Toward Object Selection with a Pointer

Henry Hexmoor* and Francisco Lopez Valverde*†
*Department of Computer Science
and
Center for Cognitive Science
State University of New York at Buffalo
Buffalo, NY 14620
†Department of Computer Science
University of Malaga
Malaga 29071, Spain

Abstract

Using a pointer as a means of disambiguating objects in the world is discussed. We offer a visual object selection algorithm and a technique for computing pointer effectiveness. We report on relevant assumptions and directions for a research path.

1. Introduction

A deixis theory is a theory for selecting objects of interest spatially and temporally. This is very useful in attention and short term memory management. Sensing and perceptual mechanisms such as vision and tactile sensing can be guided by deictic strategies. In the game of Pengi [Agre and Chapman 1987], a Penguin is chased by several bees. It must flee to prevent being stung and it and can kill the bees by pushing one ice-block against another and squashing a bee between two ice-blocks. In this game, Pengi selects and maintains visual attention on bees deictically.

Similar to sensing and perceptual mechanism, motor actions can also be guided deictically. For example, two robot arms cooperating on picking up a large object two handedly, continually use each arm movement as deictic clues (reference point in a feedback system) for parameters of movement by the other arm. Language can provide deictic strategies for both sensing and acting. In the following passage "There were three balloons. One of them is caught in the tree. That one is red. The other two are blue." The word *that* is a deictic word and refers to the red balloon caught in the tree.

So far, we identified deictic strategies in modalities of sensing, acting, and language. More complex deictic strategies may involve multiple modalities or use of external deives. External deictic devices can be used in deictic strategies to ease selection. For example, in the game of dart, the concentric markings of the target play as an external deictic devices to guide visual attention. This is true in most video-arcade games.

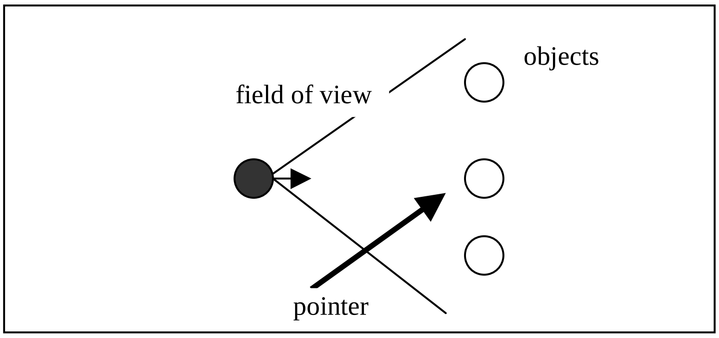
As an example of a complex multi-modal deictic strategy, consider you are holding one end of a box and there is a line drawn on the ground and the end you are holding is closer to the line than he other end. With a head-nod you say to a friend, "Help me line up that end of the box with the line". The head-nod is used in a visual deictic strategy and as the friend starts moving the other end of the box at the same time that you move your side, the line on the ground and your movements sets up the deictic strategy for the friend's moving parameters.

We have been developing intelligent agents and general architectures that provide natural language interfaces, symbol grounding, and concept embodiment as well as adaptation and learning, [Hexmoor et al 1993, Lammens et al 1995, Hexmoor and kortenkamp 1995, Hexmoor 1995, Hexmoor 1996, Hexmoor and Shapiro 1997]. Our agents have been designed to maintain attention on perceptual deictic centers and be commanded by a human supervisor. Commands are issued in conjunction of typing natural language commands on the keyboard as well as external gestures in the world. This style of interaction has been called Teleassistance [Pook 1995]. Our robotic agents go beyond teleassistance in that in absence of human commands, they are fully autonomous. They can communicate with other robots. Within their intended environment and tasks these robots are self-sufficient. In this proposal we focus on teleassistance capability of our robots. Our work has differed from other work in that we have always worked with real systems and never had to simulate perception in order to demonstrate deixis. Others have taken denotation in a symbolic system to mean reference and never bother to build a system that actually picks out an object in the world.

