Evidential Reasoning for Object Recognition in Wavelet Scale Space*

Ivan Marsic
CAIP Center, Rutgers University
Piscataway, NJ 08855–1390
e-mail: marsic@caip.rutgers.edu

Abstract

Evidential reasoning is the central problem of selective visual perception. It consists of gathering the evidence and integrating the information from successive fixations. Both data–driven and knowledge–driven influences in selecting the areas of interest and gathering the evidence are studied, as well as their cooperative function. We use the Bayesian networks for knowledge representation, and a decision theoretic approach in selecting where to shift the attention and for integrating the information from successive fixations. Depending upon the task, the system decides what information to gather from the scale–space decomposition of an input image.

Keywords: *object recognition, selective perception, Bayesian networks, multiresolution analysis.*

*The research reported here was supported by the Center for Computer Aids for Industrial Productivity (CAIP). The CAIP Center is supported by the New Jersey Commission on Science and Technology and the Center's Industrial Members.

Evidential Reasoning for Object Recognition in Wavelet Scale Space

Abstract

Evidential reasoning is the central problem of selective visual perception. It consists of gathering the evidence and integrating the information from successive fixations. Both data–driven and knowledge–driven influences in selecting the areas of interest and gathering the evidence are studied, as well as their cooperative function. We use the Bayesian networks for knowledge representation, and a decision theoretic approach in selecting where to shift the attention and for integrating the information from successive fixations. Depending upon the task, the system decides what information to gather from the scale–space decomposition of an input image.

SUMMARY

1. The contribution of this work is in proposing the framework to manage the inherent intractability of visual object recognition. The model integrates both data–driven and knowledge–driven influences in directing the selective perception.

2. The model scales up with the complexity of the scenes and the requirements of a particular task, and it is domain independent. Thus, it offers an approach in solving of intractability of general recognition problem.

3. The most closely related work is by Rimey and Brown (1992–1994). Their model of selective perception almost exclusively relies on knowledge for directing the attention, whereas our model relies mostly on data–driven influences in deciding about the next attention spot. We also addresses the integration of data–driven and knowledge–driven influences.

4. The model can be used in data–driven selection of the regions of interest, as well as for tracking of objects or their features. The model also provides a scheme for integrating visual information from successive fixations.

1 Introduction

In order to solve the difficult problem of recognition, the visual system must apply some simplifications. It appears that biological vision escapes the trap of overwhelming complexity by varying the level of descriptive abstraction—the amount of detail captured—depending upon the specific task. Experiments of Yarbus [12] have shown that the eye movements of the observer will depend upon the task the observer is required to perform. Moreover, additional time spent on perception is not used to examine the secondary elements, but to re–examine the most important elements of the picture.

There is typically an abundance of evidence in images, which however is always partial and sometimes incorrect due to occlusion, deformation, noise, and imperfect processing algorithms. When combined, the data may be used to deduce facts about the physical objects. This is why uncertain methods of inference, particularly Bayesian networks [6], are suitable for visual tasks to rank order conflicting scene interpretations that arise from unreliable evidence [1, 2, 3, 8, 11].

The central hypothesis of this work is that any object can be assigned the same form of representation with the same level of detail, regardless of its size or shape complexity. This is a minimum representation for an object or its feature. By means of wavelet scale–space, the image can be decomposed into the pieces having the minimal object representation. These pieces are "recognition quanta" or elementary pieces of evidence in our model. Based on wavelet analysis, we derived [5] that this representation is the wavelet coefficients of object's characteristic scale. These coefficients determine the object's base class. The same is valid for object features which determine the object's sub-class. This hypothesis leads to the

specific architectural constraints for a visual system, and specific issues to be solved within this framework.

The whole object is represented with one or more *icons*, which is the collection of wavelet coefficients in a corresponding region of scale space. The number of the regions that needs to be processed for a particular object depends upon its shape complexity. Also, the selection of a particular region to be processed is task–dependent.

This paper is organized in the following way. Section 2 first provides a brief overview of the scheme for decomposing an image into the scale–space cells and computing the saliency maps. The scheme is presented elsewhere in detail [5]. Section 3 deals with knowledge representation using Bayesian networks, whereas Section 4 deals with task–specific evidential reasoning and cooperative function of data–driven and knowledge–driven influences in selecting the areas of interest and gathering the evidence. The basis for the knowledge–driven selective perception and evidential reasoning part of the model presented here is Rimey and Brown's work [7, 8]. The implementation is described and the results are reported in Section 5. The implications of the results are discussed in Section 6.

2 Scale–Space Cells and Saliency Maps

We propose that low–level vision processes should decompose and represent image structures (objects and their features) with a minimum amount of information. For this purpose we use the wavelet representation of an object or its features which we call *scale–space cell* (SSC) [5]. The scale–space cell of an object is located at the scale which is approximately one octave below the scale at which an object becomes smoothed out and thus indistinguishable from other objects of the same size. This is illustrated in Figure 1. Small objects span small

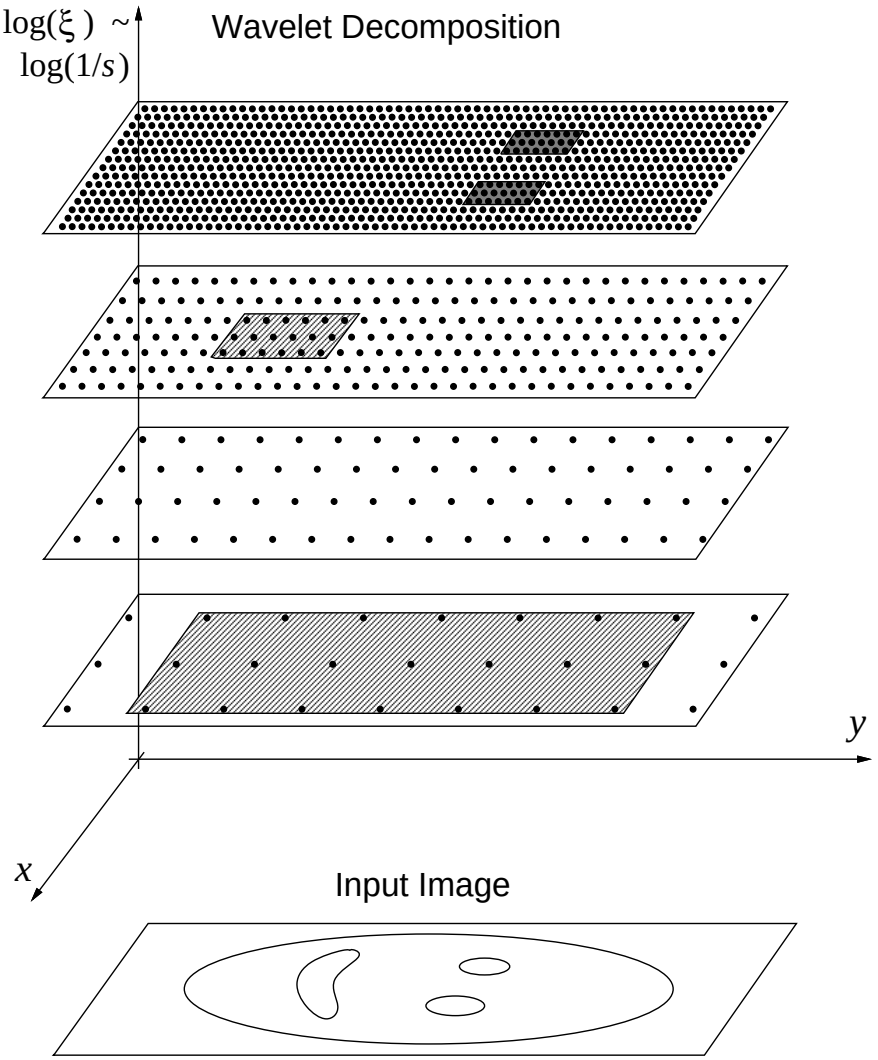

Figure 1: The proposed system uses the same amount of input data for classification purposes, no matter what the coordinates of the object's SSC in scale–space are. Note that the frequency ξ and the scale s are inversely proportional.

