The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Semantic Evaluation of Region of Interest for Intelligent Robot

Md. Rokunuzzaman, K. Sekiyama, and T. Fukuda

***Abstract*—This paper introduces the concept of semantic evaluation of Region of Interest (ROI) for intelligent robots. The intelligent robot must have the capability of understanding situations. The first step of understanding of the situation is to find where to focus on and how to behave. Focusing on some particular area or region needs selection of the objects of interaction relevant to the context. Moreover, the focused area needs to be semantically evaluated to quantify the semantic relations. In this paper, we first detect interacting objects based on dynamic interaction. Then we recognize probable objects using Dynamic Bayesian Networks. Using the probable objects and a mutual supplementation model, we determine the contextual object. We form ROIs based on possible combinations of objects and the contextual object. Finally, we semantically evaluate each ROI. Various experimental results are provided to illustrate our method.**

I. Introduction

Robots are considered intelligent agents in case they perform various tasks in various situations. However, in current systems robots interact with objects physically [1], [2], [3] rather than understanding situations. The latter task is very difficult as it is based on psychology or behavior of object-object relations. Behavior is subjective and closely related to research areas of [4], [5] and [6]. On the other hand, context recognition is related to probability-based modeling [7], [8], ontology [9], text [10], semantics [11], [12] and its evaluations [13], [14].

By the term "semantic evaluation of ROI", we mean the quantification of the significance of a ROI. This is necessary for selecting the appropriate region that contains significant information for processing. A region contains different objects as entities. These entities interact with each other either statically or dynamically. Dynamic interaction attracts human attention. Therefore, we choose this for important object detection. However, due to our subjective preference, we also sometimes feel interest in static relations or static-dynamic relations of observed objects. To solve this problem, we need to determine our cognitive boundary and evaluate these relations in this boundary. The cognitive boundary concept is close to the frame concept in Artificial Intelligence. The origin of the frame concept is the making of films and in particular the camera frame. Here the frame problem is that the film director must control the camera such that the viewer can understand the acting and no unnecessary information is displayed. The following example scenes can illustrate the emergence of the cognitive boundary problems:

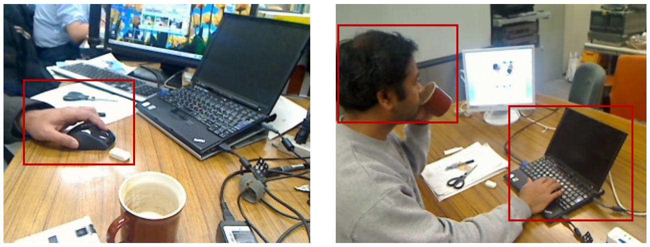


Fig.1 Illustration of cognitive boundary problems, where ROIs show (a) operating a PC mouse, (b) two events, drinking coffee and operating a PC

In Fig.1, the red rectangles are drawn arbitrarily to locate region of interest in the scenes. However, how can be sure that the drawn ROIs are appropriate? The issues of selecting the appropriate ROI can be stated as follows:

1. ROI should contain most important information
2. ROI should exclude unnecessary information
3. ROI should include semantically evaluated context
4. ROI should be compact to lower down cognitive loads

The ROI shown in Fig.1 (a), contains two more objects which are irrelevant (e.g. eraser and scissor in relation to PC mouse) and are as considered unnecessary information. Moreover, the PC is missing as semantic context, because PC mouse is functionally relevant to PC.

In Fig. 1(b), there are two ROIs detecting different events. However, since processing both ROIs simultaneously increases cognitive load, the cognitive boundaries need to be selected first. By the term "cognitive boundary", we mean the region where objects are connected semantically. In this regard, we propose an algorithm that takes care of these difficult issues and explain it in detail in this paper.