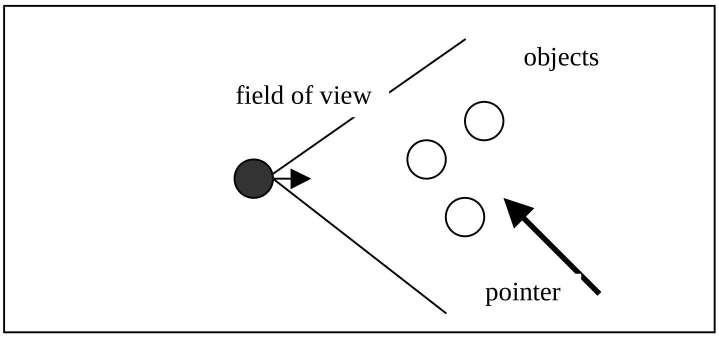
Deictic strategies can include sensing, acting, and language modalities. They may make use of external devices, or a combination of modalities and devices. Deictic strategies can be spatial or temporal. Gestures used for deictic reference may include a wide range of eye-movement, facial gestures, body gestures, external devices, and linguistic indexicals. We limit the scope of this paper to a subset of spacial deictic strategies with gestures using external man-made devices and language.

2. Basics in commands containing indexical noun phrases

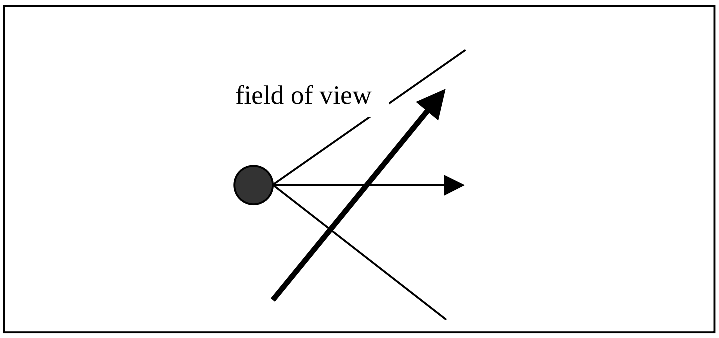
Indexical noun phrases (e.g., that red ball, this apple, that road ahead) narrow the sensory (in our case visual) space of search. Simple commands such as “Look at that red ball.” are parsed into two components :a verb and an indexical noun phrase. Consider the following diagram of a robot with the vision system having located three objects matching the correct shape and color in the field of view. The pointer is visually located and its direction is identified. Following the direction of the pointer, the object in the middle is pointed to and it is selected. This scenario is very simplistic and the pointer is clearly pointing to the object in the middle but in general the situation can be more difficult.



Consider the following scenario. The pointer and direction of pointing are difficult to visually discern. In a phrase like “this…” are we to pick out an object closer to the robot or to the pointer? When there are two objects in the direction of the pointer, should the closer or farther one be selected?

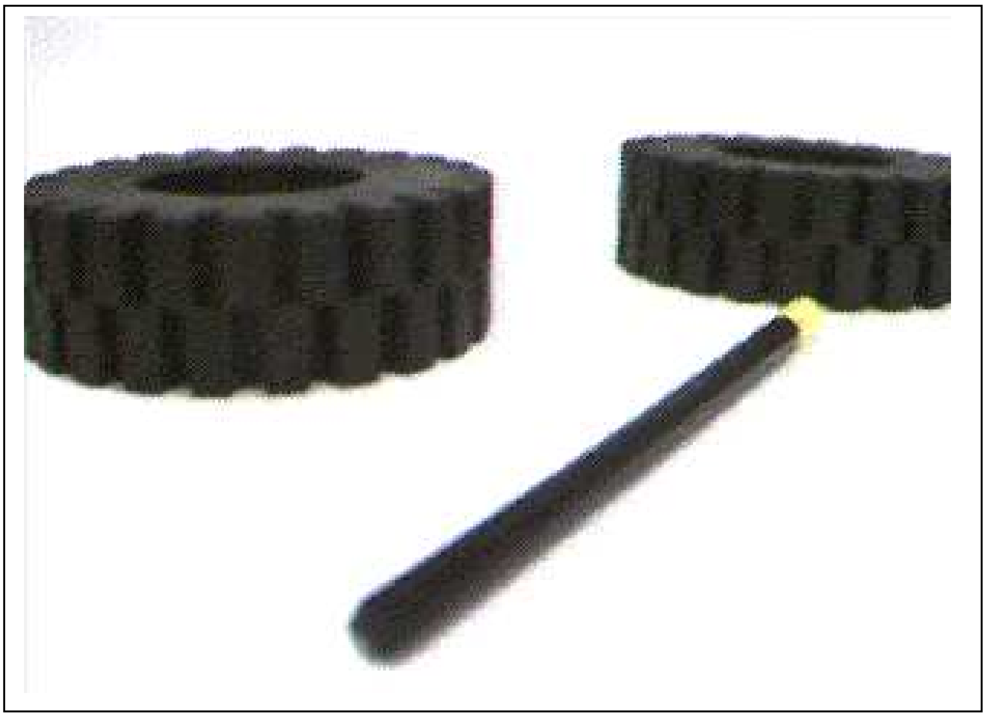


In the remainder of this paper we make the following two assumptions. The intersection of the salient direction of the robot's field of view and the pointer must form a degree less than 90 degrees. This is shown in the following diagram. An example for this assumption is when two people walk along a path and point objects to one another along the way. Another assumption is that the pointer will be held horizontally. This is to avoid computations in involving angular distortions.