portions of the image but require denser calculations of wavelet coefficients since their SSCs fall in the region of wavelets with small support. The converse is true for large objects, i.e., they span large portions of the image but require sparser calculations of wavelet coefficients. Thus, the system uses the same number of wavelet coefficients for classification purposes, no matter what the coordinates of the object's SSC in scale–space are (within the range determined by the image's resolution of the receptor mosaic and the viewport size). In other

words, the form of the SSCs is self–similar no matter what is the shape complexity of the object or feature that it represents.

The saliency computation is accomplished in four steps [5], to be executed in the following order:

Step 1—Wavelet Transform: Compute the wavelet pyramid of an image with J dyadic scales using oriented quadrature bandpass filters tuned to 16 different orientations, i.e. $\Theta = 0°, 22.5°, 45°, ..., 337.5°$. See [10] for a detailed derivation and description of computing the wavelet pyramid using the steerable basis filters.

Step 2—Local Energies: Compute the oriented local energies using the equation:

$$E(\Theta, s, x, y) = \left[G^{\Theta}(s, x, y)\right]^2 + \left[H^{\Theta}(s, x, y)\right]^2 \tag{1}$$

where $G^{\Theta}(s, x, y)$ and $H^{\Theta}(s, x, y)$ are the outputs of a quadrature pair of analyzing wavelet filters at the scale–space coordinate (s, x, y), oriented at the angle Θ. For each image point, 16 different oriented local energies are calculated.

Step 3—Saliency Maps: Compute J saliency maps. First reset all the points of a saliency map to zero. Compute the saliency of each particular SSC using the equation:

$$\text{saliency SSC}(s, x, y) = \max_{\Theta} \{E(\Theta, s, x, y)\} * \vartheta(x, y) \tag{2}$$

where $\vartheta(x, y)$ is the filter kernel obtained by computing the highest oriented energies at each location at the finest scale for a disc with a radius of 4 points. The obtained number is located at the SSC center.

Step 4—Peaks in Saliency Maps: Find the peaks in each saliency map so that there is no overlapping between the salient SSCs in a particular map.

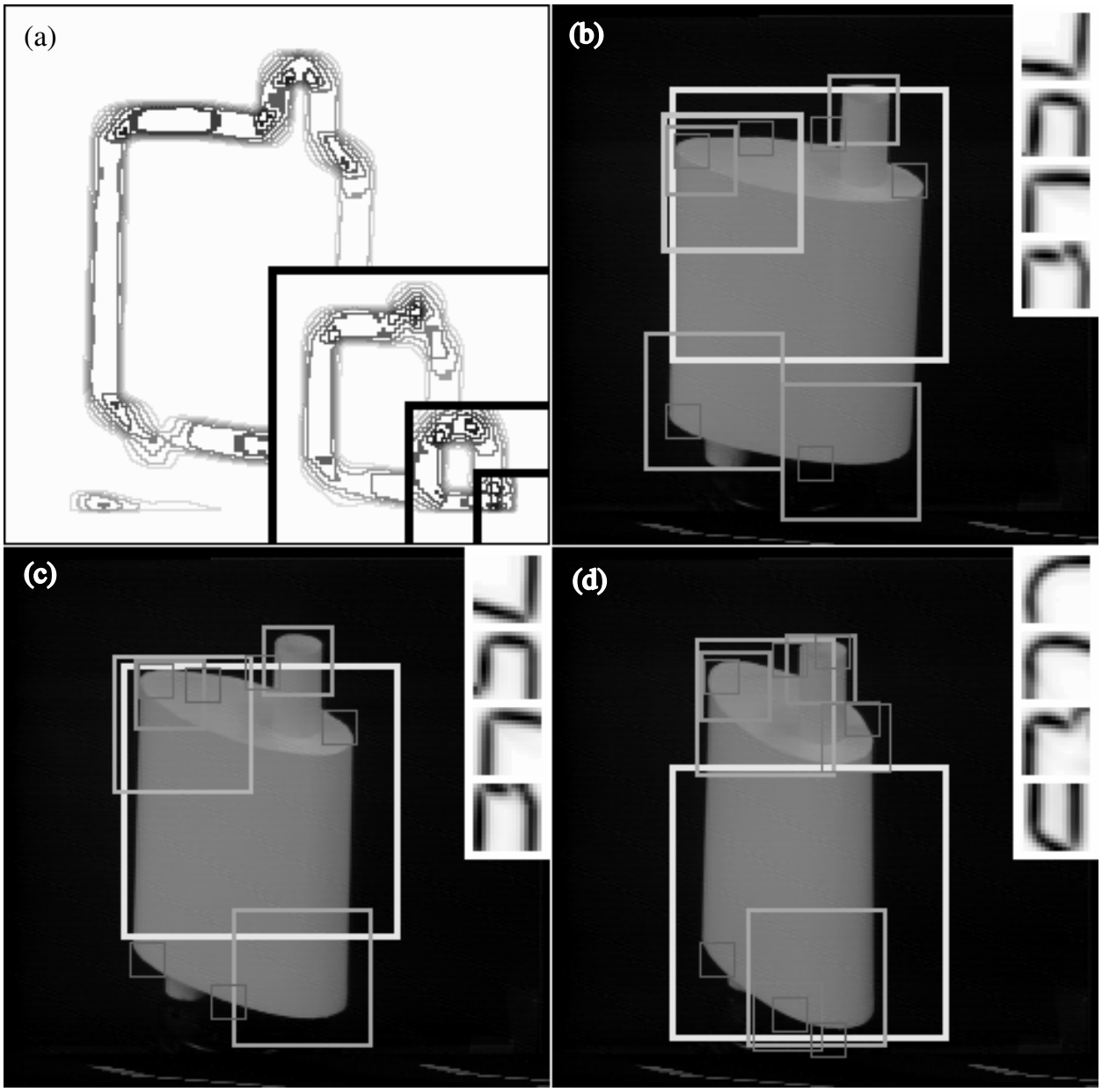

Figure 2: (a) Contour plots of the saliency maps at four scales of the "muffler" image. (b) Selected salient SSCs remain consistent as the object rotates by 20° (c), and as much as 40° (d). The most salient icons from the finest to the coarsest scale, going from the top to the bottom, are shown as insets in the corresponding images.

There are no parameters to be tuned experimentally during the algorithm execution. Figure 2 shows an example of searching for the most salient SSCs. It can be observed in Figure 2a that objects and their features produce peaks in the saliency maps. We are interested in detecting those icons which are most informative about the objects in an image[1]. The results of the quantitative measure of informativeness coincide well with what we intuitively consider to be the informative details, with some notable exceptions [5]. These simple examples and the more complex ones presented in [5] demonstrate the robustness of the scheme.

