A Framework for Cooperative Object Recognition

N. Oswald and P. Levi
Institute of Parallel and Distributed High-Performance Systems,
Applied Computer Science - Image Understanding
70565 Stuttgart
Germany

Abstract

This paper explores the problem of object recognition from multiple observers. The basic idea is to overcome the limitations of the recognition module by integrating information from multiple sources. Each observer is capable of performing appearance-based object recognition, and through knowledge of their relative positions and orientations, the observerrs can coordinate their hypotheses to make object recognition more robust.

A framework is proposed for appearance-based object recognition using Canny edge maps that are effectively normalized to be translation and scale invariant. Object matching is formulated as a non-parametric statistical similarity computation between two distribution functions, while information integration is performed in a Bayesian belief net framework. Such nets enable both a continuous and a cooperative consideration of recognition result. Experiments which are reported on two observers recognizing mobile robots show a significant improvent of the recognition results.

1 Introduction

The fast and robust recognition of objects is a central task in robotic applications in traffic, manufacturing or services. In such applications, various objects with a priori unknown identity, position and orientation are seen with changing background, projective distortions, and varying illumination conditions from observers in arbitrary view positions. Thus, the design of a recognition system requires the avoidance of misinterpretations caused by the just mentioned difficulties. Consequently, the presumption seems likely that an observation with distributed visual input from several cooperative observers will increase robustness in such applications.

Information integration is a kind of multi-sensor fusion with homogenouos data. In applications concerning multi-sensor fusion this process can typically be assigned to different levels of abstraction [8] [11]. Unlike in [2], where cooperation is distinguished between the levels of sensing, processing, manipulation, behaviors, and agents, we consider different aims of cooperation and its prerequisites, and develop a relation to the levels of abstraction. At low-level processing, the aim of cooperation is to take advantage of available required sensory information in a team of observers. There do not exist any prerequisites for a single observer. Cooperation leads obviously to an increasement of knowledge of individual observers if they are not able to look at the required region. Up to now, most of the literature concerning cooperation can be assigned to the intermediate-level. The aim of cooperation at this level is the composition of object components provided that a shared and overlapping field of view between combining views exists. [14] describes a method to generate a 3D reconstruction of an unknown static scene by combining stereo and focus data. In [7] the 3D reconstruction is done in a dynamic manner by fusing data gained from stereo and optical flow computations. Another approach [1] builds a formal object description from vision and touch that is used as a base for the recognition task. Recent approaches [12] [13] [4] that deal with multiple perspective interactive videos try to enable a 3D scene analysis with the background of influencing the individual perspective view of an observer. Cooperation at high-level processing assumes an object that is observed from different view positions and deals with the integration of object hypotheses. Such integration or fusion of data at a higher level is supported by few agent architectures [8]. Cooperation aims at this level are of great interest in multi-agent applications like in [17] in order to build observers that operate autonomously by calculating their own estimations. Nevertheless, research with regard to this level can hardly be found.

The recognition scenario (Fig. 1) discussed in this article concentrates at high-level processing on the cooperative recognition with several observers. In such a shared environment each observer calculates recognition results of a common target object. The object

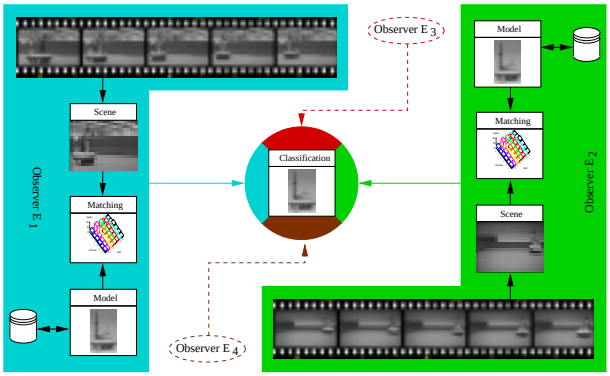

Figure 1: Scenario of cooperative recognition

recognition receives relevant data in the form of areas of motion calculated by a tracking algorithm[1] similar to the one proposed by [22]. The correspondence between object and model data is done by a new statistical approach that supplies a list of hypotheses. The hypotheses of each observer were finally combined by the use of bayesian belief nets. Experimental results will indicate the influence of cooperation on the quality of hypotheses.

