From contextual knowledge to computational constraints

Shaogang Gong [*]

Dept. of Comp. Sci., QMW, Uni. of London
Mile End Road, London E1 4NS

Hilary Buxton [*]

COGS, Uni. of Sussex
Falmer, Brighton BN1 9QH

Abstract

In this work we address the issue of focused computation in computer vision for effectiveness and efficiency. In particular, we propose a scheme that links scene-oriented contextual knowledge with the computational constraints required in visual motion segmentation and tracking. The approach uses Bayesian belief revision techniques to map explicit scene knowledge onto implicit causal dependent constraints in controlling computational parameters. We discuss our experimental results from applying this method in improving existing techniques in traffic surveillance applications.

1 Introduction

In the past, research in computer vision was greatly influenced by the theory of David Marr [12]. Visual processing modules in the Marr framework operate at different levels of abstraction. Typically, then, a high level visual task, such as recognising and tracking a vehicle, is performed by an assembly of self-contained modules. However, these modules impose very little processing constraint on their predecessors and successors, and their performance is usually judged in isolation by a set of "optimal" criteria. Although this approach to computer vision has developed sophisticated algorithmic procedures for individual visual competences, it is a clumsy approach to build integrated vision systems. The computational complexity required by individual competences leads to ineffective and inefficient performance for the high level tasks of a system.

In recent years, Ullman [19] has argued for the importance of integration amongst visual modules. More specifically, Ballard [3] has suggested an *animate vision* approach for two main reasons: first, vision is better understood in the context of the visual behaviours in which the vision system is engaged and these behaviours often do not require elaborated representations of the three dimensional world; second, it is important for "vision as behaviour" to have a system framework that integrates visual processing within the task context. These arguments are supported by early work in Bajcsy and Allen's concept of *active vision* [2], and more recently, by Brooks' integrated architectures for task-oriented behaviours in robotics [5, 6]. Many researchers have shown the potential for building vision systems with integrated *purposive* frameworks [17, 18, 1, 7, 15, 9].

Our approach claims that *perception is really an opinion on the state of affairs in the world rather than a passive response to sensory stimuli.* In this work, in order to "put vision into context", we emphasise the importance of focused vision and address the issue of controlling the focus by mapping explicit contextual

[*] This work was carried out under the funding of ESPRIT EP2152 (VIEWS) project.

BMVC 1993 doi:10.5244/C.7.23

knowledge to implicit computational constraints in an architectural framework which *dynamically determines the way that the visual modules function.* Then, the notion of vision as behaviour indicates that accumulated knowledge about the past and reasoned prediction about the future should dictate the very basis of any process in order to effectively overcome ill-conditioned computation [1]. Visual knowledge may appear in conceptual and symbolic descriptions, but often it is computationally attractive and feasible to associate explicit knowledge with appropriate implicit numerical measures that give rise to the emergent behaviour [4].

Our initial studies have indicated that contextual knowledge in visual behaviour can be reassembled by an appropriately linked network of chosen parameter sets [10]. The functionality of such a network is a continuous process of initiating visual modules with chosen parameter values and updating such values with new evidence. If visual behaviour is regarded as a process of providing a coherent, *most probable explanation* of all the evidence at hand, all modules involved can then be regarded *individually* as units for, on the one hand, providing its predecessors with updated evidence based on the input from its successors and, on the other hand, invoking chosen parameter values locally. The issue of mapping knowledge to computational constraints resides in: *(1) how explicit contextual knowledge can be represented as distributed implicit parameter sets, and (2) what computational mechanisms are required for effective distribution and invocation of the parameters.* Early studies by Levitt *et al* [11] suggested the use of Bayesian networks for knowledge representation. Recent work by Murino *et al* [13] further exploited such techniques for using knowledge in the control of camera operations.

In section 2, we briefly review some basics of the Bayesian nets and associated belief revision mechanisms before presenting, in section 3, a specific scheme in which scene-oriented contextual knowledge is mapped onto Bayesian nets for the control of a selective and focused segmentation and tracking of moving objects. In section 4, we discuss our experimental results and evaluate our approach against an existing technique. We conclude this work in section 5.