In addition to the recognition of the content of any of the individual SSCs, we can identify two major issues that will be addressed in this paper:

(i) Selecting a particular SSC to be processed at the next moment.

(ii) Integrating visual information from successive fixations.

Selection and integration are the major constituents of evidential reasoning. We give an example of evidential reasoning where the particular task to be solved is visual object recognition.

3 Knowledge Representation Using Bayesian Networks

We assume that a problem to be solved can be represented with a set of probabilistic variables and their dependencies. Each probabilistic variable consists of two things: a name, and the values or states. The steps necessary to create a Bayesian network are the following [6, 2]:

[1]The objects in an image do not necessarily correspond to the physical objects in the scene. For instance, shadows may be structurally significant in the image while they do not exist as independent objects in the scene.

1. Decide on the direct cause–effect relations in the domain in question and represent them as a directed acyclic graph. The nodes in the graph correspond to variables.

2. Decide the set of possible states (values) for each node in the graph, e.g., in the network presented below, the set of states for **muffler-quality** are *good, rework*, and *scrap*.

3. Decide on the conditional probabilities for all states in a node, conditioned on the states in the parent nodes. That is, the strength of the cause–effect relations are modeled as conditional probabilities.

The problem that we consider is to determine whether a valid muffler is present in an input image and reporting whether it is *good*, flawed but *reworkable*, or flawed and must be *scrapped*. Here is an example of how we design the network. Let us assume our problem has just three variables: **muffler-quality**, **valid-configuration**, and **damage-present**. Let us denote them respectively as Q, V, and D. We have defined the problem if we know how to calculate the joint probability distribution $P(Q, V, D)$. Since we want to infer muffler quality from evidence about validity and degree of damage, we prefer the form of the joint distribution where we have a marginal distribution only for Q. So, applying the chain rule formula [6] we have:

$$P(Q, V, D) = P(Q)P(V, D|Q) \tag{3}$$

Since we know that whether the muffler is valid or not is independent of whether it is damaged or not, for the second part we get $P(V, D|Q) = P(V|Q)P(D|Q)$. Therefore, we have:

$$P(Q, V, D) = P(Q)P(V|Q)P(D|Q) \tag{4}$$

This equation can be represented with the graph shown in Figure 3a. Note that we could orient these links either way by applying Bayes' rule. What matters in orienting the links this way is that we want to perform abductive reasoning about quality from evidence about validity and from the degree of damage. In other words, knowing the state of Q generates expectations about V and D.

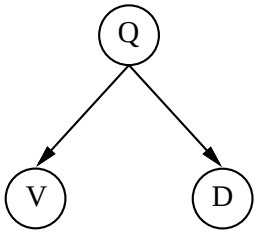

Figure 3: The initial Bayesian network representing the task (see text).

We could continue building this network by describing the task in terms of more and more "localized" evidence. However, since we are dealing with two kinds of knowledge (one is task–specific, and thus temporary, and the other is long–term system knowledge), we separate these two kinds of representations. For this purpose we use the composite network introduced by Rimey and Brown [7, 8] which consists of four types of networks: PART–OF, IS–A, expected area (EA) network, and TASK network. The reason for introducing different types of networks is to have a better structuring of the knowledge. It is natural to expect that different types of representations would be best suited for different types of knowledge.

An example PART–OF network for the muffler quality problem is shown in Figure 4. This network simply encodes all the knowledge about any particular class of objects known to the system (domain knowledge). It does not depend upon the task currently being solved by the system.

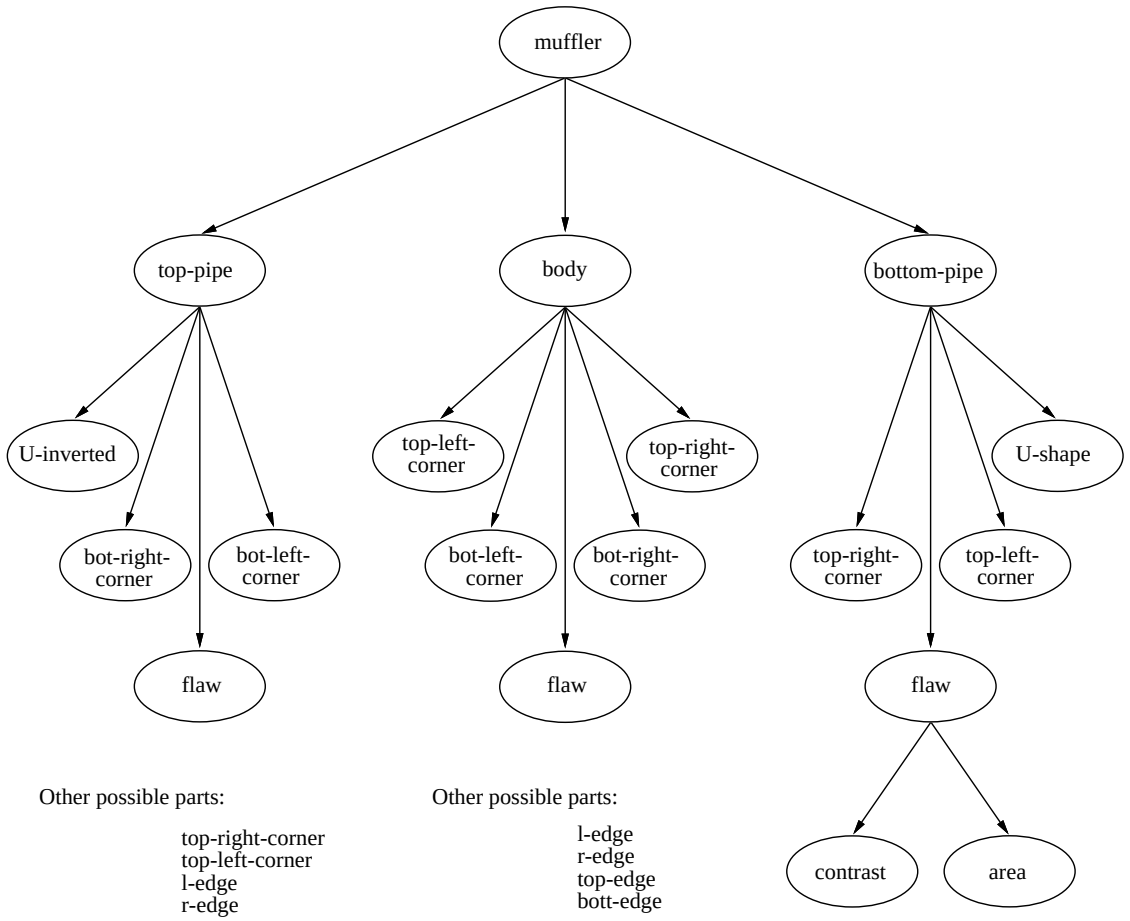

Figure 4: The PART–OF network showing the geometric decomposition of a muffler. It answers about the object/part presence.

A flaw may be a part of the pipes or the body. Since it is irrelevant whether a flaw is present for the abductive reasoning about the muffler presence, the distribution $P(\textbf{flaw}|\textbf{m-part})$ is made uniform, where **m-part** can be the pipes or the body. Note also that a flaw is a *property*: it is defined as an object anywhere inside the muffler shape rather than by its shape like other objects and parts in Figure 4. Since EA networks represent relative spatial position rather than other types of relationships (e.g. "inside", "between", etc.), we solve this problem by representing the flaw as an *ad hoc* chosen icon.