2 Statistical object recognition

There are mainly two approaches of object modelling, the appearance-based and the geometric one[9] [19]. The appearance-based approach has the advantage of not requiring a formal description. Thus, any kind of object can be modelled, but, in general, object hierarchies do not exist. On the other hand, hierarchies are the advantage of the geometric object modelling model. But so far, this approach is limited to simple objects [15]. Thus, the approach to choose depends on the complexity of the objects. Because of the robots in Fig. 1 being complex the appearance-based approach was chosen.

Aside from most isolated applications, the recognition process in a natural environment has to cope with a number of difficulties. A target object can be taken from arbitrary viewing positions with different illumination conditions and varying background. Current approaches suggest only solutions to parts of these requirements. In [16] an eigen-value method is proposed to identify various objects and their orientation, but this method requires a constant background. Based on that [3] extended this approach to cope with partly occluded objects but have not shown yet results in a real scenario like the one of Fig. 1. [6] presents a recognition method using quasi-invariants based on segments.

[1] A detailed description is beyond the scope of this article

To analyse complex objects, a lot of effort has to be investigated in the generation of segments. A lambertian model is proposed by [5] to recognize more complex objects. But this method assumes fixed locations and does not permit background information.

2.1 Formal object descriptions

Starting out of an intensity image I - received automatically from the tracking algorithm - that contains mainly the target object T iconic features like gray values, edges, or corners can be determined to build a formal object description. Features are put on geometric attributes indicating their location in image I. So far we found edges generated by the canny filter as useful features.

Each object is formally described by a set M of n_M features.

$$M = \{M_1, M_2, \ldots, M_{n_M}\} \tag{1}$$

The geometric position of each feature M_i in I is determined by its edge coordinates $\vec{c_i}$ in x and y direction with $\vec{c_i} = (c_{x,i}, c_{y,i})$. Thus, a feature is defined as a tupel M_i, consisting of its edge coordinates $\vec{c_i}$ and optional a set of further attributes D_i to specify the feature additionally

$$M_i = \{\vec{c_i}, D_i\} \qquad i = 1, \ldots, n_M. \tag{2}$$

Features are standardized by transformation from I into a scale and translation invariant space, the configuration space κ. For this geometric transformation the location of a feature M_i is set in connection with all remaining feature locations M_j with $j \neq i$ to get the relative location of M_i. Because I is 2-dimensional, 2^2 possible transformations into κ exist. With a criterion f, two arbitrary features M_i and M_j are marked with the $<$ relation along x or y as follows

$$f_{\{x,y\},ij} = \begin{cases} 1 & c_{\{x,y\},i} < c_{\{x,y\},j} - \varepsilon_{\{x,y\}} \ \wedge \ i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$\varepsilon_{\{x,y\}}$ enables an overlapping tolerance. With $f_{\{x,y\},ij}$ the relative location of a feature M_i in relation to the remaining features out of $M \setminus \{M_i\}$ along x and y is given with

$$b_{\{x,y\},i} = \frac{1}{n_M - 1} \sum_{j=1}^{n_M} f_{\{x,y\},ij} \qquad n_M > 1 \tag{4}$$

For each feature M_i the relative location builds a two-dimensional destination vector $\vec{b_i}$ valid in the range $[0, 1]$ with $\vec{b_i} = (b_{x,i}, b_{y,i})^T$. Thus the configuration space κ is a normalized continuous space in the same

range and with the same dimension as I. The transformed features from I to κ build the configuration space representation of an object $(\vec{b_1}, \ldots, \vec{b}_{n_M})$.

