On Fusion of Multiple Views for Active Object Recognition

Frank Deinzer*, Joachim Denzler, and Heinrich Niemann

Chair for Pattern Recognition, Department of Computer Science,
Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen
{deinzer,denzler}@informatik.uni-erlangen.de
http://www.mustererkennung.de

Abstract. In the last few years the research in 3–D object recognition has focused more and more on active approaches. In contrast to the passive approaches of the past decades where a decision is based on one image, active techniques use more than one image from different viewpoints for the classification and localization of an object. In this context several tasks have to be solved. First, how to choose the different viewpoint and how to fusion the multiple views.

In this paper we present an approach for the fusion of multiple views within a continuous pose space. We formally define the fusion as a recursive density propagation problem and we show how to use the CONDENSATION algorithm for solving it.

The experimental results show that this approach is well suited for the fusion of multiple views in active object recognition.

Keywords. Active Vision, Sensor Data Fusion

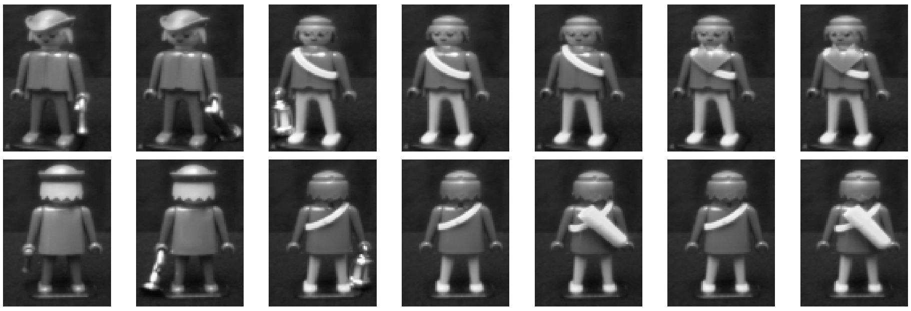
1 Introduction

Active object recognition has been investigated in detail recently [4,8,1,7,2]. The main motivation is that recognition can be improved if the right viewpoint is chosen. First, ambiguities between objects can be avoided that make recognition difficult or impossible at all. Second, one can prevent to present views to the classifier where in the mean worse results are expected. Those views depend on the classifier and can be recognized right after training, when the first tests are performed.

One important aspect in active object recognition — besides the choice of the best viewpoint — is the fusion of the classification and localization results of a sequence of viewpoints. Not only for ambiguous objects, for which more than one view might be necessary to resolve the ambiguity (examples are presented in the experimental sections), the problem arises how to fuse the collected views to finally return a classification and localization result. Also a sequence of views will improve recognition rate in general if a decent fusion scheme is applied. In this paper we present of a fusion scheme based on the CONDENSATION algorithm [5]. The reason for applying the CONDENSATION algorithm is

* 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.

B. Radig and S. Florczyk (Eds.): DAGM 2001, LNCS 2191, pp. 239–245, 2001.



object o1 gun | object o2 trumpet | object o3 lamp | object o4 band/ruff | object o5 quiver/ruff | object o6 band/bib | object o7 quiver/bib

Fig. 1. Examples of the seven toy manakins used for the experiments. Please note that objects o4 to o7 cannot be classified with one view due to the complex ambiguities

threefold: first, inherently one has to deal with multimodal distributions over the class and pose space of the objects. Second, moving the camera from one viewpoint to the next will add uncertainty in the fusion process, since the movement of the camera will always be disturbed by noise. Thus, in the following fusion process of the classification and localization results acquired so far with the results computed from the current image, this uncertainty must be taken into account. 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 let us believe, 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: in viewpoint fusion the dynamics is given by the known but noisy camera motion between two viewpoints.

In the next section we summarize the problem and propose our sensor data fusion scheme based on the Condensation. The performed experiments and an introduction to the classifier used in the experiments are resented in Section 3 to show the practicability of our method. Finally, a conclusion is given in Section 4.

2 Fusion of Multiple Views

In active object recognition object classification and localization of a static object is based on a sequence or series of images. These images shall be 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 in fact a well directed combination of a funded fusion of images and an active viewpoint selection.

This section deals with the principles of the fusion of multiple views. Approaches for active viewpoint selection will be left out in this paper. They have been presented in [3,2].