While the other types of Bayesian networks have minimal memory requirements, the *expected area* network requires relatively large amount of memory for each node, as well as specific circuits for belief calculation (convolution operation). If one assumes that eventually all of the system's knowledge will be represented using the above defined types of Bayesian networks, it poses tremendous burden in terms of resource requirements. However, due to the particular structure of the scale–space (see Figure 5), the global coordinate maps may be omitted from node information as they contain no long–term information. The global coordinate maps are needed only during the reasoning phase. Thus, the EA nodes information (relation maps) is stored in the long term store, and particular sub-network is pulled out from the store into the *reasoning engine* during the object/scene recognition.

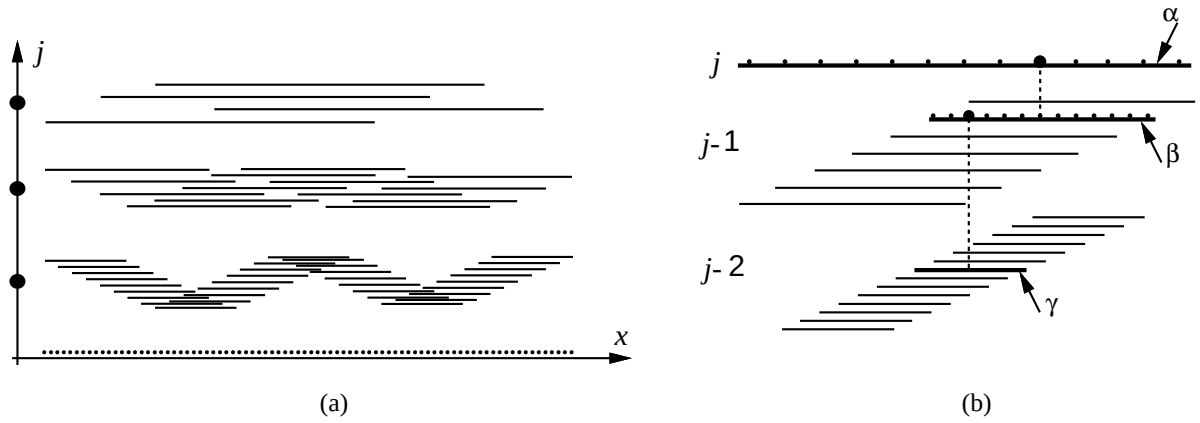

Figure 5: (a) The spatial layout of the SSCs in the one–dimensional scale–space. The abscissa represents the time (or spatial position) and the ordinate represents the scale. (b) An example where γ SSC is represented in the local coordinate system of β and β is represented in the local coordinate system of α.

The PART–OF, expected area, and IS–A networks encode all the information that can come up in any scene containing a particular class of objects (i.e. the network encodes domain specific knowledge), whereas the TASK network encodes only the information that the system is looking for, in order to answer the posed question (task–specific knowledge). A different

TASK network is created for each task the system is to solve [8].

The TASK network for the problem of inspecting muffler quality is shown in Figure 6. Note that task is simplified so that the spatial position of the flaw does not matter, i.e., a flaw is treated in the same way no matter whether it is part of the muffler body or its pipes. For this purpose the PART–OF network in Figure 4 should be modified so that the variable **flaw** becomes direct child of the variable **muffler** rather than being a child of its **top-pipe, body**, and **bottom-pipe**.

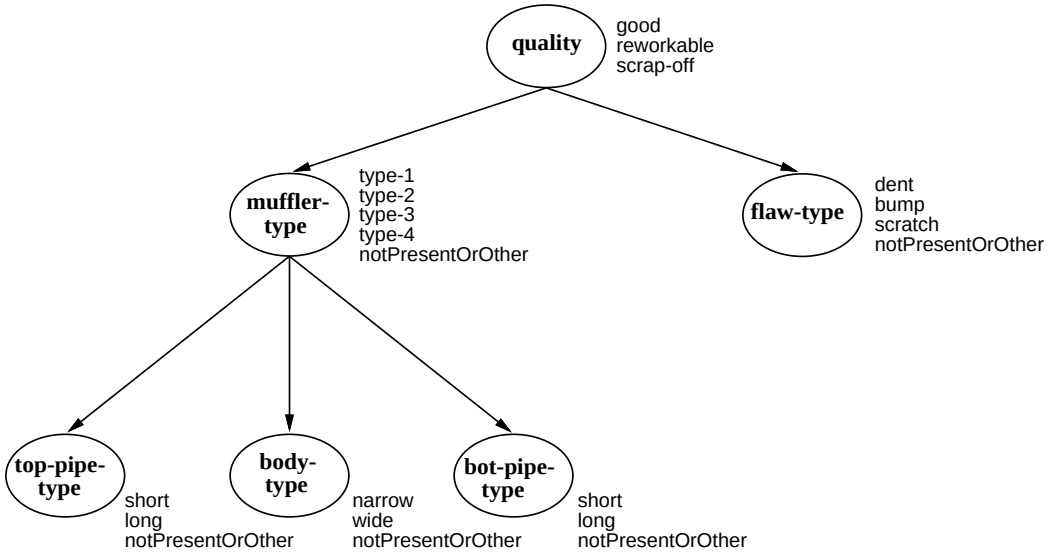

Figure 6: The TASK network for the problem of inspecting muffler quality.

A TASK network encodes what subset of scene information would be useful for solving the task, but says nothing about *how* to obtain that information. This will be considered in the next section.

BEL values are computed in the composite network as follows [8]: (1) Propagate belief in each of the networks, except the TASK network. That is, each of the separate networks in the composite network, except the TASK network, maintains its BEL values independently of the other networks. (2) Construct *packages* of BEL values from the other networks for

transfer to the TASK network. A package is treated as an evidence report that is attached to a node in the TASK network (see details in [8]). (3) Propagate belief in the TASK network.

4 Knowledge–Driven Focal Attention

In the previous section we have considered (passive) knowledge representation using belief networks. Belief network contains only the prior probabilities of the root node(s) and the conditional probabilities of the non–root nodes. We can obtain the prior beliefs about the non–root nodes by propagating this information. In order to use the network for a particular image, the system has to supply it with evidence about the image. The evidence is then propagated through the entire network until new belief values for all nodes are established. This operation is repeated until some hypothesis accumulates enough evidential support to be declared a solution of the problem. The goal of an attention mechanism is to decide where in the image the system should gather evidence. We consider knowledge–guided attention mechanisms, and integration of these mechanisms with data–driven mechanisms.