In Fig. 2 a sample transformation from a reference object R and a target object T into κ using the canny operator is shown. As can easily be verified, the emerged configuration space is translation and scaling invariant. Because of considering only relative locations, few disturbances or projective distorsion appear compensated in κ.

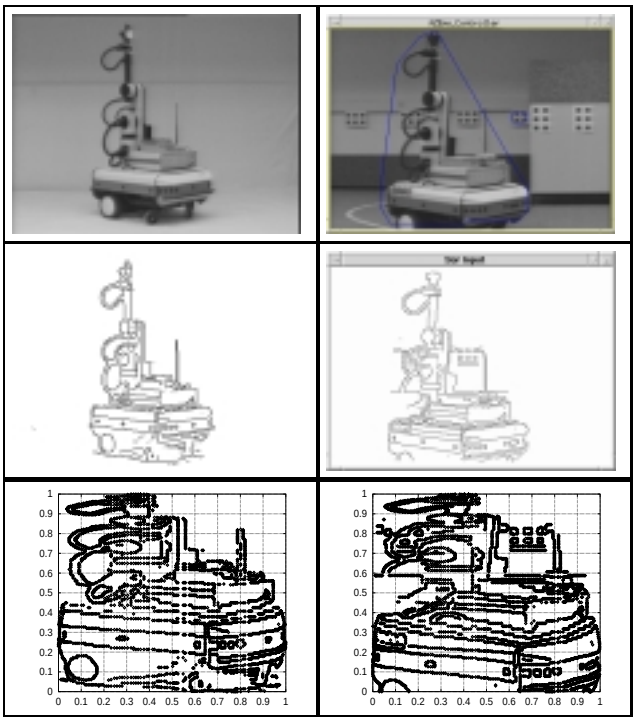

Figure 2: Transformation of reference (left) and target object (right) via canny edge images into configuration spaces κ^R and κ^T

2.2 Generation of hypotheses

Because n_M of our objects is very high we consider all $\vec{b_i}$ in κ as randomly located in the statistical sense. However, except for classification tasks, statistical methods for recognition are not widely used yet. [25] provides a recognition method based on the linear combination of normal distributed density functions from point features. In [10] a formal statistical object description based on parametric methods is introduced to identify and localize objects. Differently, we compare distribution functions of a reference object F^R with a target object F^T by the use of a non-parametric statistical test in order to have no constraints to the distribution of the point features. This test calculates a similarity measure between the two distribution functions. A steady and discrete distribution function F with $\vec{z} \in [0,1]^2$ is built from the destination vectors in κ with

$$F(\vec{z}) = \begin{cases} 0 & \vec{z} < \vec{b}_1 \\ \vec{b_i} & \vec{b}_i \leq \vec{z} < \vec{b}_{i+1} \\ 1 & \vec{z} \geq \vec{b}_{n_M} \end{cases} \quad (5)$$

In a number of experiments we found the test of Kuiper useful [21]. This test is a variant of the well-known test of Kolmogoroff and Smirnow. In contrast to latter, the maximum distance between the two distribution functions is calculated in both directions. The Kuiper statistic D between the distributions of a target object $F^T(x)$ and a reference object $F^R(x)$ is calculated by

$$D = \max_{\vec{z}}\{F^T(\vec{z}) - F^R(\vec{z})\} + \max_{\vec{z}}\{F^R(\vec{z}) - F^T(\vec{z})\} \quad (6)$$

For a sample set of size n_M, the significance is calculated by $Q(\lambda)$

$$Q(\lambda) = \begin{cases} \sum_{j=-\infty}^{\infty} (1 - 4j^2\lambda^2) e^{-2j^2\lambda^2} & \text{for} \quad \lambda > 0 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

In [21] λ is approximated depending on the sample data n_M and the distance D by

$$\lambda = \left(\sqrt{m} + 0.155 + \frac{0.24}{\sqrt{m}}\right) D \quad (8)$$

The number of samples m is calculated from the number of features of target ${n_M}^T$ and reference object ${n_M}^R$ with

$$m = \frac{{n_M}^T {n_M}^R}{{n_M}^T + {n_M}^R} \quad (9)$$

The more any two distribution functions resemble, the smaller gets $Q(\lambda)$. Thus, $h = 1 - Q(\lambda)$ is used as hypothesis of the match.