2 Beliefs and the Most-Probable-Explanation

Bayesian belief networks are Directed Acyclic Graphs (DAG) in which each node represents an uncertain quantity using variables with multi-possible values. The arcs connecting the nodes signify the direct causal influences between the linked variables with the strengths of such influences quantified by associated conditional probabilities. If we assume a variable in the network is X_i, and a selection of variables Π_{X_i} are the direct causes of X_i, the strengths of these direct influences are quantified by assigning the variable X_i a link matrix $P(x_i|\Pi_{X_i})$, given any combination of instantiations of the parent set Π_{X_i}. The conjunction of all the local link matrices of variables X_i in the network (for $1 \leq i \leq n$ where n is the total number of the variables) specifies a complete and consistent global model which provides answers to all the probabilistic queries. Such a conjunction is given by the overall joint distribution function over the variables $X_1, ...X_n$: $P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|\Pi_{X_i})$, where lower case symbols stand for a particular instantiation of the corresponding variables [2]. Then, if the behaviour of a visual process is

[1] By saying "overcome ill-conditioned" here we mean, that in a broad sense, employing high-level symbolic knowledge is equivalent to the use of global geometric or algebraic constraints in order to regularise the computation.

[2] In the rest of this article, variables will always be denoted by upper cases and specific instantiations of the variables will be denoted by lower cases.

partially defined by its processing parameters, the evaluation of these parameters shall be consistent with the visual task at hand such that the task is accomplished by associating the parameters with given beliefs.

In a belief network, if we quantify the degree of coherence between the expectations ($\mathbf{X}$) and the evidence ($\mathbf{e}$) by a measure of local belief [3] $BEL(x) = P(x|\mathbf{e})$, and define belief commitments as the tentative acceptance of a subset of hypotheses that together constitute a most satisfactory explanation of the evidence at hand, then, Bayesian belief revision amounts to the updating of belief commitments by distributed local message passing operations. Instead of associating a belief measure with each individual hypothesis locally, belief revision identifies a composite set of hypotheses that best explains the evidence. We call such a set the Most-Probable-Explanation (MPE). In computational terms, this means finding the most probable instantiations of all hypothetical variables given the observation. Let $\mathbf{W}$ stands for all the variables concerned, inclusive of those in $\mathbf{e}$, any particular instantiation of variables in $\mathbf{W}$ that is also consistent with $\mathbf{e}$ will be regarded as an *extension* or *explanation* of $\mathbf{e}$. The problem then is to find an extension $\mathbf{w}^*$ that maximises the conditional probability $P(\mathbf{w}|\mathbf{e})$. In other words, $\mathbf{W} = \mathbf{w}^*$ is the MPE of the evidence if $P(\mathbf{w}^*|\mathbf{e}) = \max_{\mathbf{W}} P(\mathbf{w}|\mathbf{e})$. Here, $\mathbf{w}^*$ is obtained by first locally computing the belief function for each variable X mentioned above, i.e. [4] $BEL^*(x) = \max_{\mathbf{W'}_X} P(x, \mathbf{w'}_X|\mathbf{e})$ where $\mathbf{W'}_X = \mathbf{W} - X$ and second, propagating local messages. The local messages are defined as: if X has n parents $U_1, U_2, ..., U_n$ and m children $Y_1, Y_2, ..., Y_m$, then node X receives messages $\pi^*_X(u_i), i = 1, ..., n$ from its parents and $\lambda^*_{Y_j}(x), j = 1, ..., m$ from its children given by

$\pi^*_X(u_i)$ is the probability of the most probable tail-extension of the hypothetical value $U_i = u_i$ relative to the link $U_i \rightarrow X$ and is known as an *explanation*,

$\lambda^*_{Y_j}(x)$ is the conditional probability of the most probable head-extension of the hypothetical value $X = x$ relative to the link $X \rightarrow Y_j$, known as a *forecast*.