Figure 7 represents the architecture of the machine vision system proposed here. Presently, this system is only partially implemented. For example, the decision making module is completely missing. Object recognition is based on classification of the icons of the most salient SSCs. The icon classification is not addressed in this work, but there are many available methods which perform well enough for the anticipated needs of the system (for instance, the technique by Simard *et al.* [9]). After an icon is classified, its class and scale–space coordinates are associated. The final stage is that of associating icon classes into a relational structure. Figure 7 shows no explicit association between the icon coordinate and its class. However, the associations are built–in implicitly since there must be exact correspondence

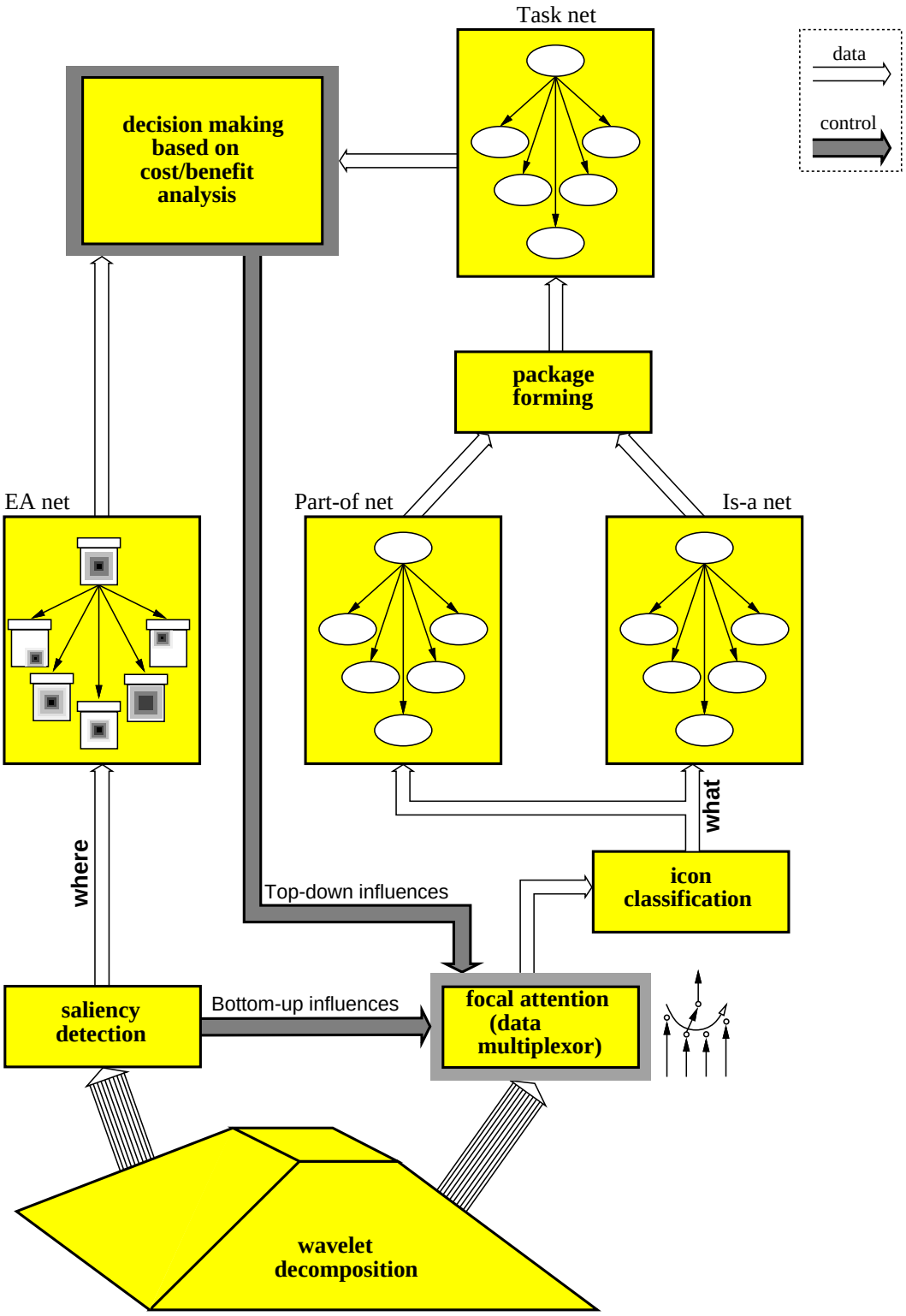

Figure 7: An architecture of a system for evidential reasoning for object recognition in the wavelet scale space.

between the PART–OF and expected area nodes.

4.1 Evidence Adding

The visual action(s) will determine all available information about an icon representing an object: object class, scale–space coordinate (s_v, x_v, y_v), and saliency value. The subscript v signifies that these are evidential values. Note that (x_v, y_v) is a coordinate of the SSC's center. The object class is determined by matching the object's icon against the object prototypes stored in the database of known objects.

If the action was saliency–initiated, the system considers adding the evidence to several Bayes nodes, as the detected object might correspond to any of the corresponding nodes. For example, if the system detects **bottom–right–corner**, it might be a part of **top–pipe** or a part of **body** (Figure 4). The system determines the list of nodes of the Bayesian PART–OF network which have the same name as the icon object class. (The expected area network should be isomorphic with the PART–OF network, and the corresponding nodes should have the identical names.) Evidence is added to only one node (since it signifies one object), and for each node, evidence is added at most once. For this reason, it is added only if it exceeds certain threshold value.

Evidence added to the nodes in the list should depend upon the expectation about the object presence at that particular scale–space coordinate (geometric constraints built in the EA net). For this purpose, the system reads the belief value $BEL(p_v)$ at the spatial coordinate $p_v = (x_v, y_v)$ of the corresponding EA node. This value is given as the percent of the highest value in the BEL array, i.e., it is normalized. To get belief value at the scale–space coordinate, we have to adjust this value for scale. The expected scale is determined from the

expected size of the object. Let the expected height and width dimensions of the object be denoted with $(E[h], E[w])$. The expected scale is given as $E[s] = log_2(\min\{E[h], E[w]\}) - 1$. The expectation for finding the object at the point $p'_v = (s_v, x_v, y_v)$ is calculated as

$$BEL(p'_v) = \frac{BEL(p_v)}{|E[s] - s_v| + 1}$$

The evidence to be added to the corresponding PART–OF node is determined as follows:

$$\lambda(present) = 0.5 + \frac{BEL(p'_v)}{2} \tag{5}$$

As an evidence to the corresponding EA node, the map is initialized with a Gaussian–shaped peak at the point $p_v = (x_v, y_v)$ with standard deviation

$$\sigma = \frac{1}{[\lambda(present)]^3} \tag{6}$$

Since $0.5 \leq \lambda(present) \leq 1$, it can be used to determine the standard deviation. The standard deviation is designed so that the peak is very sharp (almost a single point) for a high level of confidence in the object presence; it is almost uniformly spread for a very low level of confidence in the object presence.

Finally, the evidence about the object size should be reported. Since the belief values currently are not maintained for the object dimensions, the size of the detected object is stored directly as the EA property rather than posting it as an evidence. The evidential value is combined with the expected value as follows

$$s_{\text{new}} = \varepsilon s_v + (1 - \varepsilon)E[s] \quad \text{and} \quad h_{\text{new}} = w_{\text{new}} = 2^{s_{\text{new}}+1} \tag{7}$$

where weighting factor ε is determined according to the sigmoidal shape given in Figure 8. Since the scale assumes discrete values (octaves), this value has to be rounded to the nearest

integer value. The relationship indicates that the system is not willing to accept evidence that is very different from the expectation. The system should not accept some findings at all; every observation is uncertain because of the projection from a three–dimensional world to a two–dimensional image, a limited viewport size, noise, etc.