2.1 Density Propagation with the Condensation Algorithm

Given an object, a series of observed images $\boldsymbol{f}_n, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{f}_0$ and the camera movements $\boldsymbol{a}_{n-1}, \ldots, \boldsymbol{a}_0$ that lead to these images, one wants to draw conclusions from these observation for the non-observable state $\boldsymbol{q}_n$ of the object. This state $\boldsymbol{q}_n$ contains the *discrete* class and the *continuous* pose of the object

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

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

with the normalizing constant

$$k_n = p(\boldsymbol{f}_n, \boldsymbol{a}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0). \tag{2}$$

The density $p(\boldsymbol{q}_n|\boldsymbol{a}_{n-1}, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0)$ can be written as

$$\begin{aligned} &p(\boldsymbol{q}_n|\boldsymbol{a}_{n-1}, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0) = \\ &\int\limits_{\boldsymbol{q}_{n-1}} p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}) p(\boldsymbol{q}_{n-1}|\boldsymbol{a}_{n-1}, \boldsymbol{f}_{n-1}, \ldots, \boldsymbol{a}_0, \boldsymbol{f}_0) d\boldsymbol{q}_{n-1} \end{aligned} \tag{3}$$

with 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. This probability depends only on the camera movement $\boldsymbol{a}_{n-1}$. The inaccuracy of the camera movement is modeled with a normally distributed noise component so that 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. If one deals with *discrete* states $\boldsymbol{q}_n$, the integral in equation (3) simply becomes a sum

$$p(\boldsymbol{q}_n|\boldsymbol{f}_{n-1}, \ldots, \boldsymbol{f}_0) = \sum_{\boldsymbol{q}_{n-1}} p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}) p(\boldsymbol{q}_{n-1}|\boldsymbol{f}_{n-1}, \ldots, \boldsymbol{f}_0) \tag{4}$$

that can easily be evaluated in an analytical way. For example, to classify an object Ω_κ in a sequence of images with $\boldsymbol{q}_n = \left(\Omega_\kappa\right)$, $p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1})$ degrades to

$$p(\boldsymbol{q}_n|\boldsymbol{q}_{n-1}, \boldsymbol{a}_{n-1}) = \begin{cases} 1 & \text{if } \boldsymbol{q}_n = \boldsymbol{q}_{n-1} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

since the object class does not change if the camera is moved, and consequently equation (4) must have an analytically solution.

But we want to use the fusion of multiple view for our viewpoint selection approach [3,2] where we have to deal with localization of objects in continuous pose spaces and consequently states $\boldsymbol{q}_n$ with continuous pose parameters. For that reason it is no longer possible to simplify equation (3) to equation (4).

The classic approach for solving this recursive density propagation is the well-known Kalman Filter [6]. But in computer vision the necessary assumptions for the Kalman

Filter, e.g. $p(\boldsymbol{f}_n|\boldsymbol{q}_n)$ 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) [5]. It uses a 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 Condensation Algorithm for Fusion of Multiple Views

In this section we want to show, how to use the CONDENSATION algorithm for the fusion of multiple views.

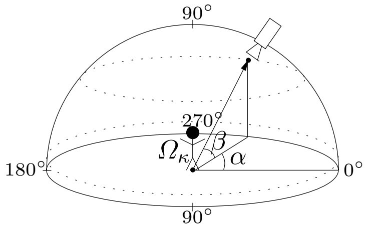


Fig. 2. Experimental setup and the possible pose space

As we want to classify and localize objects, we need to include the class and pose of the object into our state $\boldsymbol{q}_n$. In our experimental setup we move our camera on a hemisphere around the object (see Fig. 2). Consequently, the pose of the object is modeled as the viewing position on a hemisphere (azimuthal and colatitude angles). This leads to the following definitions of the state $\boldsymbol{q}_n = (\Omega_\kappa \ \alpha^n \ \beta^n)^T$ and the samples $\boldsymbol{c}_i^n = (\Omega_\kappa \ \alpha_i^n \ \beta_i^n)^T$ with the class Ω_κ, the azimuthal $\alpha \in [0°; 360°)$ and the colatitude $\beta \in [0°; 90°]$. In Fig. 2 the pose space is illustrated. The camera movements are defined accordingly as $\boldsymbol{a}_n = (\Delta\alpha_n \ \Delta\beta_n)^T$ with $\Delta\alpha_n$ and $\Delta\beta_n$ denoting the relative azimuthal and colatitude change of the viewing position of the camera

