CrossMark

Revisiting active perception

Ruzena Bajcsy[1] · Yiannis Aloimonos[2] · John K. Tsotsos[3]

Abstract Despite the recent successes in robotics, artificial intelligence and computer vision, a complete artificial agent necessarily must include active perception. A multitude of ideas and methods for how to accomplish this have already appeared in the past, their broader utility perhaps impeded by insufficient computational power or costly hardware. The history of these ideas, perhaps selective due to our perspectives, is presented with the goal of organizing the past literature and highlighting the seminal contributions. We argue that those contributions are as relevant today as they were decades ago and, with the state of modern computational tools, are poised to find new life in the robotic perception systems of the next decade.

1 Introduction

Some evening, long ago, our ancestors looked up at the night sky, just as they had done for thousands of years. But this time, it was different. For the first time, human eyes noticed

All authors contributed equally to all aspects of this manuscript.

✉ John K. Tsotsos
  tsotsos@cse.yorku.ca

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA

[2] Department of Computer Science, University of Maryland, College Park, MD, USA

[3] Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada

patterns in the stars. They saw caricatures of living things and as they scanned the heavens more appeared. These were eventually named, once language permitted, and we now know them as constellations of stars. This perhaps represents the first instance of *connecting the dots*, in a literal as well as figurative manner. How did this come about? What is needed to enable it? That one needs the ability to find a dot, and link one dot to another creating a group, is clear. Underlying this, however, is the critical ability to hypothesize a pattern: that group of stars looks like something I know—let me see if I can find other stars to complete the pattern. This is the essence of active perception—to set up a goal based on some current belief about the world and to put in motion the actions that may achieve it.

Through the years, the topic of perception, and particularly vision, has been a great source of wonder and study by philosophers and scientists alike. This history cannot be reviewed here and the interested reader should see Pastore (1971) and Wade and Wade (2000), among others. This paper presents the history of the computational perspective on the problem of active perception, with an emphasis on visual perception, but broadly applicable to other sensing modalities. Those interested in a biological perspective on active perception should see Findlay and Gilchrist (2003) and Poletti and Rucci (2013). What follows is a brief and selective history of the birth of the computational active perception paradigm. However, there is one source of motivation for the topic that deserves mention. The early work of Gibson (1950) proposed that perception is due to the combination of the environment in which an agent exists and how that agent interacts with the environment. He was primarily interested in optic flow that is generated on the retina when moving through the environment (as when flying) realizing that it was the path of motion itself that enabled the perception of specific elements, while dis-enabling others. That path of motion was under the

control of the agent and thus the agent chooses how it perceives its world and what is perceived within it. He coined the term *affordance*, which refers to the opportunities for action provided by a particular object or environment. These motivations play a role in an overall view of active perception, but as we will show, there is more to it as well. Gibson's notions pervade our computational perspective; it is useful to note however, that Gibson in later works (1979) became a proponent of direct perception, distinctly against an information processing view, in antithesis to what we present here.

Active Perception is a term that represents quite a broad spectrum of concepts. The SHAKEY robot, developed at the Artificial Intelligence Center of the Stanford Research Institute between 1966 and 1972, made history as the first general-purpose mobile robot to be able to reason about its actions (Nilsson 1969). Employing cameras, range-finders and bumpers as sensors, it could be given a task and then plan how to deploy its resources, specifically in our context, its sensing resources, to complete that task. From within this team, emerged perhaps the earliest Doctoral Dissertation on active perception by Tenenbaum (1970), where he writes:

> The author describes an evolving computer vision system in which the parameters of the camera are controlled by the computer. It is distinguished from conventional picture processing systems by the fact that sensor accommodation is automatic and treated as an integral part of the recognition process. Accommodation improves the reliability and efficiency of machine perception by matching the information provided by the sensor with that required by specific perceptual functions. The advantages of accommodation are demonstrated in the context of five key functions in computer vision: acquisition, contour following, verifying the presence of an expected edge, range-finding, and color recognition.

Barrow and Popplestone (1971) also acknowledged that vision is active by writing:

> …consider the object recognition program in its proper perspective, as a part of an integrated cognitive system. One of the simplest ways that such a system might interact with the environment is simply to shift its viewpoint, to walk round an object. In this way, more information may be gathered and ambiguities resolved. A further, more rewarding operation is to prod the object, thus measuring its range, detecting holes and concavities. Such activities involve planning, inductive generalization, and indeed, most of the capacities required by an intelligent machine.

They did not accompany this with any strategy or method that would enable such abilities and the community let these

words fade. Directions towards an embodiment would wait for quite a few more years until Bajcsy (1988)[1] wrote:

> Active sensing is the problem of intelligent control strategies applied to the data acquisition process which will depend on the current state of data interpretation including recognition.