In traditional approaches, object detection and context recognition are done separately. Moreover, meaning is represented in a more linguistic form than its relation to visual objects is. Current trends of ontology give only one to one relations for each object, which is insufficient for generality in recognition, perception or inference. Furthermore, semantic evaluation of the ROI is not investigated in relation to object recognition. To solve this problem, we propose a new methodology for semantic evaluation of the ROI, which is interactively bonded with probabilistic object recognition

Md. Rokunuzzaman is currently a doctoral student in the Department of Micro-Nano Systems Engineering at Nagoya University, Japan (corresponding author to provide phone: 052-789-4481; fax: 052-789-3115; e-mail: rzaman@ robo.mein.nagoya-u.ac.jp).

K. Sekiyama is currently as an Associate Professor in the Department of Micro-Nano Systems Engineering at Nagoya University, Japan (e-mail: sekiyama@mein.nagoya-u.ac.jp).

T. Fukuda is the Professor in the Department of Micro-Nano Systems Engineering at Nagoya University, Japan (e-mail: fukuda@mein.nagoya-u.ac.jp).

978-1-4244-6676-4/10/$25.00 ©2010 IEEE

and multiple object-verb-object relations.

The rest of the paper is organized as follows. Section 2 gives an overview of our method. Section 3 presents experimental results. We evaluate our method in section 4 and conclude in Section 5.

II. Semantic Evaluation of ROI

A. Overview of the method

Figure 2 shows the overall process for semantic evaluation of the ROI.

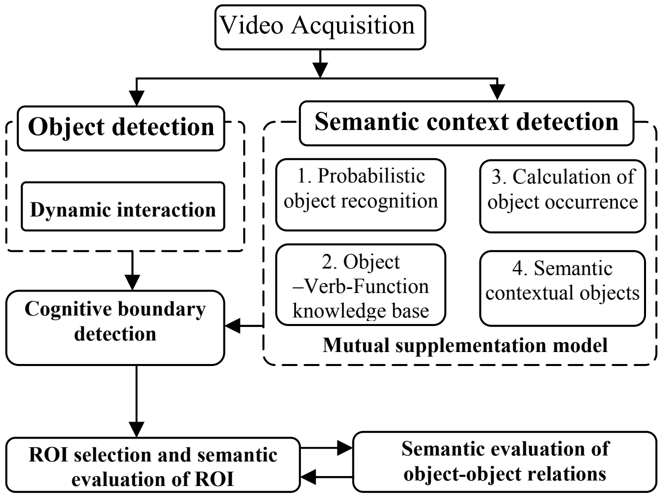

Video Acquisition
Object detection
Dynamic interaction
Semantic context detection
1. Probabilistic object recognition
3. Calculation of object occurrence
2. Object –Verb-Function knowledge base
4. Semantic contextual objects
Mutual supplementation model
Cognitive boundary detection
ROI selection and semantic evaluation of ROI
Semantic evaluation of object-object relations

Fig.2 Overall process for semantic evaluation of the ROI

In our proposed method, we first determine objects that are interesting to observe by dynamic interaction. Then we determine the semantic context by counting object's occurrence in object set based on mutual supplementation of object-verb-object relations. In this method, semantic context and objects from dynamic interaction are bonded together to form the cognitive boundary. We select the ROI and then evaluate each object-object relation semantically, and update our choice. We explain the detailed process in the following sections.

B. System configuration

We capture several videos of different contexts and save them into memory of a Personal Computer (PC). The video frames are further processed using various image-processing programs developed with Intel's Open Computer Vision (OpenCV) Library. The codes are compiled by Visual C ++ shipped with Microsoft Visual Studio 2005. Table I shows the system configuration that is used to implement our method.