3 Visual algorithms for computing intended objects

The following picture taken by a Connectrix vision system which costs about $200 running on a PC with Linux OS shows an image of a pointer and two disk-shaped objects. The pointer is pointing to the object in the right.



Since objects and the pointer are of fixed size, we can run experiments to discover object sizes at various distances and various angles. This is shown in the diagram below. For example, object 1 will appear larger and in the center of image (since it is in the center of field of view) whereas object 2 will appear in one side of the image (since it is in the corner of field of view) and smaller. This technique of recovering distances was performed successfully in a previous project. We will use our a-priori information about objects and distances to recover depth information from images. What can be measured from the image is distance from center of image d and image size s. If S is the actual object size and t is the angular distance from the object. t is computed from d. Distance D of object is a function of d and s. After training the vision system, this function is a table look up we gather from experimenting with known D, d, and s. If we keep the pointer in fixed number of orientations and distances by experimentation we can gather data about relationship of its orientation to how it appears in the image. Using a-priori relationships the pointer distances and orientation can be recovered from images.

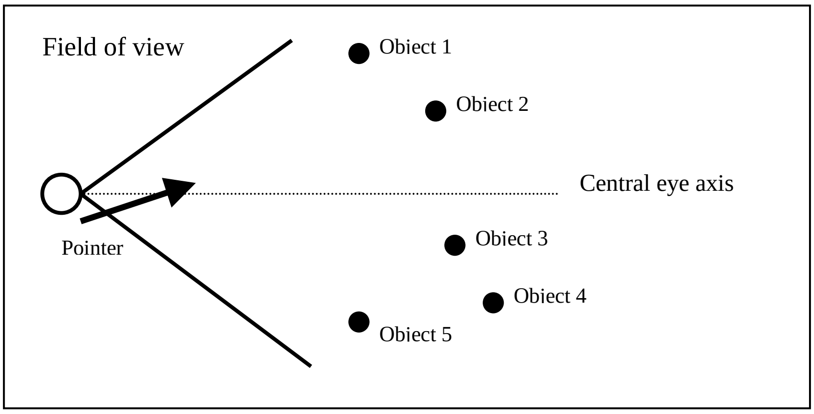
Having recovered the distances and angles from the image it is like using 2D image from robot's field of view to generate a "god's eye" (looking down view pictures we showed earlier) view. This becomes another assumption (our third so far) for our algorithm in the following section. For our fourth assumption, we assume object selection from the point of view of the pointing agent. The pointer in the possession of the pointing agent might be a human arm or a pointer device like a laser pointer or a just an elongated object like a stick. With this assumption, the pointer and agent' eye are relatively close. This is in contrast to considering two agents, one being guided by pointing and the second pointing where the distances might be comparable to distances to objects being selected.

4. An Algorithm

This algorithm is a first approach and it is based on the most simple scene and it is focused to select the object understanding that pointing is made in the usual way like the human being pointing. This scene consist:

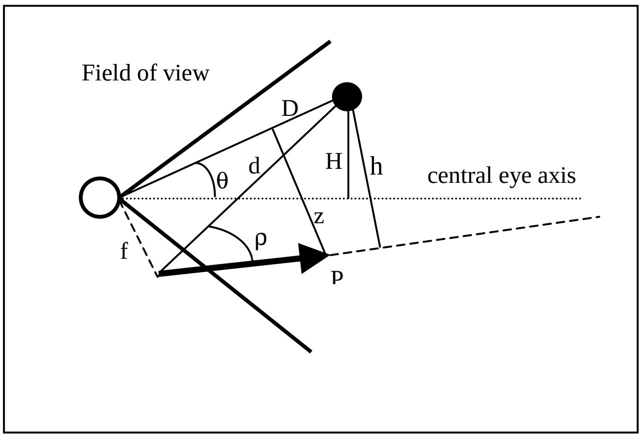
- The environment has the room size dimensions.
- The illumination is uniform over the entire room.
- There are n objects, $n \geq 2$.
- All objects are identical in size, shape, and color. They are spherical objects with radii.
- There is a pointer to select the object. This can be an arm from the secondary view person
- There are two views: Primary and Second view.
    - Primary: Point of view from the person (or Robot) that is going to select and acquire the object.
    - Secondary: from the person (or Robot) that is going to point an object.
- The objects to be selected must be in the field of view (Primary View)
- The distance between the centers of primary and secondry viewers is small.
- Every scene and pointer-base configuration has an efficiency level. The revelant parameters will be discussed later.