She went on in that seminal paper to point out that this is not simply control theory. The feedback is performed not only on sensory data but on processed sensory data, and further, feedback is dependent on a priori knowledge, on models of the world in which the perceiving agent is operating. She summarizes the process nicely by saying:

> …we have defined active perception as a problem of an intelligent data acquisition process. For that, one needs to define and measure parameters and errors from the scene which in turn can be fed back to control the data acquisition process. This is a difficult though important problem. Why? The difficulty is in the fact that many of the feedback parameters are context and scene dependent. The precise definition of these parameters depends on thorough understanding of the data acquisition devices (camera parameters, illumination and reflectance parameters), algorithms (edge detectors, region growers, 3D recovery procedures) as well as the goal of the visual processing. The importance however of this understanding is that one does not spend time on processing and artificially improving imperfect data but rather on accepting imperfect, noisy data as a matter of fact and incorporating it into the overall processing strategy.

Aloimonos et al. (1988) add further structure to the concept:

> An observer is called active when engaged in some kind of activity whose purpose is to control the geometric parameters of the sensory apparatus. The purpose of the activity is to manipulate the constraints underlying the observed phenomena in order to improve the quality of the perceptual results.

These perspectives naturally lead to a definition of active perception. Considering the wealth of insight gained from decades of research since these early papers, from many perspectives, the following emerges that will form the skeleton for this paper:

> An agent is an active perceiver if it knows why it wishes to sense, and then chooses what to perceive, and determines how, when and where to achieve that perception.

Virtually any intelligent agent that has been developed to date satisfies at least one component of the active pen-

---

[1] *An earlier version of this perspective appeared in* Bajcsy (1985).

Table 1 The five main constituents of an actively perceiving agent are defined

Active perception	Definition
Why	The current state of the agent determines what its next actions might be based on the expectations that its state generates. These are termed Expectation-Action tuples. This would rely on any form of inductive inference (inductive generalization, Bayesian inference, analogical reasoning, prediction, etc.) because inductive reasoning takes specific information (premises) and makes a broader generalization (conclusion) that is considered probable. The only way to know is to test the conclusion. A fixed, pre-specified, control loop is not within this definition
What	Each expectation applies to a specific subset of the world that can be sensed (visual field, tactile field, etc.) and any subsequent action would be executed within that field. We may call this Scene Selection
How	A variety of actions must precede the execution of a sensing or perceiving action. The agent must be placed appropriately within the sensory field (Mechanical Alignment). The sensing geometry must be set to enable the best sensing action for the agent's expectations (Sensor Alignment, including components internal to a sensor such as focus, light levels, etc.). Finally, the agent's perception mechanism must be adapted to be most receptive for interpretation of sensing results, both specific to current agent expectations as well as more general world knowledge (Priming)
When	An agent expectation requires Temporal Selection, that is, each expectation has a temporal component that prescribes when is it valid and with what duration
Where	The sensory elements of each expectation can only be sensed from a particular viewpoint and its determination is modality specific. For example, how an agent determines a viewpoint for a visual scene differs from how it does so for a tactile surface. The specifics of the sensor and the geometry of its interaction with its domain combine to accomplish this. This will be termed the Viewpoint Selection process

tuple *why, what, how, when, where* and thus without the further connective constraint—'and then'—this definition would not be helpful. The key distinguishing factor is the *why* component—the counterpart to hypothesizing a pattern of stars with the wish to complete it. An actively perceiving agent is one which dynamically determines the *why* of its behavior and then controls at least one of the *what, how, where* and *when* for each behavior. This explicit connection of sensing to behavior was nicely described by Ballard (1991) in his *animate vision* concept, writing:

An animate vision system with the ability to control its gaze can make the execution of behaviors involving vision much simpler. Gaze control confers several advantages in the use of vision in behavioral tasks:

1. An animate vision system can move the cameras in order to get closer to objects, change focus, and in general use visual search.
2. Animate vision can make programmed camera movements.
3. Gaze control systems can be used to focus attention or segment areas of interest in the image precategorically.
4. The ability to control the camera's gaze, particularly the ability to fixate targets in the world while in motion, allows a robot to choose external coordinate frames that are attached to points in the world.
5. The fixation point reference frame allows visuomotor control strategies that servo relative to their frame.
6. Gaze control leads naturally to the use of object centered coordinate systems as the basis for spatial memory.

Ballard's points should be taken in the broader context. That is, they are not restricted to only vision but rather apply to other sensory modalities. Further, although the emphasis is on external or observable gaze, Ballard also hints at internal or non-observable components, specifically attention and choice of coordinate systems. Without these the generality seen in human visual systems cannot be achieved as argued by Tsotsos (2011). Table 1 presents more detail on these five basic elements, the active pentuple.