TABLE I
System Configuration

No	Items	Specifications
1	Vision Processor	Intel Core2 Duo, 2,20 GHz, 2.0 GB of RAM
2	Vision Sensor	Canon PTZ Camera Model: VC50i
3	Development Platform	Microsoft Visual Studio 2005
4	Programming	C++, Visual C++
5	Code development	Intel's OpenCV Library

C. Cognitive boundary detection

Cognitive boundary is the region that confines important objects in interaction. Cognitive boundary formation needs to detect the following:

1) Object detection: The aim of object detection is to determine which objects are interesting to observe. Interacting objects tend to capture human attention. Therefore, we consider dynamic interaction for object detection. To determine dynamic interaction we need to find the following:

1.1) Motion saliency: Motion is an important cue for dynamic interaction. Using a blob filter, we detect the objects as blobs. Then, we determine the motion M of the blobs by associating blobs between frames and their euclidean distance as

$$M = \sqrt{(dx_n^{n+1}(i))^2 + (dy_n^{n+1}(i))^2} \quad (1)$$

Where, $dx_n^{n+1}(i)$ and $dy_n^{n+1}(i)$ are the center to center distances of the ith object from n to $n+1$ frame in x and y-coordinates respectively. Now we need to determine motion saliency that denotes the conspicuous state of an object in a video. In our method, motion saliency is expressed as a value namely that difference between motion of each object and the minimum among all the objects at that point of time. If M_i is the motion of the ith object and $M_{\min}$ = min $\{M_i \ldots.. M_n\}$, where, i = 1, 2,....n are the number of objects in the frame at that instant, then motion saliency value can be expressed as

$$M_{sv}(i) = (M_i - M_{\min}) \quad (2)$$

To obtain the value as a factor ranging from 0 to 1, we normalize it with its maximum value as

$$M'_{sv}(i) = M_{sv}(i) / \max(M_{sv}(i)) \quad (3)$$

Insignificant motion saliency value can cause the system to be irresponsive to interaction. Therefore, we need to set a weight for this. This weight can be pixel information of the object in motion. This is because; one of the aspects of human vision system is that it attains objects with larger area as it covers most of the portion of the retina. Based on this concept, we introduce the term information density for this weight that is defined as

$$I_D = A_O / A_{OBB} \quad , 0 \le I_D \le 1 \quad (4)$$

Where, A_O is the area measured in number of pixels inside the object and A_{OBB} is the rectangular area of the box that fits the periphery of the object. Hence, the weighted motion saliency value is

$$M''_{sv}(i) = M'_{sv}(i) * I_D \quad (5)$$

1.2) Proximity: Interaction between objects depends on proximity based on their relative distance. This distance is a measure of relevancy as well as proximity for interaction. If we denote relative distance D_R then it is simply the Euclidean distance of the surrounding objects from the most salient object and can be formulated as

$$D_R = \sqrt{(Cx_{MaxSalObj} - Cx_i)^2 + (Cy_{MaxSalObj} - Cy_i)^2} \quad (6)$$

where, $Cx_{MaxSalObj}$ is the center x-coordinate of the most salient object and Cx_i is the ith object's center x-coordinate except most salient object. Analogously the second term applies to y-coordinates.

1.3) Dynamic interaction factor: Based on the psychological behavior we devise an interaction detector which we name it "dynamic interaction factor". The value of this factor will determine how much the object is interacting. We define it as a ratio of the weighted motion saliency value of an object to its proximity to most motion salient object. If we denote Interaction Factor as *DIF*, then it is expressed as

$$DIF = M''_{sv}(i) / D_R \quad (7)$$

where $M''_{sv}(i)$ and D_R are defined in Eq. (5) and Eq. (6) respectively.

2) Semantic context determination: The aim of the semantic context determination is to find the context in which objects are significant. The following steps are needed to do this:

2.1) Mutual supplementation model: Since the object recognition task is very complicated and time consuming, we propose a more flexible way to solve it. We consider a Dynamic Bayesian Network (DBN) in a mutual supplementation framework as shown in Fig.3.