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

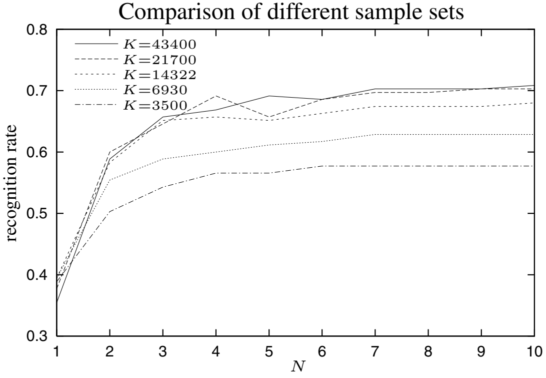
$$\frac{p(\boldsymbol{f}_{n-1}|\boldsymbol{c}_i^{n-1})}{\sum\limits_{j=1}^{K} p(\boldsymbol{f}_{n-1}|\boldsymbol{c}_j^{n-1})} \tag{6}$$

2. propagated with the sample transition model

$$\boldsymbol{c}_i^n = \boldsymbol{c}_i^{n-1} + \begin{pmatrix} 0 \\ r_\alpha \\ r_\beta \end{pmatrix} \quad \text{with} \quad \begin{array}{l} r_\alpha \sim \mathcal{N}(\Delta\alpha_n, \sigma_\alpha) \\ r_\beta \sim \mathcal{N}(\Delta\beta_n, \sigma_\beta) \end{array} \tag{7}$$

Table 1. Recognition rates for different sizes K of the sample set. The transition noise parameters are set to $\sigma_\alpha = 1.8^\circ$ and $\sigma_\beta = 1.5^\circ$. N denotes the number of fusioned images

Object	$K = 43400$			
	$N{=}1$	$N{=}2$	$N{=}5$	$N{=}10$
o1	32%	16%	16%	16%
o2	48%	84%	92%	92%
o3	16%	36%	60%	60%
o4	24%	56%	64%	68%
o5	64%	88%	88%	92%
o6	24%	52%	76%	80%
o7	40%	80%	88%	88%
ϕ	35%	59%	69%	71%



and the variance parameters σ_α and σ_β of the azimuthal and colatitude Gaussian noise $\mathcal{N}(\Delta\alpha_n, \sigma_\alpha)$ and $\mathcal{N}(\Delta\beta_n, \sigma_\beta)$. They model the inaccuracy of the camera movement under the assumption that the error of the azimuthal and colatitude movements of the camera are independent of each other.

3. evaluated in the image by $p(f_n|\boldsymbol{c}_i^n)$.

For a detailed explanation on the theoretical background of the approximation of equation (1) by the sample set C^N cf. [5].

It is important to note that it is absolutely necessary to include the class Ω_κ into the object state $\boldsymbol{q}_n$ (and therewith also into the samples $\boldsymbol{c}_i^n$). An obvious idea that would omit this is to set 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 Experiments

For the experiments presented in this section we have decided for an appearance based classifier using the Eigenspace approach in a statistical variation similar to [1]. As already proposed the Condensation algorithm is independent of the used classifier as long as the classifier is able to evaluate $p(\boldsymbol{f}_n|\boldsymbol{q}_n)$. Our classifiers projects an image $\boldsymbol{f}_n$ into the three-dimensional Eigenspace and evaluates the resulting feature vector for the trained normal distribution with pose parameters that are closest to the given pose. The intention of three-dimensional Eigenspace is to force big importance to the fusion aspect as the chosen low dimensional Eigenspace of course is not suited to produce optimal feature vectors.