More precisely, given the fixed local probability $P(x|u_1, ..., u_n)$ and the best value of X as x^*, the propagation concerns with:

Updating BEL^*: For $F(x, u_1, ..., u_n) = \prod_{j=1}^m \lambda^*_{Y_j}(x) P(x|u_1, ..., u_n) \prod_{i=1}^n \pi^*_X(u_i)$, $BEL^*(x) = \beta \max_{u_k} F(x, u_1, ..., u_n)$, $1 \leq k \leq n$; $x^* = arg \max_x BEL^*(x)$.

Parent-bound n messages to $U_1, ..., U_n$: $\lambda^*_X(u_i) = \max_{x, u_k : k \neq i} \frac{F(x, u_1, ..., u_n)}{\pi^*_X(u_i)}$, $i = 1, ..., n$.

Child-bound m messages to $Y_1, ..., Y_m$: $\pi^*_{Y_j}(x) = \beta \frac{BEL^*(x)}{\lambda {Y_j}^*(x)}$, $j = 1, ..., m$.

Boundary conditions: three types of nodes set up the boundary conditions.
1) Anticipatory nodes: uninstantiated variables with no children. For such a node X, $\lambda^*_{Y_j}(x) = [1, \ldots, 1]$. 2) Evidence nodes: instantiated variables. For variable $X = x'$, it is regarded as X being connected with a dummy child Z such that $\lambda^*_Z(x) = \begin{cases} 1 & \text{if } X = x' \\ 0 & \text{otherwise} \end{cases}$ and other real children of X, $Y_1, Y_2, ..., Y_m$, receives the same message $\pi^*_{Y_j} = \lambda^*_Z(x)$ from X. 3) Root nodes: variables with no parents. Similarly, for each root variable, a dummy

[3] In this article, all the incoming evidence will be denoted by $\mathbf{e}$ and be regarded as a set of instantiated variables $\mathbf{E}$. Symbol α will be used to denote a normalising constant and β will be used for an arbitrary constant.

[4] This $BEL^*(x)$ represents the probability of the most probable extension of $\mathbf{e}$ that is also consistent with the hypothetical assignment $X = x$.

parent U with permanent 1 instantiation is introduced and $P(x|u) = P(x) = \pi^*(x)$.

It is important to understand the conceptual essence of such a propagation mechanism. For each hypothetical value of a single variable X, there exists a best extension of the complementary variables $\mathbf{W'}_X$. The problem of finding the best extension of $X = x$ can be decomposed into finding the best complementary extension to each of the neighbouring variables according to the conditional independencies between X and the rest. This information can then be used to decide the best instantiation of X. The very process of this decomposition resembles the principle of optimality in dynamic programming in that it is applied recursively until it reaches the network's boundary where evidence variables have predetermined values.

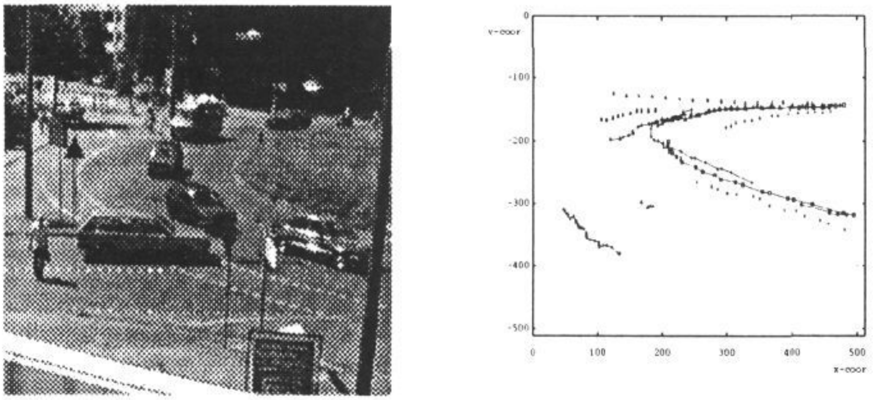

Figure 1: *Left: a traffic roundabout scenario and its traffic flow. Right: correlated spatio-temporal constraints on the movements of individual objects are imposed implicitly by this scene layout.*