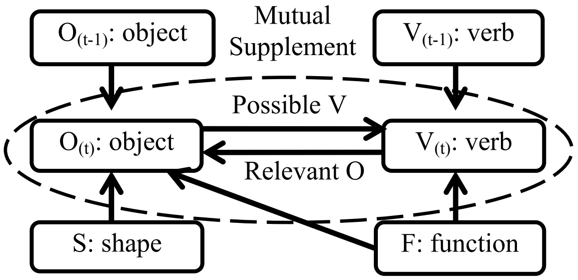


Fig.3 Object-Verb-Function Mutual Supplementation Model

For our model, we have the following similar expressions for probabilities of Object *O* and verb *V* at current frame *t* as

$$P(O^t \mid S,F,O^{t-1},V^{t-1}) = \alpha P(O^{t-1})P(O^t \mid S)P(O^t \mid V^{t-1},F)P(V^{t-1}) \quad (8)$$

$$P(V^t \mid S,F,O^{t-1},V^{t-1}) = \alpha P(O^{t-1})P(V^t \mid O^{t-1},F)P(V^{t-1}) \quad (9)$$

where α is a normalization constant, *S* and *F* represent shape and function corresponding to object as prior information and O^{t-1} and V^{t-1} represent object and verb at frame *t-1* respectively.

For probabilistic object recognition our model needs the following steps:

2.1.1) Prior knowledge about function of objects: Each object can have different functions. Based on the object-set we find several functions that we categorize into 10 classes. The classes are as follows:

i) ***Generation***: the function which generates some entity(object) or event

ii) ***Identification***: this function makes some mark for entity detection

*iii)****Separation***: this function makes entity division or act of disintegration

iv) ***Get in***: this function acts as a possession of entity by entailment

v) ***Take away with***: this function moves the entity or has conveyance ability

vi) ***Give positive feeling***: this function makes the entity which supports its existence

vii) ***Get information***: this function enables the entity to sense information

viii) ***Control***: this function make the entity to perform some actions on another entity

ix) ***Transform***: this function changes the state of the entity

x) ***Put-together***: this function helps to integrate the entities as opposed to function 3

2.1.2) Prior knowledge about object shape: We consider 10 objects as object set of a typical computer lab of our laboratory to test our algorithm. For creating a database of the object set, we choose shape as a function of two invariant properties: aspect ratio (i.e. object width divided by object height) and circularity (i.e. 2π*Area/perimeter^2). Then we map these two parameters for each object and store them in a database.

2.1.3) Object-Verb-Function knowledge base: We propose a new approach for building a knowledge base for the objects based on mutual supplementation of Object-Verb-Function dependence. A sample of each matrix is shown in Table II.

This approach gives a more generalized view both on recognition and context determination by relating multiple functionalities of the object. We make several object-verb-function matrices, which describe our objects based on their verbs and functions. Relevant objects are associated to verbs according to the usualness or frequency of use in daily life. Then we grouped similar verbs in a function

TABLE II
OBJECT-VERB-FUNCTION MATRICES

Functions	Verbs	Relevant Objects
Generation	Write	Pen , Paper , PC , Keyboard, Scissor , Monitor
	Communication	Cell phone , PC , Pen , paper, Keyboard , Monitor
Identification	Indite	Pen , Scissor
	Mark	Pen , Scissor
Separation	Cut	Scissor, Pen
	Wipe	Eraser, Paper
Get in	Enclose	Cup , Paper , Pen
	Wrap	Paper
Take away with	Convey	paper , power cable , cell phone ,Cup , PC , Monitor
	Browse	PC , Cell phone , Monitor
Give positive feeling	Support	Power cable , PC , Pen , Keyboard
	Correct	Eraser , Pen , PC , Paper
Get Information	Display	Monitor , PC , Cell phone , paper
	Present	Monitor , PC , Cell phone , paper , Pen
Control	Control	PC , Monitor , Keyboard , Power cable , Pen
	Check	Monitor , PC , Cell phone
Transform	Cook	Cup
	Prepare	PC , keyboard , Monitor , Paper ,pen , Cup
Put together	Adhere	Pen , Eraser
	Fix	Pen , Scissor , Eraser, PC , Keyboard , Monitor

class with the help of cognitive synonyms (synsets) from Word Net (http://wordnet.princeton.edu).

2.2) Probabilistic object recognition: After observing the object, the shape is determined by computing the aspect ratio and circularity of the object. Then we calculate the relative probability of the object $P^{(t)}(S \mid O)$ by comparing to the aspect ratio-circularity map (database) based Euclidean distance.