A typical scene could be the following which has five objects:



In this case, the selected object must be the one number 2.

Suppose that we arrange the top view (God's eye view) from the secondary view point. The parameters that are interesting in the scene are seen in the following diagram:



The parameters meaning are:
- D : Object axis – the line connecting the center view of secondary and the of object.
- h : Distance between object and pointer axis
- H : Distance between object and eye axis.
- f : Distance between the primary view center and secondary view center.
- d : Distance between object and pointer-base.
- θ : Angle between eye axis and object axis (D)
- ρ : Angle between pointer axis and the line from object to pointer-base
- z : Distance between object axis and pointer-tip.
- P: Pointer-tip, (x,y) location

Note : All object Distances refer to the object center.

Then, if we consider a scene with n objects, the algorithm has n + 1 inputs:

- (x,y) location for each object in the scene. (n inputs)
- (x,y) location of point p. (Pointer Tip). (1 input).

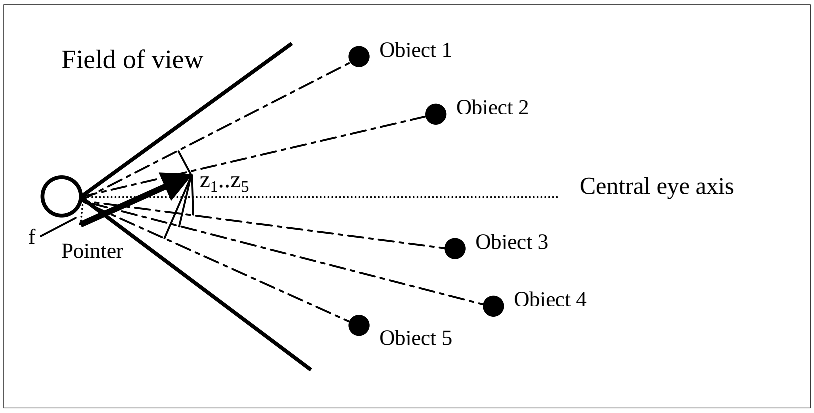
Then heuristic is:

1. Calculate the object axis for each object.
2. Calculate (z) the distance between the object axis and P for each object.
3. Select the object with the minimum distance z.

The output is:

The selected object in step 3.

We can see the heuristic graphically in the following drawing.



If the minimum distance is not unique, then the algorithm will not uniquely determinate the object. It instead will report a number of candidate objects. Further refinement must account additional information

The pointer will not always be effective. We offer the following statistical computation as a measure of pointer effectiveness

We define several indices of effectiveness E_1 & E_2.
For this propose we consider the distance between object axis and pointer-tip (z)

$E_1 = \min\{z_1, z_2, ..., z_n\}$ Where n is the number of objects.

$E_2 = \frac{z_1 + z_2 + ... + z_n}{n}$ Where n is the number of objects, i.e. E_2 is the average of all z values.

In a later phase we will develop more sophisticate Efficiency parameters and the basic scene will be extended to a more versatile one.

5. A brief discussion of parameters

We'll start with the object axis. We believe the distance between the eye and an object (i.e., D is the Figure) is not that important in our use of object selection scheme.

The lines from objects perpendicular to the central eye axis are just as important as the lines from objects perpendicular to pointer axis. This parameter is very similar to object shadow lines. However, we feel that object shadow lines better reflect the orientation of the pointer.

The pointer size and the lines between objects and the pointer base are important. Long pointers must be more accurate than short ones, i.e., we expect a greater effectiveness form a long pointer than a short pointer.

6. Related work

Pointing precedes language and notation in human evolution as can be seen in writings found in http://www.seas.gwu.edu/seas/institutes/nel/. Other basic understandings for pointing can be found in human visual perception [Bundesen and Shibuya 1995]. Deixis was born out of studies of theories of reference and making sense of notation and connections of symbolic systems to the physical world. This problem is sometimes known as the reference problem. As such it has been studied in Linguistics, Psychology, Neuro-biology, and AI. By engineering a robotic agent that can use deixis we find ourselves at a place to decide attention mechanism, language and communication, and sensor and perception management. A proper discussion of deixis goes beyond the scope of this space in this proposal.