3 Motion segmentation and tracking

In VIEWS, a vision system for surveillance applications, one of the key objectives is to segment detected optic flow field into dynamic regions corresponding to possible moving objects and to track these regions effectively and consistently over time. Wenz [20] applied a scheme based on estimated frame displacements of the extremal loci of a bandpass filter. "Similar" displacement vectors are grouped intc different moving regions (bounding boxes) in each frame and the similarity is defined by four parameters 1) neighbourhood range, 2) neighbourhood displacement magnitude ratio, 3) neighbourhood orientation difference and 4) neighbourhood vector numbers. In Wenz's approach, these similarity parameters are set as *independent constants* across the entire image. Although this direct approach is computationally rather straightforward, it is unable to deliver effective and consistent interpretations, especially in images of crowded scenes such as at a traffic roundabout shown in the left picture of figure 1. Left frames in figure 5 illustrate some typical defects in the sensitivity and consistency of such an approach. A more detailed analysis will be given in section 4. We propose that in order to obtain both effectiveness and efficiency, scene-oriented contextual knowledge has to be incorporated into the control of parameter values for focused computation.

VIEWS uses a fixed camera for collecting visual input at each scenario. Under such static camera configurations, three dimensional scene layout imposes indirect, but nevertheless invariant, constraints on both possible loci of appearances, sizes, speeds of bounding boxes and the overall traffic flow (see the right picture of figure 1). Therefore, scene layout defines visual expectations and constrains the setting of processing parameter values. In other words, the following correlated measures are constrained probabilistically with respect to image coordinates: 1) between object orientation and optic flow vector orientation; 2) between object size and flow vector neighbouring speed ratio, 3) between neighbouring orientation difference, object dx, object dy and object bounding box width or height. Such probabilistic constraints on a bounding box set a compound network of coherent hypothetic variables (figure 2) that increases resistance to incompleteness and inconsistency in the flow fields. Such a network can best be modelled by a Bayesian belief network with dynamic setting of the hypotheses using belief revision propagation. With this approach, we regard segmenting similar flow vectors into possible moving regions in the image and tracking them down in time as providing a coherent, Most-Probable-Explanation of the detected flow fields by actively revising the distributed beliefs according to the dependent causal constraints.

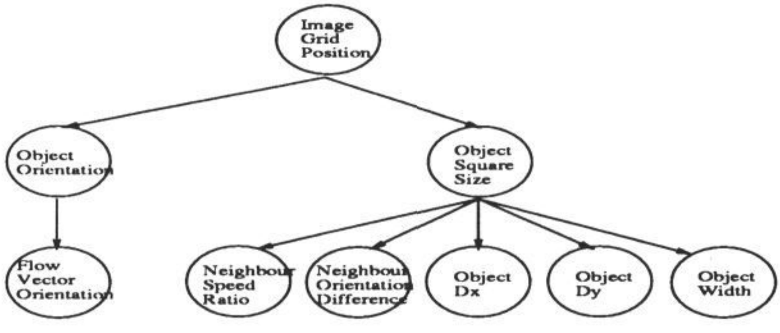

Figure 2: *A belief network that captures the dependent relationships between the scene layout and relevant measures in motion segmentation and tracking.*

The belief network in figure 2 is purposively constructed with a tree structure, a special type of "singly connected" network, in order to guarantee the propagation of message passing in belief revision to be tractable [14]. With the image (512 × 512) being divided into 25 grids, the root node IGP (Image Grid Position) in this tree represents the probabilistic expectation in the occurrence rate of objects in image grid positions. Nodes OSS and OOR represent respectively the probabilistic expectations in the square size and orientation of bounding boxes in image grids. The six leaf nodes at the bottom level of the tree represent, respectively, the expectations in flow vector orientation (FVO), neighbouring vector speed ratio (NSR), orientation difference (NOD), x component in object bounding box's displacement (ODX), y component in bounding box displacement, and the width of a bounding box (OWD).