The probability of relevant function of a given object $P^{(t)}(F = f_{rel} \mid O)$ can be calculated by observing function as categorized in section 2.1.1 and associated verbs mentioned in Table II.

Therefore, using mutual supplementation model, the probability of the object class is recognized by

$$P^{(t)}(O \mid F, S) = \alpha P^{(t)}(F = f_{rel} \mid O) P^{(t)}(S \mid O) P^{(t-1)}(O) \quad (10)$$

2.3) Calculation of object's occurrence in object set:

In perspective of vision, visual words can be considered as words. We relate word's occurrences in text to object's occurrence in relevant object's set. This can be formulated as:

Given a set O of n objects in an image, I and a set V of m possible verbs corresponding to n objects can be expressed mathematically as

$$O = \{o_1, o_2, o_3, ...o_n\} \quad (11)$$

$$V = \{v_1, v_2, v_3, ...v_m\} \quad (12)$$

Let denote $Q(V)$ and $Q(O)$ as a power sets of all possible verbs and relevant objects, then by mutual supplementation, we have the following relations expressed by

$$v : O \rightarrow Q(V) \quad (13)$$

$$r : V \rightarrow Q(O) \quad (14)$$

This implies that verb v is mutually correspondent function of object that maps to $Q(V)$.

Denoting $V_{assoc.}$ as a set of associated verbs for a set of recognized object $O_{recog.}$ such that for each element $rec \in O_{recog.}$ Using set theory, $V_{assoc.}$ can be expressed by

$$V_{assoc.} = \{X \in Q(V) \mid \exists rec \in O_{recog.} : v(rec) = X\} \quad (15)$$

We can define a relevant object corresponding to X as

$$\forall_{o \in O};\quad rel_o(X) = \left\{ \begin{matrix} 1 & if\; o \in r(X) \\ 0 & otherwise \end{matrix} \right\} \quad (16)$$

The object occurrence $C(o)$ in the relevant object set can be found as

$$C(o) = \sum_{X_i \in V_{assoc.}} \sum_{x \in X_i} rel_o(X) \quad (17)$$

2.4) Semantic contextual object:

Semantic contextual object is the relevant object corresponds to maximum number of occurrence. Assuming i is the index of each image in a video, the contextual object θ_i can be determined by

$$\theta_i = \arg\max_{o \in O} C(o) \quad (18)$$

D. ROI selection

In our method, we define ROI as a cognitive boundary that confines the interacting objects with a contextual object. Assuming interacting object $z_{\text{int}} \in Z$ is the relevant object detected by interaction. Therefore, ROI at a given context can be written by

$$ROI_{\theta_i} = \{Z \in_i Q(O) \mid \forall z_{\text{int}} \in Z : \exists x \in v(\theta_i) : z_{\text{int}} \in r(x)\} \quad (19)$$

E. Semantic evaluation of object-object relations

We obtain possible ROIs from eq.(19). In order to select an appropriate ROI for a given semantic context, we need to evaluate each ROI. For this purpose, we propose an information theoretical approach for functional similarity measurement as an evaluation of semantic relations between objects in the ROI.

Let $S_E(O_1, ...O_n)$ denote semantic evaluation among n objects then the functional similarity with regard to contextual object θ_i in mutual supplementation can be expressed as

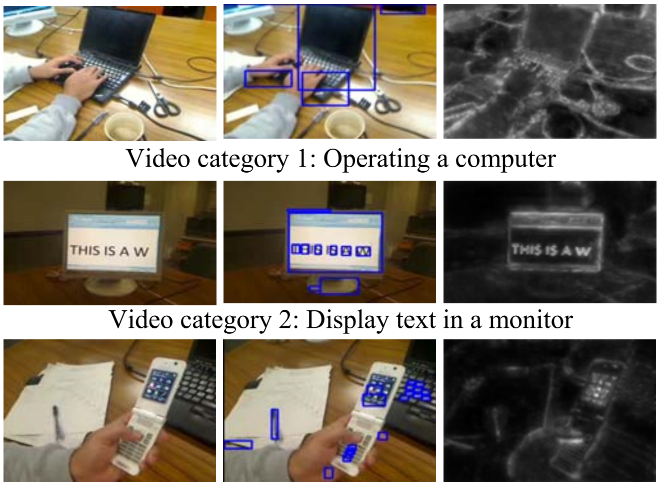
$$S_E(O_1, ...O_n) = \left(\frac{2 \times \log P(F_1 \bigcap F_n)}{\sum_{i=1}^{n} \log P(F_i)} \right)_{\theta_i} \quad (20)$$