In robotics, there are recent works in gesture recognition. For example, a group in UCSD is developing a hand-gesture scheme. The interpreted hand-gesture is transmitted to a remote robot vehicle for execution. Typically parameters for vehicle operation such as speed and turns are transmitted. Our proposed work in combining pointing with natural language is novel and we don't know of anyone developing such a system. IS Robotics (www.isr.com) has begun developing a natural control interface between human and robot using a wearable computer. The wearable interface seems to include speech and gesture recognition, eye-tracking, and augmented reality heads-up displays. IS robotics is using interesting technology and in our follow on work we may consider using wearable computers to add to the functionality of our proposed deictic system. Our work is "looking where the pointer points". Perhaps an interesting companion to our work is "pointing to where we are looking" work described in [Marjanovic et al 1996]. This work describes MIT's COG physically pointing in the direction of foveation.

Summary and Conclusion

We have use of a pointer in selecting an object in the field of view. We feel that our algorithm is a good approximation of natural use of pointer devices. We plan on experimenting with an implementation of the proposed algorithm. An alternative to our algorithm is a learning system that will learn from a given set of examples that identify objects being selected by a pointer.

References

P. Agre and D. Chapman, Pengi: An implementation of a theory of activity, Proceedings of **AAAI-87**, pp. 268-272.

C. Bandera, H. Hexmoor, S. Shapiro, A. Izatt, F. Du, K. Kramer. (1997) FEVAHR final report. Technical report available from Amherst Systems Inc, Buffalo, NY.

R.P. Bonasso and D. Kortenkamp (1995). Characterizing an architecture for intelligent, reactive agents. In **AAAI Spring Symposium** on Lessons Learned From Implemented Software Architectures for Physical Agents, pages 1-6. AAAI press.

C. Bundesen and H. Shibuya (editors, 1995). **Visual Selective Attention**. Lawrence Erlbaum associates, Hillsdale, USA.

H. Hexmoor, W. E. Underwood, On the Development of a Sensory-based Reactive System, Proc. GSFC Conference on Space Applications of Artificial Intelligence, Greenbelt, MD, June 1989.

H. Hexmoor, J. Lammens, G. Caicedo, S. C. Shapiro (1993), Behavior Based AI, Cognitive Processes, and Emergent Behaviors in Autonomous Agents, In Applications of AI in Engineering VIII, Vol.2, Applications and Techniques, (Eds) G. Rzevski and J. Pastor and R. Adey, pp. 447-461, CMI/Elsevier.

H. Hexmoor and D. Kortenkamp (1995). Issues on building software for hardware agents. **Knowledge Engineering Review** 10:3. Cambridge University Press, London.

H. Hexmoor (1995) Representing and Learning Routines Activities. PhD thesis, State University of New York in Buffalo, 1995.

H. Hexmoor (1996), Learning Routines, In Agent theories, architectures and languages, Michael Wooldrige, Joerg Mueller and Milind Tambe (editors), Springer Verlag.

H. Hexmoor and S. Shapiro (1997). Building skillful agents. In Ford, K., Hoffman, R., and Feltovich, P., editors, **Human and Machine Expertise in Context**., MIT press.

J. Lammens, H. Hexmoor and S. C. Shapiro (1995), Of Elephants and Men, NATO ASI Series on biology and technology of intelligent autonomous agents, Volume 144, Luc Steels (Editor), Springer-Verlag.

M. Marjanovic, B. Cassellati, and M. Williamson (1996) Self-taught Visually guided pointing for a Humanoid Robot, MIT TR.

P. Pook (1995). Teleassistance: using Deictic Gestures to Control Robot Action, PhD dissertation, University of Rochester.

S. Shapiro and W. Rapaport (1987). SNePS considered as a fully intensional propositional semantic network. In Cercone, N. and McCalla, G., editors, **The Knowledge Frontier**, pages 263-315. Springer-Verlag, New York.

R.M. Shiffrin and W. Schneider (1977). Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending, and a General Theory. In **Psychological Review**, volume 84-2, pages 127-190. The American Psychological Association, Inc, Washington DC.

S. Shapiro and W. Rapaport (1987). SNePS considered as a fully intensional propositional semantic network. In Cercone, N. and McCalla, G., editors, **The Knowledge Frontier**, pages 263-315. Springer-Verlag, New York.