It is important to point out that first, leaf nodes are the evidence nodes and it is desirable to relate them to *qualitative* measures by representing *relative* quantities of flow vectors. This is designed to overcome the instability of individual vectors in optic flow fields. Second, great effort was made to reduce the number of causal connections and the number of hypothetical variables to the minimum at the expense of approximations in the representation of certain variable nodes. This is because the computational load increases by an order of $2^n - 1$ where n

is the number of variable nodes in a network [8]. Node OSS is also taken as the approximation for the neighbouring vector searching range. The actual size of a bounding box is determined by the grouped number of neighbouring flow vectors and the loci of these vectors. The location of a bounding box is given by the centre of gravity of the member vectors and its initial velocity is estimated by the mean velocity of the vectors grouped in the initial frame. Third, in order to have efficient computation, it is crucial to balance the compromise in the approximation of hypothetical values and the accuracy of their representations. It is computationally attractive to approximate any continuous variable with a set of few discrete values. Fourth, the conditional probability distribution matrices between any two nodes are usually subject to probabilistic estimation based on extensive test examples. Statistical studies in the past [8] suggest that if well controlled number of variables are built into a Bayesian network, the estimated distribution matrices are not just merely appropriate numbers that can explain away a set of examples without capturing the general characteristics of the phenomena. Still, accurate estimation of these parameters remains one of the important factors for computational success of a belief network. Recent studies by Spiegelhalter [16] have shown techniques for updating and learning of the distribution matrices dynamically in order to provide more accuracy in their estimation. Finally, the algorithmic steps of our approach for the segmentation and tracking of object bounding box from optic flow fields are: 1) Set the maximum expected number of object in a scene and initialise such a number of belief nets. 2) Set $\lambda^*_{Y_j}(x_i) = [1, \ldots, 1]$ where $X_i = [FVO, NSR, NOD, ODX, ODY, OWD]$ and $P(x|u) = P(x) = \pi^*(x)$ where $X = [IGP]$, then initial equilibrium of a belief tree is obtained by (a) propagating all the *lambda* messages upwards, (b) propagate all the π messages downwards, (c) estimate the local beliefs throughout the tree, and (d) obtain a composite set of local instantiations of each variable that together is the best interpretation of the initial. "no evidence", condition. 3) For the first image frame, vectors are grouped according to the best value assignments associated with beliefs corresponding to their image grid position. For successive frames in the sequence, vectors are grouped according to best values, either to beliefs associated with previous tracked bounding boxes, or to beliefs associated with image grid positions. 4) For each calculated measure in the similarity test procedure, the value instantiates the associated node and revises local belief as well as other nodes' beliefs by propagation until the tree reaches equilibrium. 5) Revise locally every node's best value assignment so that the bounding box will set the most probable similarity threshold values for grouping vectors that are near to its expected location in the next image frame. Repeat steps 3 to 5.

4 Experiments and evaluation

The current design of the belief network has been tested extensively on image sequences from the traffic roundabout scenario. In the following, we measure the performance and computational cost of both the belief revision and the direct approach and discuss their effectiveness against their efficiency.

In assessing performance, we first show the sensitivity of the techniques by measuring their false alarm rate before we measure the consistency of both techniques in tracking individual objects over time. The false alarm rate was taken over an image sequence of 400 frames.

The left graph in figure 3 shows the false alarm rate on both techniques over time. It gives a good indication that the belief revision approach increases the true identifications significantly without introduce excessive false alarms. Throughout