where F_i is the function implied by O_i. The numerator of the equation represents commonality whereas the denominator represents the individual description of the objects based on object functions for a given contextual region θ_i.

III. Experimental Results

A. Object detection based on dynamic interaction and comparison with saliency maps

We take several videos in our computer laboratory. The average length of each video is 1 min 15 sec taken at a resolution of 320×240 @1024kbps. We check the validity of our proposed object detection with existing saliency algorithms [15] as shown in Fig. 4.



Video category 1: Operating a computer

Video category 2: Display text in a monitor

Video category 3: Browsing in a cell phone

Fig.4 Interactive object detection with saliency maps

In Figure 4 shows three categories of dynamic interaction scenarios. The first image of each category is the original image, second image illustrates the object detection and last one is the corresponding saliency map. The blue rectangles in second images define boundaries where dynamic interaction occurs. These boundaries are potential regions containing objects of interest. From the saliency map, we can observe that the saliency value is high at the objects of interaction.

In order to choose appropriate regions we need to evaluate each region. The next results give a meta-data calculation needed for semantic evaluation of ROIs.

B. *Contextual object detection*

Based on probabilistic object recognition, we determine the semantic context using our method. Table III summarizes the results for the three video categories as follows:

TABLE III
PROBABILISTIC OBJECT-CONTEXT RECOGNITION

Video Cat.	Possible recog. of objects/Event	Context recog.	Ground truth	Recog. Acc. (%)	Avg. Acc. (%)
1	Keyboard (KB), Power cable, Monitor, scissor	PC	PC	100	
2	Monitor, Generation (Event)	PC, Monitor	Monitor (display)	94	93.33
3	Pen, Paper, Cell phone, Generation (Event)	Pen PC	PC (browse)	86	

The accuracies are calculated based on functional relevance as context recognition is based on it.

C. *Semantic evaluation of ROI*

We evaluate the ROI semantically by applying the functional similarity measure with regard to the contextual object in mutual supplementation.

Table IV shows the similarity check metric with regard to the contextual object for video seq.1 of video category 1.

TABLE IV
SIMILARITY CHECK METRIC WITH REGARD TO CONTEXTUAL OBJECT "PC"

Verbs	PC	KB	Power cable	Monitor	Scissor
Write	○	○	×	○	○
Communicate	○	○	×	○	×
Convey	○	×	○	○	×
Browse	○	×	×	○	×
Support	○	○	○	×	×
Correct	○	×	×	×	×
Display	○	×	×	○	×
Present	○	×	×	×	×
Control	○	○	○	○	×
Check	○	×	×	○	×
Prepare	○	○	○	○	×
Fix	○	○	×	○	○

In video sequence 1 of video category 1, the contextual object is *PC*. For this object, we find the associated verbs from Table II. Then we compare the possible recognized objects with regard to the verbs of contextual objects. The ○ marks indicate similar and × marks indicate non-similar verbs in the table. By counting the marks and using eq. (17) we have for Video seq.1 of video category 1:

$$S_E(PC, KB, Powercable) = \left(\frac{2\times\log(3)}{\log(12)+\log(6)+\log(4)}\right)_{PC} = 0.38$$

With those possible combinations, we can form ROIs and obtain semantic evaluation of each ROI as shown in Table V

TABLE V
SEMANTIC EVALUATIONS OF POSSIBLE ROIS FOR VIDEO CATEGORY-1

Possible ROIs	S_E
PC-KB-Power cable	0.38
PC-KB-Monitor	0.49
PC-KB-Scissor	0.28
PC-Power cable-Monitor	0.36
PC-Monitor-Scissor	0.13
PC-Power cable-Scissor	0.00

This result suggests that objects with relevant functions give higher values of semantic evaluation and should be a good choice for ROI. We are interested in seeing how the semantic evaluation varies with different number of object-object combinations. Fig. 5 depicts this investigation.

