100	100
hole punch green	0	10	20	30	30	30	30
hole punch red	40	50	85	100	100	100	100
NaCl-bottle	20	30	65	90	100	100	100
pillbox	75	60	90	100	100	100	100
cup	0	15	10	25	35	30	40
cup with saucer	0	0	0	0	0	0	0

Table 1: Recognition rates (in percent) of the single objects.

The density $p(\boldsymbol{f}_n|\boldsymbol{c}_i^n)$ from Section 2.2 is evaluated by projecting the image into the eigenspace and evaluating the normal distribution $\mathcal{N}(\boldsymbol{\Phi}\boldsymbol{f}_n|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ whose

- class Ω_{κ_m} is the same as in sample $\boldsymbol{c}_i^n$
- pose $\boldsymbol{\theta}_i$ has a minimum distance to the pose of the sample $\boldsymbol{c}_i^n$.

In our experiments we restrict the dimension of our eigenspace from equation 11 to $D = 3$ to complicate the object recognition even more.

3.2 Data Set and Results

Our data set consists of the 10 office and hospital objects shown in Figure 2. The images of the objects were originally taken in front of homogeneous, black background and pasted into separately taken pictures of 314 backgrounds. The training of the classifier described in Section 3.1 was performed with 18600 images. We want to note that this relative large number of images for training is a consequence of our eigenspace-based classifier. If one would use a classifier such as [11] the number of required training images would decrease significantly.

The evaluation of our fusion approach was done with 20 sequences of 25 images each per object which leads to a total number of 5000 test images used in our experiments. The camera movements $\boldsymbol{a}$ were chosen *randomly* from set of pre-taken test-images. In our experiments the variance parameters for the transition noise (see equation (9)) were

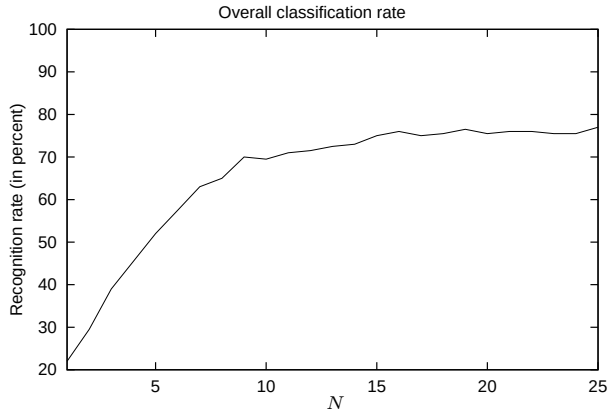


Figure 3: Overall recognition rate. N denotes the number fusioned images.

set to $\sigma_1 = 3.0°$ and $\sigma_2 = 3.0°$. Our sample sets C^n consisted of $K = 25000$ samples (2500 for each of the 10 possible classes).

In Table 1 we show the recognition rates for an increasing number of fusioned images. The results are separated into the single objects. As expected, the quality of classification increases with the number N of fused images, except for the "cup with saucer" that is always misclassified as "cup". This is a property of the used statistical eigenspace classifier and the used background treatment. Smaller objects are always preferred to big objects. This is the reason why the hypotheses for the "cup" is always rated higher by $p(\boldsymbol{f}_n|\boldsymbol{c}_i^n)$ than the one for "cup with saucer". In Figure 3 the overall classification rate is illustrated for all performed fusion steps.

The results of the experiments for the localization accuracy of correct classified images are shown in Figure 2. The accurateness is given with the so called *percentile values*, which describe the limits of the localization error if the classification is correct and only the x% best localizations are taken into account. For example, the percentile value Percentile_{75} expresses the largest localization error within the 75% most accurate localizations. As it can be seen in Table 2, the Percentile_{75} localization error, for example, drops from $81°$ in the first image down to $15°$ after 25 images.

The computation time needed for one fusion step is about 59 seconds on a LINUX PC with an AMD Athlon XP 2000 (1667 MHz) processor. Most of the computational effort is used for the evaluation of the samples which takes about 2.4 milliseconds per sample. Almost all of the time is spend in the handling of the masks for the elimination of the background, the only part of our system that is not optimized for runtime at the moment. In [5] we have shown for a less computational expensive fusion problem that in principle fusion of multiple views is possible for real-time applications. And as the computational effort scales linear to the size of the sample set it is always possible to decrease computing time by decreasing the size of the sample set. Furthermore, as the CONDENSATION algorithm can be parallelized very well this is another way realize time critical applications.