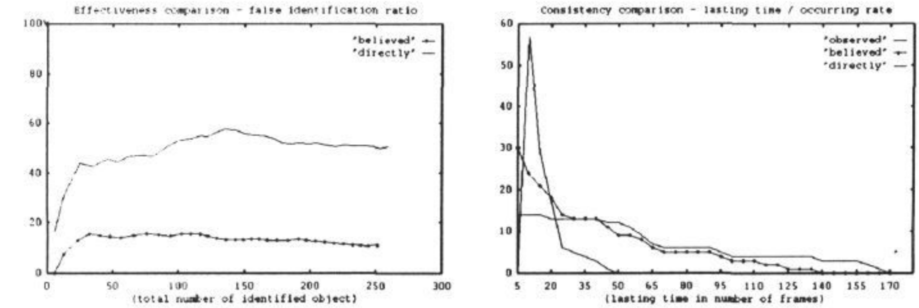

Figure 3: *Left: the false alarm rate. Right: the "ground truth" and detected number of objects and their durations in the scene over a sequence of 170 image frames.*

the whole sequence, the maximum false alarm rate from the belief revision approach is about 16 %, which is below the minimum rate from the direct approach. The maximum false alarm rate of the direct approach, on the other hand, reaches 60 % and its average rate is nearly 50 %!

For measuring consistency, we compile the histories of tracked objects from both techniques and compare them with the "ground truth" of a 170 frames image sequence we collected independently. In the right graph of figure 3, the flat and long lasting line shows the ground truth of the number of objects against their durations in the scene. For example, 1 object that has stayed for the entire 170 frames, 13 objects which have lasted for 14 frames, etc. The sharp pulse line shows that the direct approach has taken fragments of objects with long durations and tracked them as a large number of objects with very short histories. There is no object being tracked for more than 50 frames. That shows poor consistency. In contrast, the dotted line shows that the belief revision approach provides with much accurate measure of both the number of objects and their durations.

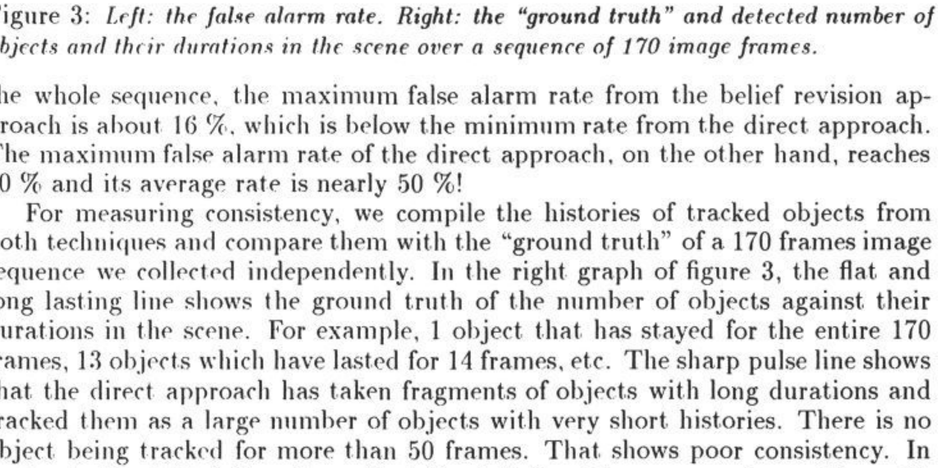

Figure 4: *Left: time consumption in seconds for the belief revision and the direct approaches respectively, and their first order derivatives over time. Right: percentage increase in belief revision approach's time consumption.*

For estimating the computational cost, we first measure the absolute time consumption (in seconds) of both schemes over the 400 frames sequence, see the two near linear increasing lines in the left graph of figure 4. The divergence between the two lines is rather deceptive since it appears to show a continuous increase of processing time in the belief revision scheme. However, it actually shows the accumulated cost of bootstrapping belief networks over time. The frame by frame computational cost is more realistically given by the first order derivative of time over those two lines, which are shown by the two step lines. This can be seen more clearly by measuring the percentage of the increased time consumption in the belief revision approach from the direct approach (see the right graph in