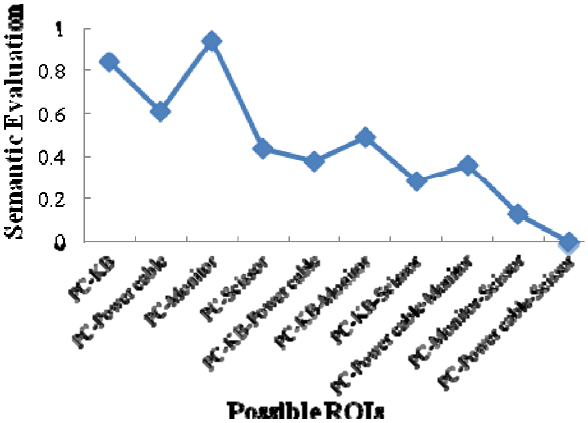


Fig.5 Semantic evaluations of different ROIs

From this figure, it is clear that ROI with less objects have higher semantic value. However, these values vary significantly causing instability. The ROIs with more objects have less variation in semantic value and are more stable.

D. *Object recognition with event observation*

The main advantage of our proposed system is the ability of object recognition by evolution of observation of events. For instance, in video category 3, the probability of being a *monitor* increases when the generation function is observed in video sequences as an event that is detected as the number of rectangles increase on the *monitor*.

E. *Possible inference by mutual supplementation*

Unlike conventional ontology systems, our system can make multiple probabilistic inferences based on object-object relations. With limited data or erroneous information, our system can produce multiple possible inferences so that it can adapt to various contexts or situations. Moreover, our system can predict possible contexts if it gets only partial information.

Table VI shows possible inferences for video category 2 and 3.

TABLE VI
POSSIBLE INFERENCES PRODUCED BY MUTUAL SUPPLEMENTATION

Video Cat.	Possible recognized object	Context	Recognized event	Possible inferences
2	Monitor	PC	Generation	Write, communicate, convey, browse, display, present, control, check, prepare, fix
3	Pen, Paper, Cellphone	PC Pen	Generation	Write, communicate, present, display, correct, prepare

For example, in video category 2 only *monitor* is detected and our system predicts that there is a possibility of having a PC nearby. This method differs from the conventional ontology approach where we have to define that *monitor* is a part of PC in a hierarchical architecture.

IV. PERFORMANCE EVALUATION

Our proposed approach is unique and therefore, we have compared our ROI selection as salient region detection to state of the art algorithms as shown in Table VII.

TABLE VII
COMPARATIVE SALIENT REGION DETECTION EFFICIENCIES

Methods	Implementation	Efficiency
Proposed method	**Real time/C++**	**0.86**
Our method [16]	Real time/C++	0.81
Neuro Vision Tool[17]	Real time/ C++	0.75
Saliency Tool Box	Offline/ Matlab	0.74
Informax[18]	Offline/ Matlab	0.72

The efficiencies are tabulated by calculating the area under ROC (Receiver Operating Curve) after detecting the salient region for each method.