According to the appearance-based approach, each reference object R_i is modelled by a number of views n_A. Thus, for a statistical match between a target object T and a reference object R_i, the configuration space κ^T has to be compared with all configuration spaces κ^{R_i}. The comparison of T with R_i supplies a hypotheses vector $\vec{h}_{R_i}$ with $\vec{h}_{R_i} = (h_{R_1}, \ldots, h_{R_{n_A}})^T$ containing the calculated hypotheses of all matches between views of R_i and T. The comparison of T with n_R reference objects leads to the resulting hypotheses set H_i with $H_i = \{\vec{h}_{R_1}, \ldots, \vec{h}_{R_{n_R}}\}$.

The results of this method are very robust compared to an ideal scenario with varying distances as well as few variations in illumination, orientation, and identical background like in [16]. However, in real scenes with arbitrary viewing positions, variable background, and changing illumination conditions the quality of the segmentation depends on the tracking algorithm. Results are sometimes not exact, data are either not relevant or missing. This may lead to misinterpretations. To overcome this problem it is useful to consider recognition results over a time period.

3 Information integration

Various methods in the field of information integration were presented and compared e.g. in [24] [23]. In contrast to the traditional use of information integration, the scenario deals with changing object positions and orientations as well as with insufficient quality of hypotheses and thus requires dynamic assignments. For these reasons we found the method of bayesian belief nets [18] most useful because belief nets are rather robust in applications with imprecisions [20] [23] and react rather fast to changes [24].

In our application belief nets are used to verify recognition result H with $H = (H_1, \ldots, H_{n_O})$ of up to n_O observers. In the case of one observer the belief net estimates the current orientation of a target object T in the scene based on H_i. In the case of several observers, hypotheses of n_O involved participants are combined in the belief net to estimate the objects identity and orientation.

A belief net consists of a number of single belief nodes that represent the state of the net at a time step t. The design of a single belief node Z is shown in Fig. 3. Each node Z is connected with a predecessor node A and a successor node B that represent states at time steps $t-1$ and $t+1$ respectively. To determine the orientation of a moving object only top-down propagation is essential. In the single node Z the causal support $\pi(Z_i)$ is calculated from the propagated values $\pi_Z(A_j)$ of A and the conditioned probabilities $P(Z_i|A_j)$ with

$$\pi(Z_i) = \sum_{j=1}^{n_A n_R} P(Z_i|A_j)\pi_Z(A_j) \qquad (10)$$

To get conditioned probabilites $P(Z_i|A_j)$, we use $n_A n_R \times n_A n_R$ aspect transition matrices (ATM) which can be different for each belief node:

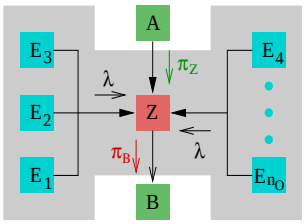

Figure 3: Single belief node Z

$t_{k-1} \backslash t_k$	$R_{1,1}$	...	R_{1,n_A}	...	$R_{n_R,1}$	...	R_{n_R,n_A}
$R_{1,1}$	$p_{1,1}$	...	p_{1,n_A}	...	0	...	0
...	...	...	...	...	...	...	...
R_{1,n_A}	$p_{n_A,1}$	...	p_{n_A,n_A}	...	0	...	0
...	...	...	...	...	...	...	...
$R_{n_R,1}$	0	...	0	...	$p_{1,1}$	...	p_{1,n_A}
...	...	...	...	...	...	...	...
R_{n_R,n_A}	0	...	0	...	$p_{n_A,1}$	...	p_{n_A,n_A}