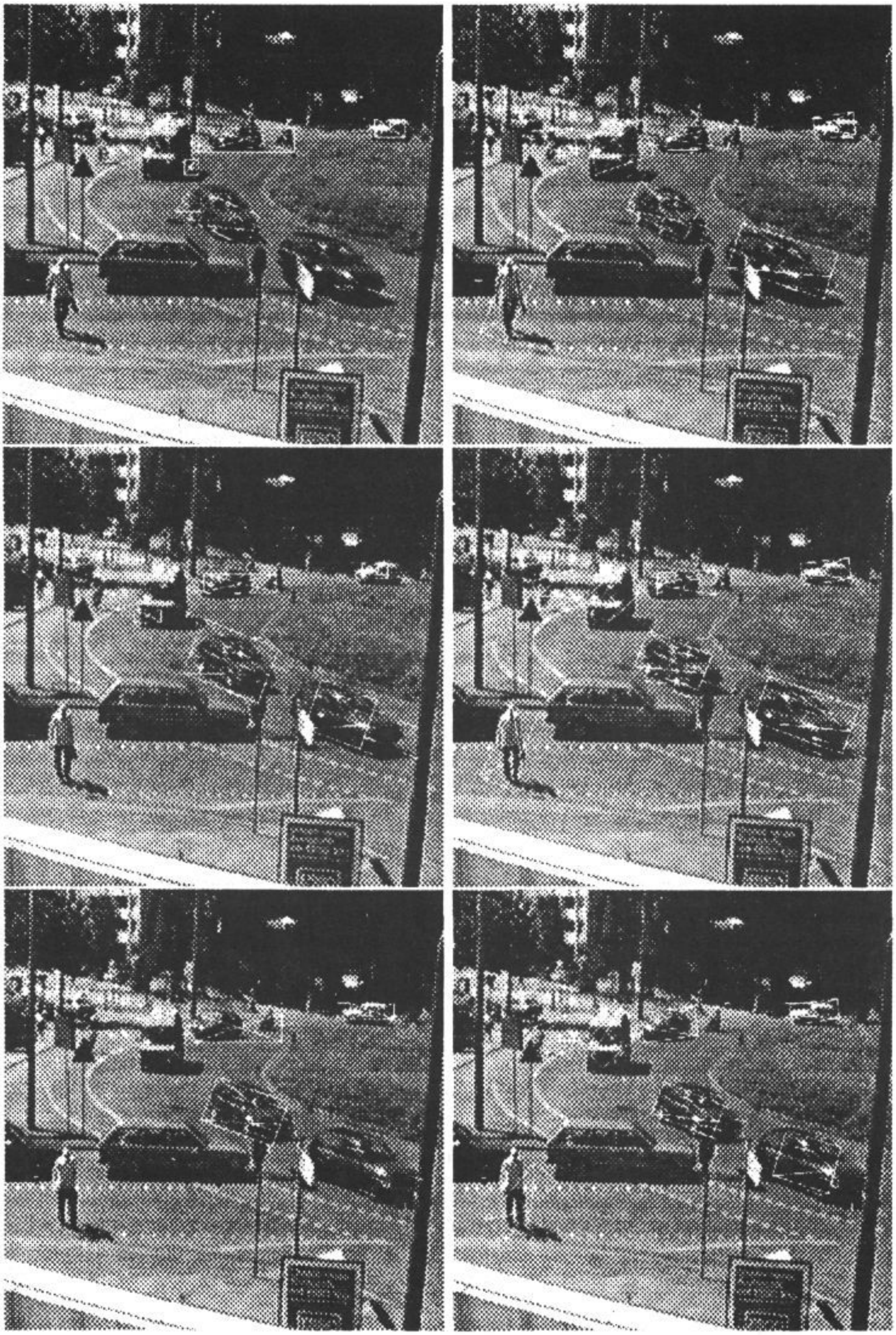

Figure 5: Results on frames 140, 145 and 150. Left: by the direct approach; right: by the belief revision approach.

figure 4). The former's frame by frame computational overhead against the later throughout the whole sequence is below 13 %, and it is worth pointing out that providing more accurate segmentation and tracking of objects instead of missing identifications requires "extra" computational cost.