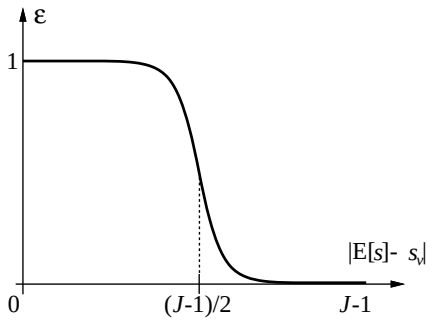

Figure 8: The shape of the weighting factor as a function of the absolute difference between the expected and detected object scale.

As the new SSCs are analyzed, the newly acquired information is fused with the already available information and used to further determine the next locus of attention. This effectively results in a sequence of fixations which depends on the image data, the object–model representation, and the visual task being performed.

The problem with relying only on data–driven influences (saliency) is that system fails to check for absence of important features. If the task is inspecting the muffler quality (Figure 6), then the system has to make sure that that all parts of the muffler are present, in addition to checking whether there are any flaws. The saliency–based module will direct the attention only to those parts which are present and salient. If a part is not present, no attention will be payed to the location of the missing part.

In order to circumvent this shortcoming, the decision–making module of the system should traverse the PART–OF network and check whether all evidence nodes have posted

evidence. If an evidence node has no posted evidence, the search should be initiated in order to check whether the corresponding part is missing, or it is present but not salient enough to draw the attention.

4.2 Summary of the Saliency–Guided Recognition Algorithm

The saliency–guided recognition starts with the saliency maps which are determined as summarized in Section 2. The beliefs in the EA network should be propagated ahead of recognition in order to set prior beliefs. These five steps are repeated for each salient SSC, in the following order:

Step 1—Next Salient Icon: Find the local maxima of each saliency map so that there is no overlapping between the salient SSCs. Obtain the icon associated with the next most salient SSC.

Step 2—Icon Classification: Classify this icon to determine the object class to which it belongs. If the icon does not correspond to any of the classes, add it to the knowledge base (learning). After this step the system has all data about the icon: object class, scale–space coordinate (s_v, x_v, y_v), and saliency value.

Step 3—Evidence Adding: Determine the list of nodes of the Bayesian PART–OF network with the same name as icon object class (expected area network should have the same nodes). Add evidence to the node for which the belief in presence of the corresponding object (Eq. (5)) is the highest using the equations (5)–(7).

Step 4—Belief Propagation: Propagate the evidence in both PART–OF and expected area networks.

Step 5—Package Propagation: Construct packages of BEL values from the other net-

works for transfer to the TASK network. Propagate the evidence in the TASK network.

5 Implementation and Results

The system is designed to be an experimental environment. It is organized as a set of processes which communicate via TCP/IP sockets [4]. Each module has a graphical user interface, allowing observation and direct control over the most critical points of the recognition process, thus providing the user with better insight into the particular stages of the process and enabling the test and improvement of hypotheses.

The initial setup of the entire system is shown in Figure 9. The figure shows two programs (`simple` and `Object Recognition`) on top, and below them two Bayesian networks: PART–OF and expected area networks. The nodes are initialized with prior probabilities. `Object Recognition` comprises the lower part of Figure 7 (wavelet decomposition, saliency detection and icon classification), whereas `simple` comprises the upper part of Figure 7 (Bayesian networks and decision analysis).

There are a total of thirteen salient SSCs for the image of the damaged muffler (upper right corner in Figure 9). The sequence of locations visited by the spot of attention during the recognition process is shown in Figure 10. This sequence might look different if the decision analysis module of the system were implemented (Figure 7). For example, the order might be different, and some steps may be skipped as not being informative enough.

As the attention shifts to the next most salient SSC, the corresponding evidence nodes of both networks get updated. The evidence that is gathered here is the identity of an object/part within the SSC and its coordinate in the scale–space. As explained in Section 4.1, evidence is added to the PART–OF nodes based on the expectation about their presence at

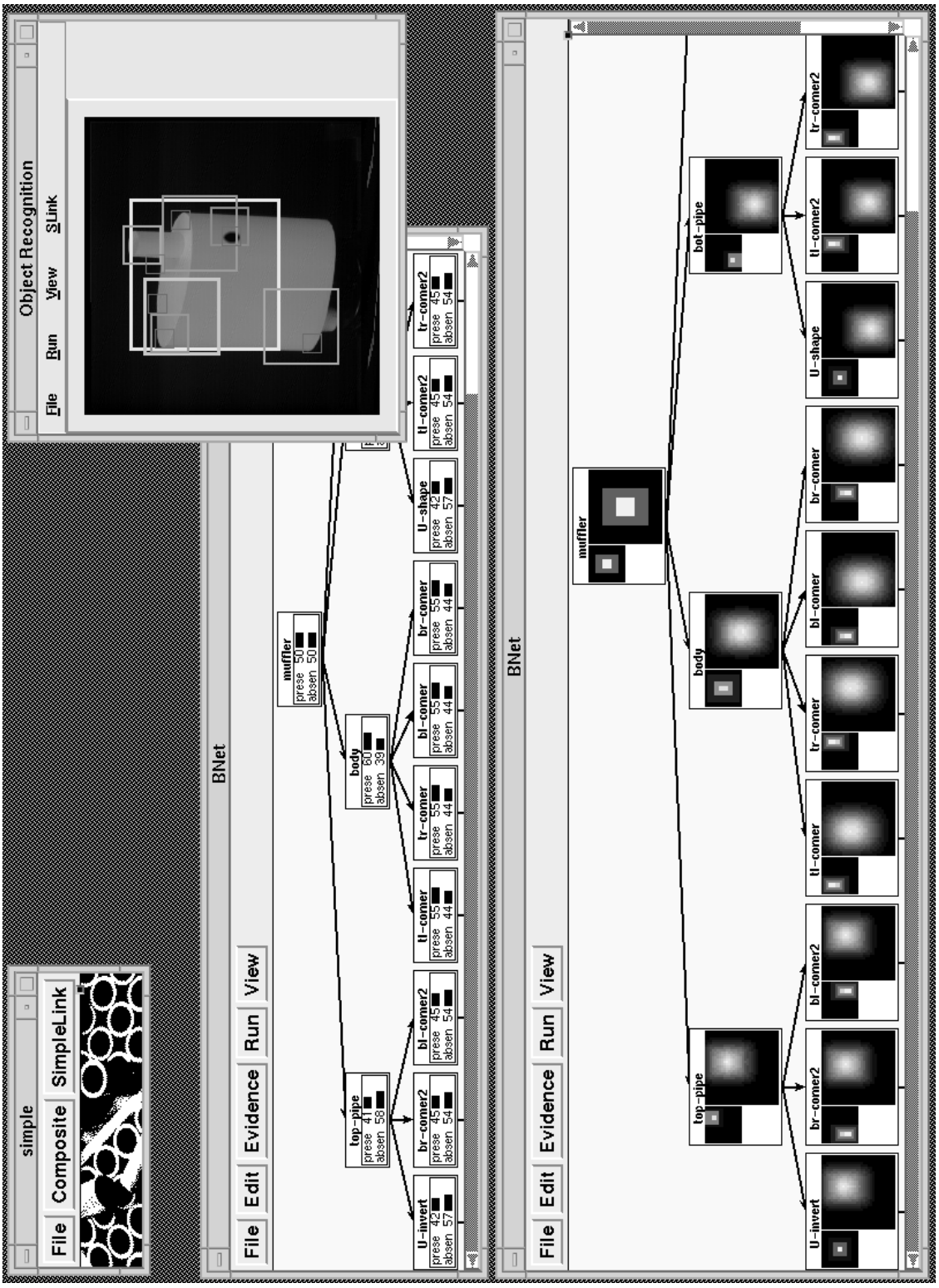

Figure 9: The initial setup of the entire system. Upper right corner shows the results of the saliency detection algorithm [5]. Squares in the figure identify the most salient regions (SSCs) at four octave scales with wider and brighter lines corresponding to greater saliency. The size of the square corresponds to the particular scale.