Our data set consists of the seven toy manikins shown in Fig. 1. The objects have been selected in a way that they are strongly ambiguous from some viewpoints. The objects o4 to o7 even cannot be classified with one view so that a fusion of multiple views is essential. The evaluation of our fusion approach was done with 25 sequences of

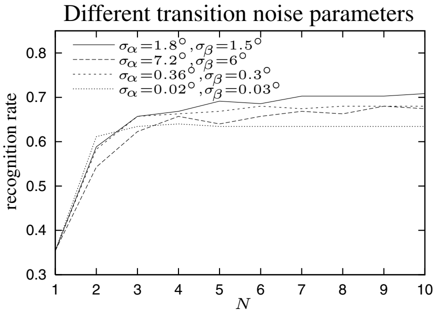


Fig. 3. Recognition rates for different settings of the transition noise parameters σ_α and σ_β. The size of the sample set is K= 43400. N denotes the number of fusioned images

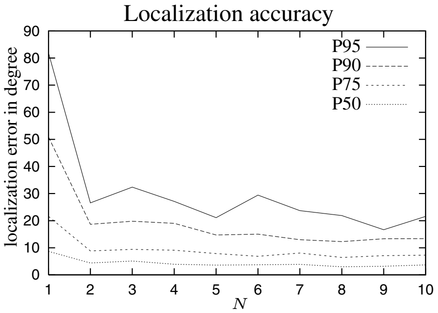


Fig. 4. Accuracy of localization for percentile values of 95% (P95), 90% (P90), 75% (P75), 50% (P50). Size of sample set K= 43400, transition noise parameters σ_α= 1.8°, σ_β= 1.5°

10 images each per object. The camera movements $\boldsymbol{a}$ were chosen *randomly* from the – within the mechanical limits of 0.03° – continuous space of possible movements.

In Table 1 we show the recognition rates for different sizes K of the sample set. As expected, the quality of classification increases with the number N of fused images. It also turns out that the size of the sample set has a noticeable influence on the recognition rates as the approximation of equation (1) is more accurate for larger sample sets.

Another important point we investigated was the influence of the noise parameters σ_α and σ_β from equation (7) on the recognition rate. In Fig. 3 the recognition rates for different transition noise settings are shown. As it can be seen, too much transition noise (large σ_α and σ_β) performs better than insufficient transition noise. The reason for that is that small σ_α and σ_β cause the samples in the sample set to be clustered at a very "narrow" area with the consequence that errors in the camera movement and localization are not sufficiently compensated. In contrast, too much noise spreads the samples too far.

The results of the experiments for the localization accuracy are shown in Fig. 4. 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 P90 expresses the largest localization error within the 90% most accurate localizations. As it can be seen in Fig. 4, the P90 localization error drops from 50° in the first image down to 13° after ten images.

The computation time needed for one fusion step is about 1.8 seconds on a LINUX PC (AMD Athlon 1GHz) for the sample set with $K = 43400$ samples. As the computational effort scales linear to the size of the sample set, we are able to fuse 7 images per second for the small sample set with $K = 3500$ samples which already provides very reasonable classification rates. We also want to note that the CONDENSATION algorithm can be parallelized very well so that even real-time applications can be realized using our approach.

4 Conclusion

In this paper we have presented a general approach for the fusion of multiple views for active object recognition. Using the Condensation algorithm we are independent of the chosen statistical classifier. Other advantages of our approach are its scalability of the size of the sample set and possibility to parallelize the Condensation algorithm. In the experiments we have shown that our approach is well suited for the fusion of multiple views as we were able to double the overall classification rate from 35% to 71% and increase of the classification rate for single objects of up to 233%.

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 [3,2]. 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 for the given task of classifying ambiguous objects.

Open questions in our approach are the minimal necessary size of the sample set and the optimal parameters for the noise transition models. Furthermore other sample techniques are to be evaluated.

References

1. H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Appearance based active object recognition. *Image and Vision Computing*, (18):715–727, 2000.
2. F. Deinzer, J. Denzler, and H. Niemann. Classifier Independent Viewpoint Selection for 3-D Object Recognition. In G. Sommer, N. Kr"uger, and Ch Perwass, editors, *Mustererkennung 2000*, pages 237–244, Berlin, September 2000. Springer.
3. 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.
4. J. Denzler and C.M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (in press)*, 2001.
5. M. Isard and B. Andrew. CONDENSATION—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
6. R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, pages 35–44, 1960.
7. 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.
8. B. Schiele and J.L. Crowley. Transinformation for active object recognition. In *Proceedings of the Sixth International Conference on Computer Vision*, Bombay, India, 1998.