Our quantitative measures presented here illustrate that: with very limited cost in computational efficiency, significant gains are obtained in effectiveness by using the belief revision technique. A more visual comparison between the two approaches can be seen in figure 5. Three successive frames from our test sequence are shown with the results from the direct approach on the left and from the belief revision approach on the right. It is worth noticing that: first, the belief revision approach is very robust against incomplete evidence (see the tracked cyclist behind a sign post to the left hand side of frames 145 and 150). Second, it is capable of segmenting very closed moving objects (see the cyclist and the two cars close to its right). Third, one of the reasons for unnecessary time being taken in the current belief revision process is caused by the simulation of distributed message passing procedures in the belief propagation on sequential machines.

5 Conclusions

Our main argument in this paper concerns the need to build in knowledge even at the earliest stages of visual processing in order to deliver both effective and efficient performance on visual tasks. The specific example elaborated here uses scene-oriented contextual knowledge to improve the sensitivity and consistency of the segmentation and tracking of moving objects in the image. We have proposed the Bayesian belief revision network as an appropriate model for representing such conceptual knowledge and the associated belief propagation as the suitable mechanism for effective and efficient constraint propagation within such a framework. We examined relationships between our conceptual knowledge of the traffic scenes and of the image sequences for the roundabout scenario. We then identified the required implicit computational constraints to the beliefs that specify the dependencies between processing parameters involved. We presented the way in which a specific belief network can be designed for grouping optic flow fields at a traffic roundabout scenario.

In conclusion, the results obtained so far show that the computational overhead introduced by mapping explicit knowledge to implicit constraints for controlling selective processing is small considering the correct number of moving objects identified and the improved consistency in both segmentation and tracking. The belief revision increases the sensitivity to incomplete evidence so that finds moving objects missed by the direct approach. An extension of this approach will be examined for model-based object recognition, tracking and behavioral evaluation in 3D space.

References

[1] Y. Aloimonos, I. Weiss, and A. Bandopadhay. "Active Vision". In *ICCV*, 87.

[2] R. Bajcsy and P. Allen. "Sensing Strategies". In *US-France Robotics Workshop*, 84.

[3] D. Ballard. "Reference frames for animate vision". In *IJCAI*, 89.

[4] V. Braitenberg. *Vehicles: Experiments in Synthetic Psychology.* 84.

[5] R. Brooks. "A Robust Layered Control System for a Mobile Robot". *IEEE J. Robotics and Auto.*, RA-2, 86.

[6] R. Brooks. "Intelligence without reason". In *IJCAI*, 91.

[7] P.J. Burt. "Smart Sensing in Machine Vision". In *Machine Vision: Algorithms, Architectures and Systems*. 88.

[8] E. Charniak. "Bayesian Networks without Tears". *AI Magazine*, 12, 91.

[9] S.G. Gong. "Visual Observation as Reactive Learning". In *SPIE International Conference on Adaptive and Learning Systems*, 92.

[10] S.G. Gong and H. Buxton. "On the Expectations of Moving Objects". In *ECAI*, 92.

[11] T.S. Levitt, T.O. Binford, and *et al.* "Probability based Control for Computer Vision". In *DARPA IUW*, 89.

[12] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 82.

[13] V. Murino, M.F. Peri, and C.S. Regazzoni. "Distributed Belief Revision for Adaptive Image Processing Regulation". In *ICCV*, 92.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference*. 88.

[15] R.D. Rimey and C.M. Brown. "Selective Attention as Sequential Behavior: Modeling Eye Movements with an Augmented Hidden Markov Model". Technical report, Comp. Sci. Dept., Uni. of Rochester, 90.

[16] D.J. Spiegelhalter and R.G. Cowell. "Learning in Probabilistic Expert Systems". In *Bayesian Statistics 4*. 92.

[17] J.K. Tsotsos. "Knowledge Organisation and its Role in Representation and Interpretation for Time-Varying Data: the ALVEN System". *Computing Intelligence*, 1, 85.

[18] J.K. Tsotsos. "On the Relative Complexity of Active vs. Passive Visual Search". *IJCV*, 7, 92.

[19] S. Ullman. "Visual routines". *Cognition*, 18, 84.

[20] G.H. Wenz. *Parallel Realtime Detection and Tracking*. Master's thesis, QMW, Uni. of London, 92.