These matrices describe a priori suspected dependencies of n_A aspects and n_R models between two time steps. Dependencies are usually only sensible between aspects of the same model. All other dependencies are set to zero. Each cell in ATM indicates the probability $p_{u,v}$, that aspect u of model R_i (written as $R_{i,u}$) at time t_{k-1} will change to aspect v of model R_i at t_k $(R_{i,v})$. All dependencies or assumptions can basically be described randomly. Assumptions are e.g. gaussian or uniform distributions applied to a limited range because ususally the observed object will move in a continouos manner. Distributions then assess the transition probability $p_{u,v}$ between selected or all aspects of R_i. All other transition probabilities within one model R_i are marked with 0. As a constraint, each row in ATM has to sum to 1.

At each state and thus at each time step diagnostic support from up to n_O observers can be processed. This is modelled in Fig. 3 by nodes E_i that supply the diagnostic support $\lambda(E_i)$ which contains H_i. The collection of all observer's hypotheses is done by

$$\lambda(Z_i) = \prod_l \lambda(E_l) \qquad (11)$$

The belief values for each belief node are calculated from causal and diagnostic support normalized by α

$$BEL(Z) = \alpha\lambda(Z_i)\pi(Z_i) \qquad (12)$$

From the current belief values the top-down propagation values $\pi_B(Z)$ are calculated by

$$\pi_B(Z) = BEL(Z) \qquad (13)$$

Once a belief node reaches the value zero for one aspect, this value is propagated to all future belief nodes. Such a filtering effect is desired, if a really false aspect is considered. Unluckily, in cases with insufficient input, the object recognition will calculate low or zero probability values for correct aspects. As a consequence, belief values become zero too. Such a behavior is unwanted because it is not robust. To avoid misinterpretations, we extend equation (10) by a factor τ to

$$\pi(Z_i) = \sum_{j=1}^{n_A n_R} P(Z_i|A_j)(\pi_Z(A_j) + \tau_j) \qquad (14)$$

to smooth the causal support $\pi(Z_i)$. To keep the filtering effect, an aspect marking matrix similar to ATM is used as a criterion whether aspects or models are still valid or not.

4 Experimental results

In order to examine the consequences of cooperation on the quality of recognition results we will analyze a sampple scenario in detail with two observers looking at the same moving robot "Aramis". Fig. 5 shows part of that image sequence containing 74 images of size 384×288 with the target object observed from the two viewing angles of E_1 and E_2. T is marked by convex hulls automatically generated by our tracking algorithm and serves as input I for the recognition. The orientation angle between E_1 and E_2 is approximately 190°. To recognize T, both observers used a model database with five reference objects each consisting of 24 aspects taken from fixed distances in 15° steps from 0° to 360°. The recognition time measured on a SUN Enterprise amounts 220ms.

In this experiment, the cooperative recognition enables processing of two tasks, the verification and the localization (Fig. 4). The aim of the verification is to improve a single observer's recognition results by combining H_1 and H_2. To do so, known positions of observers are assumed and hypotheses are transformed to a common coordinate system, usually either to the one used by E_1 or E_2. In the localization, the suspected reference object is used to locate the relative spatial orientation between E_1 and E_2. This has to be done by using the difference of the recognition results of both observers. Obviously, the latter task requires precise values H_1 and H_2.

To examine the consequences of information integration, cooperative results are compared with single and continuous results of both observers. Fig. 6 (a) shows for each of the 74 images both, the assignment of T to a reference object and the corresponding estimated view in degree. For each t_i only the maximum