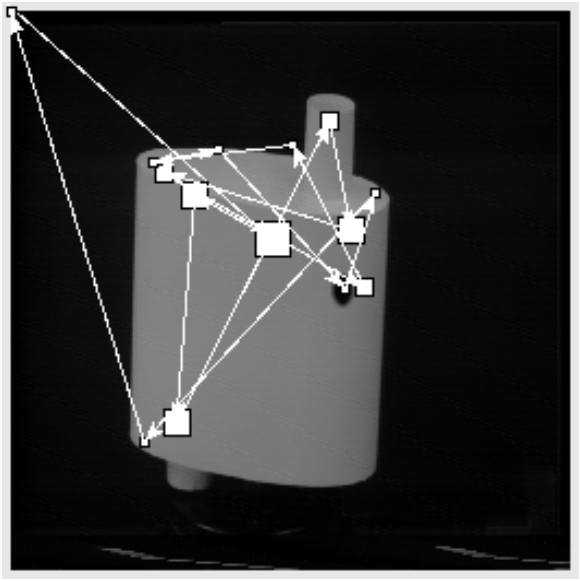

Figure 10: The shifts of focal attention during the recognition as guided by the bottom–up influences. The sequence starts and ends in the upper left corner. The squares are positioned at the centers of the SSCs that are focused upon. The size of each square corresponds to the SSC's scale. Compare this figure to the figure in the upper right corner of Figure 9.

a particular location (geometric constraints built in the EA net).

Figures 11 and 12 show how the beliefs about the muffler presence and its quality change as the new evidence get added. As can be seen, evidence is added only for four icons (**tl-corner, U-inverted, flaw**, and **br-corner2**). The remaining nine salient icons either do not meet the criteria established in Section 4.1 or the corresponding nodes already have the new evidence from a previous fixation. The decision analysis module might avoid visiting these locations at all.

It can be seen in Figures 11b and 12d that finding the flaw does not affect significantly the *scrap-off* state of the node **muffler-quality**. The reason for this is that the flaw's EA node has distribution spread over the entire muffler, and when the system finds the flaw, the value of $\lambda(present)$ in Eq. (5) is relatively low. This further causes relatively high belief in the *notPresentOrOther* state of the **flaw-type** node of the TASK network.

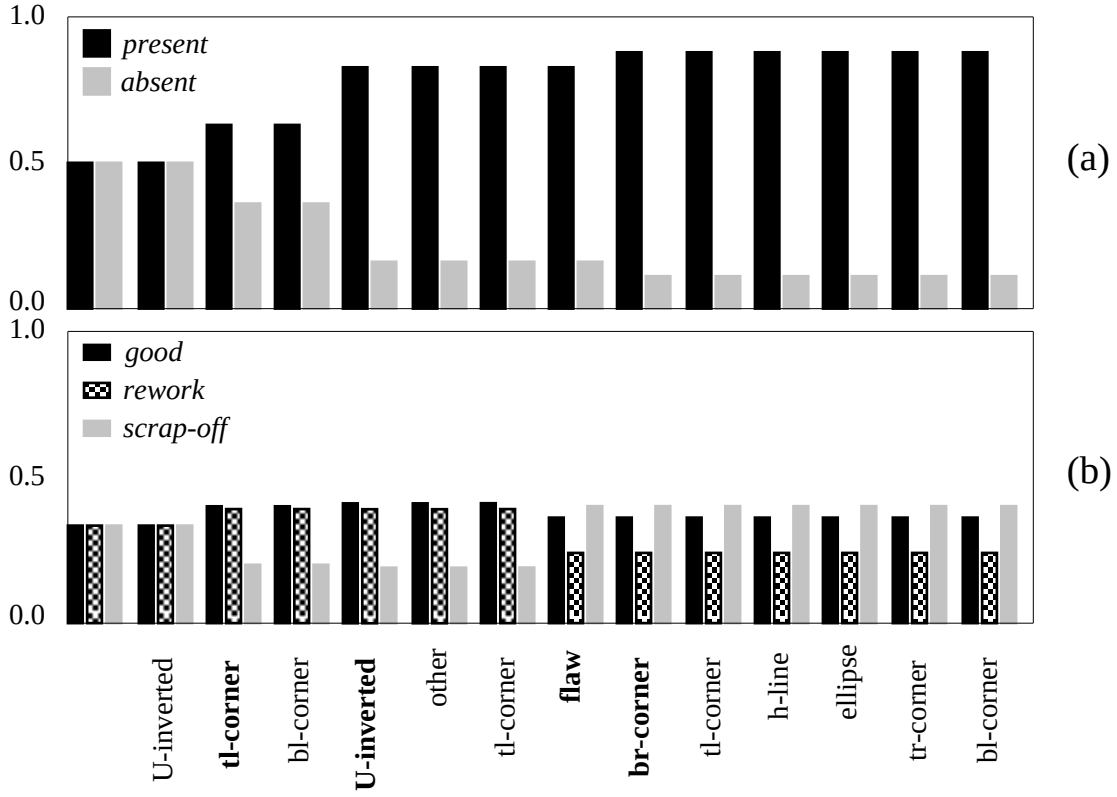

Figure 11: The diagram of the beliefs in the **muffler** node of the PART–OF network (a) and beliefs in the **muffler-quality** node of the TASK network (b) as the new evidence gets added. The names of the four icons that add evidence are shown in bold face.

One can observe that knowing *where* is (a piece of) *what* is not useful as long as it does not fit into any of known knowledge "frames". Knowing *where* an object (*what*) is located may be useless if it cannot be put in the context of the existing system's knowledge and expectations that arise from it. If the system encounters a novel situation, it cannot use gathered evidence in solving the current task. In a way, the "impedances" of both ends (evidence and expectations) should be matched. The evidence may be only stored and used later when the system encounters the same situation.