V. CONCLUSION

In our method, the object is recognized by its behavior. When an event occurs, the system calculates the probability of each object from a list of relevant objects. Existing object recognition systems assume that the object should be unique in appearance. However, this type of recognition fails when the actual condition dissatisfies that assumption. As a result, the systems cannot recover from errors or cannot make inferences to have a probabilistic value of recognition. Our method overcomes this problem by increasing the probability of recognition when there is sufficient evidence of the event relevant to each object. Moreover, context is derived from various object-verb-function mutual supplementations without sufficient data for training. We agree that it is very difficult to solve any recognition problem without any training or learning. Thereby, this research is an extension of the learning type systems that dynamically update recognition and adapt to situations. The research contributions to AI area are as follows:

i) We propose a novel, generic and unsupervised method of object-context recognition with mutual supplementation model.

ii) We detect the object-object relations with dynamic interaction and model an event based probabilistic object recognition.

iii) We devise a new structure of an ontology-like knowledge base where each object has multiple relations for making inference and can be aware of the situation.

In future, we will improve our system by integrating learning systems.

REFERENCES

[1] N. G. Muller and A. Klienschmidt, “Dynamic Interaction of Object-and Space-Based Attention in Retinotopic Visual Areas,” *J. Neuroscience*, vol.23, no.30, pp. 9812-9816, 2003

[2] B. Möller and S. Posch, “Analysis of Object Interactions in Dynamic Scenes,” in *Springer Berlin*, Pattern Recognition, vol. 2449, 2002, pp. 361-369

[3] C. W. Lin, “Dynamic Region of Interest Transcoding for Multipoint Video Conferencing,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 13, no. 10, pp. 982–992, 2003

[4] A. Felix, “The psychology of interest (II),” *Psychological Review*, vol. 13, no. 5, pp. 291-315, 2006

[5] R.T. Canolty, “Spatiotemporal dynamics of word processing in the human brain,” *Frontiers in Neuroscience*. vol. 1, issue 1, pp. 185-196. 2007

[6] T. Sevilmis and M. Bastan, “Automatic detection of salient objects and spatial relations in videos for video database system,” *Int. J. Image and Vision Computing*, vol.26, pp. 1384-1396, 2008

[7] Seung-Bin, I, Youn-Suk S., and Sung-Bae C. “Context Modeling with Bayesian Network Ensemble for Recognizing Objects in Uncertain Environments”, FSKD 2006, LNAI 4223, pp. 688 – 691, 2006

[8] N.Rasiwasia,N.Vasconcelos,Holistic Context Modeling using Semantic Co-occurrences, proc. of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,June 2009,pp.1889-1895

[9] L. Jing, L. Zhou, M.K. Ng. and J.Z. Huang, “Ontology-based Distance Measure for Text Clustering,” in *Proc. SIAM International Conference on Data Mining, Text Mining Workshop*, Maryland, 2006.

[10] K. Church, et al, “Word association norms, Mutual information and Lexicography,” *Computational Linguistics*, vol.16, no. 1, pp.22-29, Mar. 1990

[11] G. M. Ana et al, “Algorithmic Detection of Semantic Similarity,” in *Proc.14th international conference on World Wide Web , Semantic querying* , 2005, pp. 107 – 116

[12] A. Sosei, “Integrated Ambiguity Analysis Model: Detection, Representation and Optimal Meaning Selection”, *SKY journal of linguistics*, no. 20, pp. 35-79, 2007

[13] E. Khatchatourian, “Analysis of Discourse Markers: between the Semantic Stability and the Contextual Variability”, in *Proc. workshop on Formal and Computational Approaches to Discourse and Other Particles*, 2005

[14] P. McCarthy, R. Guess, and D. McNamara, “The components of paraphrase evaluations,” *Behavior Research Methods*, vol. 41 no.3, pp.682-690, 2009

[15] N. Butko, L. Zhang et al, “Visual Saliency Model for Robot Cameras,” in *Proc.of international conference of Robotics and Automation*, pp.2398-2403, 2008

[16] M. Rokunuzzaman, K. Sekiyama, and T. Fukuda, “Automatic ROI detection and Evaluation in Video Sequences based on Human Interest,” *Journal of Robotics and Mechatronics*, vol. 22 no.1, pp.65-75, 2010

[17] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20 no.11, pp.1254-1259, 1998

[18] N. Bruce, J. K. Tsotsos, “Saliency Based on Information Maximization,” *Neural Information Processing Systems*, pp. 155–162, 2005