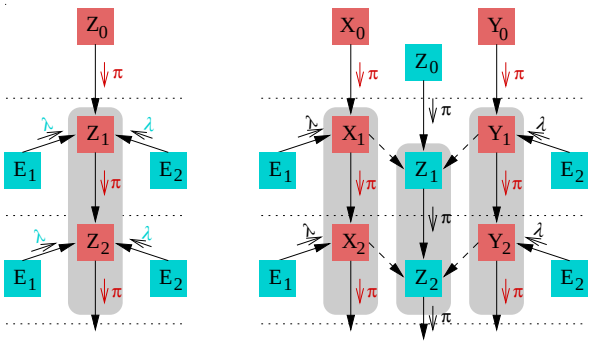

Figure 4: Belief nets for verification task (left) and localization task (right)

hypotheses of E_1 and E_2 are considered. The recognition results of the Kuiper test are quite different for E_1 and E_2. While H_1 contains 24% of misinterpretations, H_2 has 44%. The main reason for this lies in the quality of I depending on the given convex hull. For comparison the results of the continuous considerations of T by E_1 and E_2 are shown in Fig. 6 (b) and (c). The aspect transition matrices are filled according to a uniform distribution in the interval $[i-1, i+1]$ around aspect i and τ was set to 0.002. In (b) the respective estimated reference object is shown and in (c) the corresponding orientation. Note that in (c) only neighbouring hypotheses of the same reference object are connected by a line. As e.g. can be seen in (b) the last four hypotheses of E_1 indicate a wrong model estimation, where T is moving out of E_1's field of view. Although the continuous results are significantly improved, still not all best matches are correct. Observer E_1 calculates hypotheses with approximately 20% misinterpretations, either the wrong view or the wrong model. Observer E_2 could improve its recognition rate to about 29% misinterpretations.

Now these results are compared with cooperative results. The verification is done according to the left picture in Fig. 4. Values of both observers E_1 and E_2 enter the belief net simultaneously at t. Hypotheses H_2 are transformed to the camera coordinate system of observer E_1. The result of the verification is indicated in Fig. 6 (d) by the square symbols. As can be seen, the integration of information from the two different sources E_1 and E_2 lead to an improvement of the estimated current orientation of T with no significant misinterpretations. Due to chosen statistical recognition method, variations of 15° from the real aspect are in the range of tolerance. The time for cal-

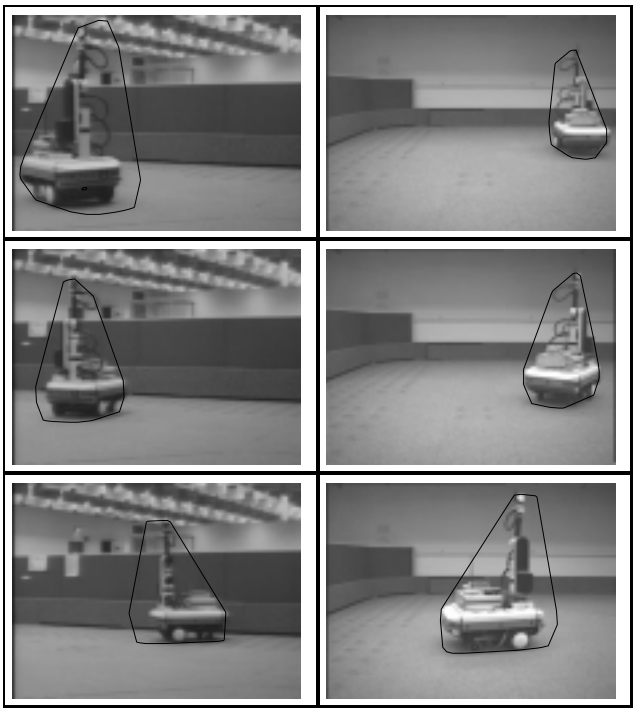

Figure 5: Images #1, #30, and #60 with convex hulls of T observed from the view points of E_1 and E_2

culations in this belief net amounts 20ms on a SUN Enterprise. Thus the total time for the verification task (without time for communication and tracking) is approx. 260ms.