	Number of fusioned images.					
	1	5	10	15	20	25
P50	$39°$	$28°$	$14°$	$14°$	$14°$	$10°$
P75	$81°$	$60°$	$46°$	$47°$	$41°$	$15°$
P90	$93°$	$98°$	$80°$	$80°$	$81°$	$46°$
P95	$105°$	$113°$	$92°$	$101°$	$96°$	$82°$

Table 2: Improvement of localization results. The localization error is given in degree. P50, P75, P90, P95 denote the Percentile_{50}, Percentile_{75}, Percentile_{90} and Percentile_{95} values.

4. Conclusion

In this paper we have presented a general approach for the fusion of multiple views for active object recognition that can help to improve classification rates substantial even under difficult conditions like heterogeneous background and not perfectly suited classifiers.

In the experiments we have shown that our approach is well suited for the fusion of multiple views as we were able to more than triple the overall classification rate to 77%. A passive approach for object recognition would have stopped after the first image and would have obtained a classification rate of only 22%.

The main advantages of our approach are the improvement of the object recognition results and that our CONDENSATION-based fusion scheme is independent of the chosen statistical classifier. Other advantages of our approach are – as we have shown in [5] – its scalability of the size of the sample set and possibility to parallelize the CONDENSATION algorithm.

Presently we use randomly chosen views for our fusion. But we expect that far better classification rates will be reached after fewer views if we combine our fusion approach with our viewpoint selection [4, 3]. The combination of these two approaches for the selection of views and their fusion will result in a system that is still independent of the used classifier and well-suited classifying ambiguous objects.

Open questions in our approach are the minimal necessary size of the sample set – a problem that has been left out in this paper – and the optimal parameters for the noise transition models. Furthermore other sample techniques are to be evaluated.

References

[1] Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association*. Academic Press, Boston, San Diego, New York, 1988.

[2] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Appearance based active object recognition. *Image and Vision Computing*, (18):715–727, 2000.

[3] F. Deinzer, J. Denzler, and H. Niemann. Classifier Independent Viewpoint Selection for 3-D Object Recognition. In G. Sommer, N. Krüger, and C. Perwass, editors, *Mustererkennung 2000*, pages 237–244, Berlin, September 2000. Springer.

[4] F. Deinzer, J. Denzler, and H. Niemann. Viewpoint Selection - A Classifier Independent Learning Approach. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 209–213, Austin, Texas, USA, 2000. IEEE Computer Society, California, Los Alamitos.

[5] F. Deinzer, J. Denzler, and H. Niemann. On Fusion of Multiple Views for Active Object Recognition. In B. Radig and S. Florczyk, editors, *Pattern Recognition — 23rd DAGM Symposium*, pages 239–245, Berlin, September 2001. Springer.

[6] J. Denzler and C. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002.

[7] M. Isard and B. Andrew. CONDENSATION—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[8] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, pages 35–44, 1960.

[9] H. Murase and S. Nayar. Visual Learning and Recognition of 3–D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[10] L. Paletta, M. Prantl, and A. Pinz. Learning temporal context in active object recognition using bayesian analysis. In *International Conference on Pattern Recognition*, volume 3, pages 695–699, Barcelona, 2000.

[11] M. Reinhold, D. Paulus, and H. Niemann. Appearance-Based Statistical Object Recognition by Heterogenous Background and Occlusions. In B. Radig and S. Florczyk, editors, *Pattern Recognition, 23rd DAGM Symposium*, pages 254–261, München, September 2001. Springer-Verlag, Berlin, Heidelberg, New York. Lecture Notes in Computer Science 2191.

[12] B. Schiele and J. Crowley. Transinformation for active object recognition. In *Proceedings of the Sixth International Conference on Computer Vision*, pages 249–254, Bombay, India, 1998.

[13] S. C. Sumantra Dutta Roy and S. Banerjee. Recognizing Large 3-D Objects through Next View Planning using an Uncalibrated Camera. In *IEEE International Conference on Computer Vision*, pages II: 276 – 281, Vancouver, Canada, June 2001. IEEE Computer Press.