The finding that only few pieces of evidence get "framed" at the first trial should not be surprising. The shifts in Figure 10 are expected to take place in the first phase of object recognition which would be mainly saliency–driven. In the next phase, which is

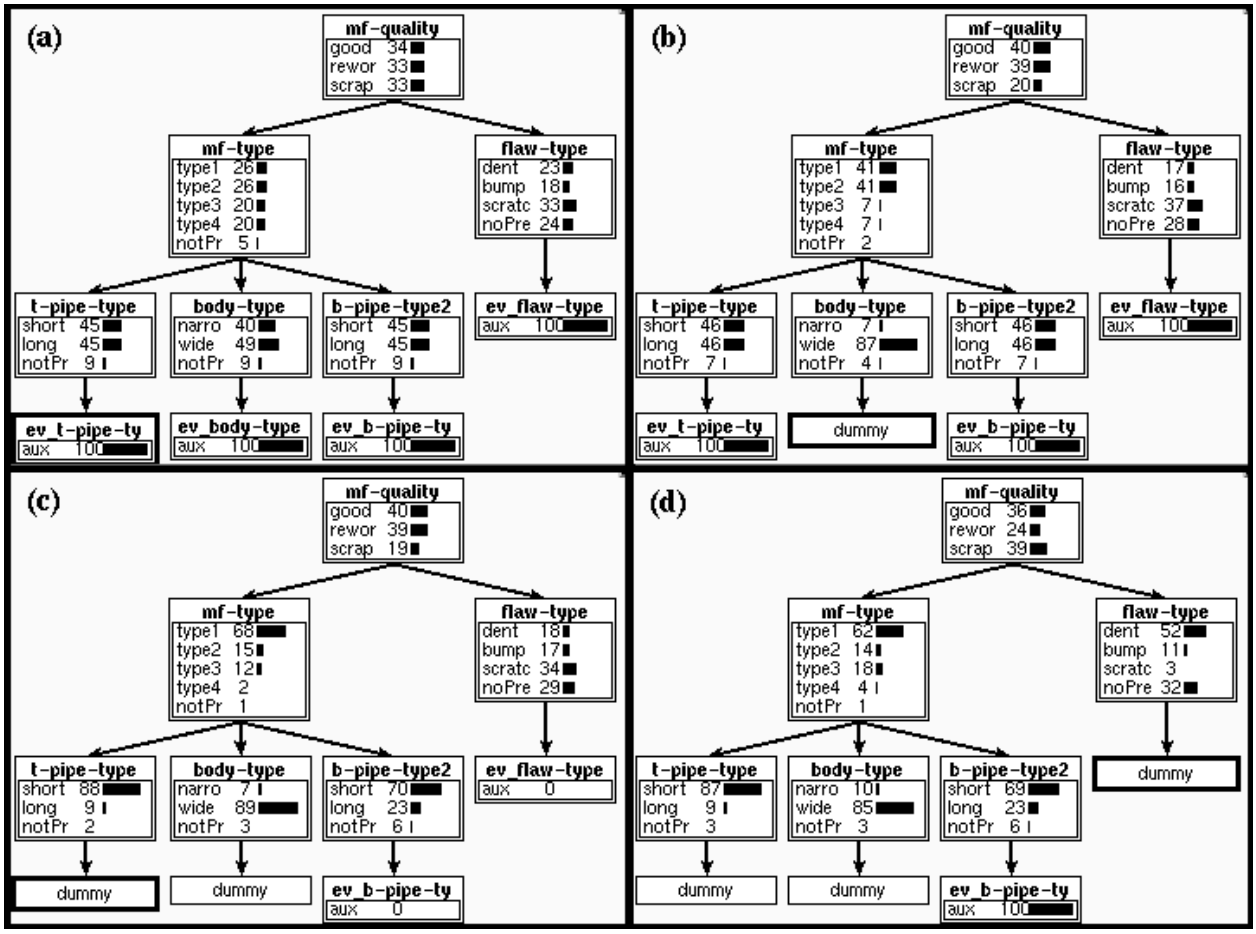

Figure 12: The evolution of the TASK network as the packages are propagated from the "knowledge networks". The initial state is shown in (a). The terminal nodes turn into the *dummy* type as the evidence gets added to them [6].

mainly knowledge–driven, some of these points might be visited several times in order to re–examine them, the same way as it happens with human eye movements [12]. Only after several trials the system would succeed to fit into some of known knowledge "frames" those pieces of evidence that remained unfit after the previous visits. Also, additional points might be visited based on the prior knowledge about the problem.

6 Conclusions

The essence of our model of selective perception is in proposing the scale–space cells, whose icons make the "quanta" of recognition. The model selects the scale which contains the *minimum* amount of information about an image object. This scale is approximately one octave below the scale at which an object becomes smoothed out and thus indistinguishable from other objects of the same (retinal) size. The computation of the icons is performed in parallel, whereas they are used serially in the recognition process.

The structure of the scale–space provides for transform–invariant object representation and recognition. It also reduces the amount of information that needs to be stored in long–term memory, and provides for easier reasoning about spatial relations.

Knowledge about the world and about the specific task to be solved is represented using multiple Bayesian networks. This way the system can tolerate unreliable evidence due to the physics of image creation and due to the mistakes in image interpretation.

The examples show that the influence of the findings on beliefs in root nodes is relatively modest. This means that the system will have to repeat fixations, this time with more focused EA distributions to make sure the previous findings were correct. This is in accordance with the observation that humans need many repeated fixations of various details when examining complex objects [12].

The model provides an observer with the means for comprehending successive views, and allows the observer to grasp the structure of a visual world that is never visually present all at once, but "exists" beyond the boundaries of each successive view. It provides expectations about the probable layout of the next attended area, as well as providing a spatially organized

storage system for integrating information from successive views.

References

[1] CALIFANO, A., KJELDSEN, R., and R.M. BOLLE, "Data and Model Driven Foveation," *Proc. 10th IEEE Int. Conf. Pattern Recognition, Vol.1*, Atlantic City, NJ, pp.1–7, June 1990.

[2] JENSEN, F., CHRISTENSEN, H.I., and J. NIELSEN, "Bayesian Methods for Interpretation and Control in Multiagent Vision Systems," *Proc. SPIE, Vol.1708: Applications of Artificial Intelligence X: Machine Vision and Robotics*, K.W. Bowyer (ed.), pp.536–548, 1992.

[3] LEVITT, T.S., BINFORD, T.O., and G.J. ETTINGER, "Utility–Based Control for Computer Vision," in *Uncertainty in Artificial Intelligence 4*, R.D. Shacter, T.S. Levitt, L.N. Kanal, and J.F. Lemmer (eds.), pp.407–422, North Holland, 1990.

[4] MARSIC, I., A Simple Bayesian Network Simulator, CAIP–TR–178, Center for Computer Aids for Industrial Productivity, Rutgers University, March 1994.

[5] MARSIC, I., "Data–Driven Shifts of Attention in Wavelet Scale Space," submitted.

[6] PEARL, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publ., Inc., San Mateo, CA, 1988.

[7] RIMEY, R.D., and C.M. BROWN, "Where to Look Next Using a Bayes Net: Incorporating Geometric Relations," *Proc. 2nd European Conf. Comp. Vision—ECCV'92*, G. Sandini (ed.), Springer–Verlag, pp.542–550, 1992.

[8] RIMEY, R.D., and C.M. BROWN, "Control of Selective Perception Using Bayes Nets and Decision Theory," *Int. J. Comp. Vision*, Vol.12, No.2/3, pp.173–207, April 1994.

[9] SIMARD, P.Y., LECUN, Y., and J. DENKER, "Efficient Pattern Recognition Using a New Transformation Distance," *Neural Information Processing Systems 5*, S. Hanson, J. Cowan, and L. Giles (eds.), Morgan Kauffman Publ., 1993.

[10] SIMONCELLI, E.P., FREEMAN, W.T., ADELSON, E.H., and D.J. HEEGER, "Shiftable Multiscale Transforms," *IEEE Trans. Inform. Theory*, Vol.IT-38, No.2, pp.587–607, March 1992.

[11] SUCAR, L.E., and D.F. GILLIES, "Probabilistic Reasoning in High–Level Vision," *Image Vision Comp.*, Vol.12, No.1, pp.42–60, January/February 1994.

[12] YARBUS, A., *Eye Movements and Vision*, Plenum Press, New York, 1967.