The localization is done according to the right picture in Fig. 4. To get robust hypotheses, each observer verifies its hypotheses by using belief net X respectively Y. To ensure, that E_1 and E_2 suspect the same reference object, the identity of T is estimated by using the maximum belief values for each reference object. These values of $BEL(X)$ and $BEL(Y)$ enter a third belief net Z that estimates the identity of T. With $\tau = 0$ and an identity function for the 5×5 matrix ATM, where aspects are replaced by models, "Aramis" remains the only candidate reference object after $t = 3$ images. To determine the orientation angle between E_1 and E_2 the three highest marked views in H_1 and H_2 are used. Such a procedure is necessary in appearance-based approach due to ambiguous interpretations. The result of the localization is indicated in Fig. 6 (d) by the plus symbols. As long as T is completely visible from E_1, the estimated orientation angle between E_2 and E_1 stays steady between 165° and 195°. After the 70^{th} image, where T run out of I, estimations declines.

5 Conclusions

We introduced a framework for cooperative recognition of objects in a multi-robot environment and showed that the use of integrated information from multiple sources will increase the robustness of the recognition process. To identify complex objects in a real scenario an appearance-based qualitative non-parametric statistical matching algorithm was developed. Based on that, belief nets were proposed to integrate information from both, a single or multiple observers. Cooperation can be used not only for verification but also for the localization of spacial orientations of several observers. Time considerations showed that the whole recognition process is rather fast even with a higher number of models.

References

[1] P.K. Allen, Robotic Object Recognition Using Vision and Touch, Kluwer Acad. Publishers, 1987.

[2] R. Bajcsy, From active perception to active cooperation - fundamental processes of intelligent behavior, GRASP Technical Report 398, 1995.

[3] H. Bischof, A. Leonardis, Robust recognition of scaled eigenimages through a hierarchical approach, Proc. of CVPR98, pp. 664-670, 1998.

[4] Q. Cai and J.K. Aggarwal, Tracking Human Motion Using Multiple Cameras, 13*th* ICPR, pp. C:68-72, 1996.

[5] R. Epstein et. al, Learning Object Representations from Lighting Variations, *Object Representation in Computer Vision II*, pp. 179-199, Springer, 1996.

[6] P. Gros, Using quasi-invariants for automatic model building and object recognition: an overview, *Object Representation in Computer Vision*, pp. 65-75, Springer, 1994.

[7] E. Grosso et. al, 3D Object Reconstruction Using Stereo and Motion, IEEE Trans. Systems, Man and Cybernetics 19(6), pp. 1465-1488, 1989.

[8] D. Hall, *Mathematical Techniques in Multi-sensor Fusion*, Artech House, 1992.

[9] M. Hebert et. al, *Object Representation in Computer Vision*, Springer, 1994.

[10] J. Hornegger, H. Niemann, Statistical Learning, Localization, and Identification of Objects, 5*th* ICCV, pp. 914-919, 1995.

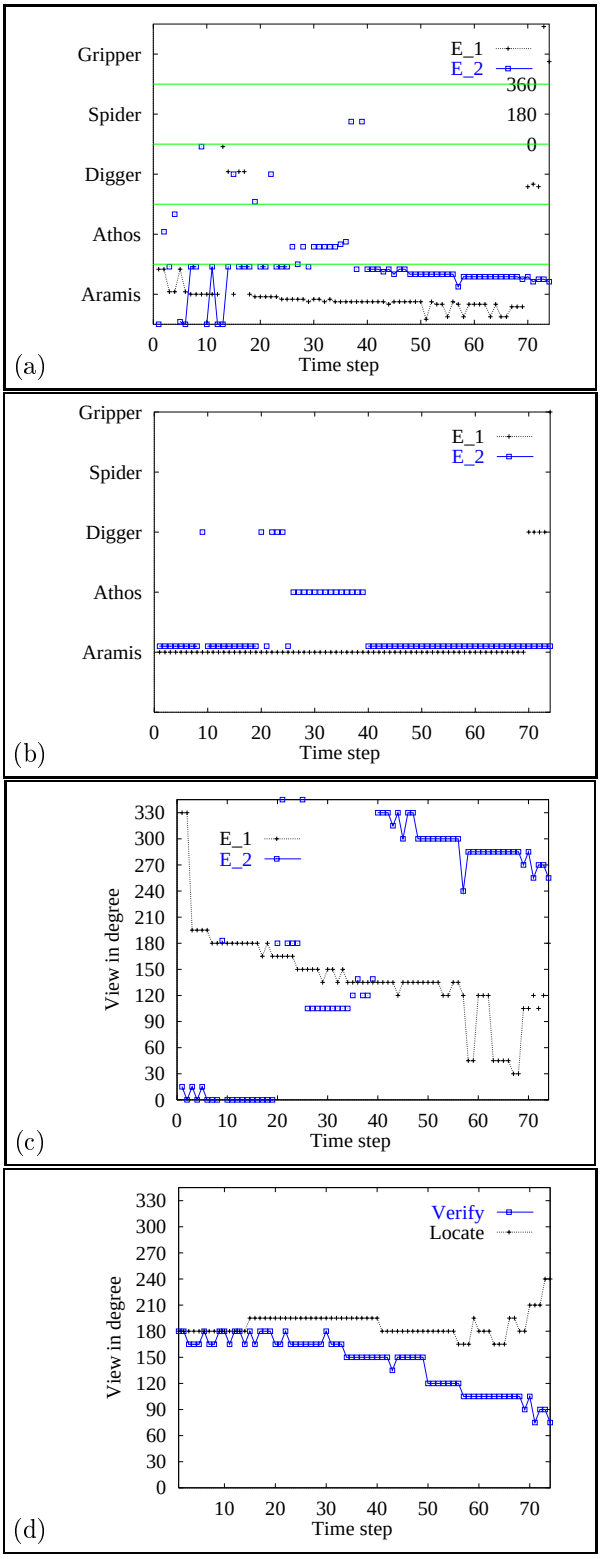

Figure 6: (a) Hypotheses resulting from test of Kuiper for E_1 and E_2 (b+c) model and aspect hypotheses resulting from belief nets for E_1 and E_2 (d) results of verification and localization

[11] S.S. Iyengar, L. Prasad, H. Min, *Advances in Distributed Sensor Integration*, Prentice Hall, 1995.

[12] R. Jain, K. Wakimoto, Multiple Perspective Interactive Video, Int. Conf. Multi-media Computing and Systems, pp. 202-211, 1995.

[13] A. Katkere, R. Jain, A Framework for Information Assimilation, *Exploratory Vision: The Active Eye*, pp. 241-256, Springer, 1996.

[14] E.P. Krotkov, *Active Computer Vision by Cooperative Focus and Stereo*, Springer, 1989.

[15] J. Mundy, An Experimental Comparison of Appearance and Geometric Model Based Recognition, *Object Representation in Computer Vision II*, pp. 247-269, Springer, 1996.

[16] H. Murase, S. Nayar, Visual Learning and Recognition of 3D Objects from Appearance, Int. Jor. of Computer Vision (14), pp. 5-24, 1995.

[17] N. Oswald, P. Levi, Cooperative Vision in a Multi-Agent Architecture, LNCS 1310, 709-716, Springer, 1997.

[18] J. Pearl, *Distributed Revision of Composite Beliefs*, Artificial Intelligence 33, pp. 173-215, 1987.

[19] J. Ponce, A. Zisserman, M. Hebert, *Object Representation in Computer Vision II*, Springer, 1996.

[20] M. Pradhan et. al, The sensitivity of belief networks to imprecise probabilities: an experimental investigation, Artificial Intelligence 85, pp. 363-397, 1996.

[21] W.H. Press, Numerical Recipes in C, Cambridge University Press, 1992.

[22] S.M. Smith, ASSET-2: Real-time motion segmentation and object tracking, Journal of Real Time Imaging, 4(1):21–40, 1998.

[23] P. Walley, Measures of uncertainty in expert systems, Artificial Intelligence 83, pp. 1-58, 1996.

[24] E. Waltz, J. Llinas, *Multisensor Data Fusion*, Artech House, 1990.

[25] W.M. Wells III, *Statistical Object Recognition*, PhD thesis, MIT, 1993.