It is important to highlight that *selection* represents an integral process of all elements of the active pentuple. As Tsotsos (1980) wrote:

Since several simultaneous [interpretation] hypotheses can co-exist, a focus of attention mechanism is necessary in order to limit the number of hypotheses under consideration.

In addition, resource constraints play an important role not only because of computer power and memory capacity, but
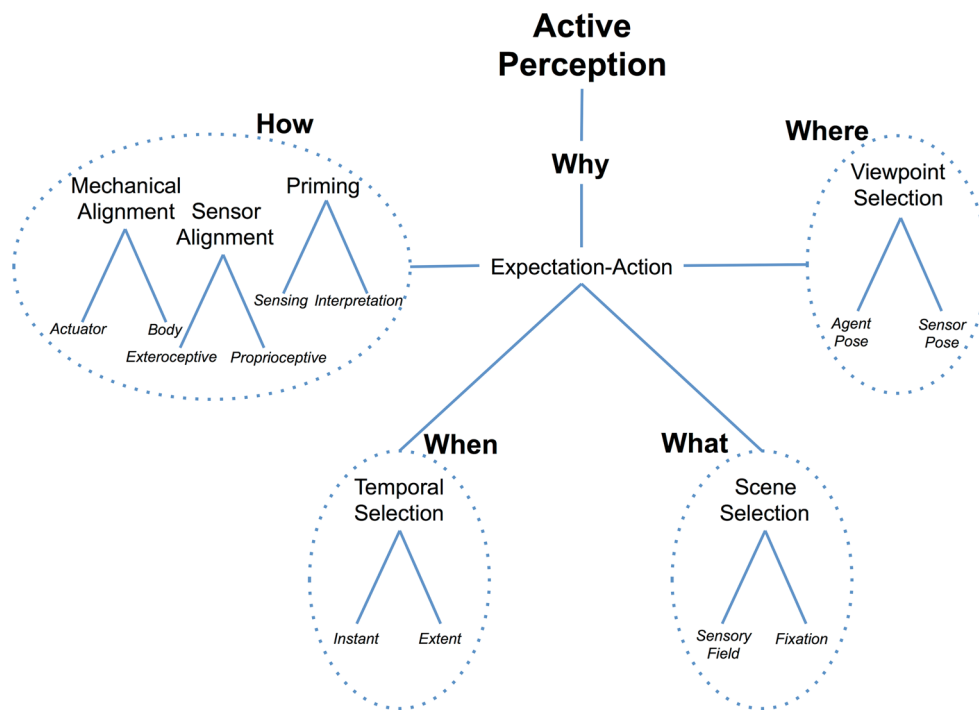
Fig. 1 The basic elements of Active Perception broken down into their constituent components. Instances of an embodiment of active perception would include the Why component and at least one of the remaining elements whereas a complete active agent would include at least one component from each

also because in practice, the number of sensors (and other physical components of an agent) is limited as well. Thus, choices must be made. In vision, the history of studies of visual attention covers centuries of thought and cannot be summarized here (see Tsotsos et al. 2005). Within computational vision, attention has played a role since the mid-1970s (e.g., Williams et al. 1977), with early proponents of the explicit link between computational and biological visual attention being found in Koch and Ullman (1985), Fukushima (1986), Tsotsos (1987), and Burt (1988).

Each element of the active perception definition can be further decomposed into the set of computations and actions it comprises, as shown in Fig. 1, noting that this decomposition is abstract and may be further detailed. Table 2 presents each of the elements of Fig. 1 along with some of the seminal works that first addressed those elements.

The remainder of this paper will present a brief historical perspective on the methods developed over the past 45+ years that address each element of what it means to be an active perceiver. The overwhelming conclusion that we draw, consistent with the seminal (Barrow and Popplestone 1971; Bajcsy 1988; Aloimonos 1990; Ballard 1991) as well as modern conceptualizations (Aloimonos 2013; Soatto 2013, among others), is that the full task of perception requires an active agent.

2 Why does an agent need to choose what to sense?

The fundamental difference between an active perception system and other perception systems lies in action, or lack of it. Whereas both types of systems include decision-making components, only the active system includes dynamic modulations to the overall agent's behavior, both external (via motors) and internal (via parameter configurations) (Coates et al. 2008). Let us give an example: consider that we have trained, using state of the art techniques in machine learning, a filter to recognize a particular object, like a knife, from images. Consider further that our filter has a success rate of 90%. This may be a breakthrough result, however, it may not be so interesting for a behaving robotic system. Indeed, by using this filter we can search for images—in a database— containing a knife and out of ten results, nine will be correct. But this is not sufficient for an active perception-action system that needs to act and make changes to the world. With 90% success, 10% of the time the system will be acting on the wrong objects. A different approach seems needed. In the best case, this filter could be used as an attention mechanism (among others) to suggest that a knife maybe in such location. This section dissects the components of an active vision system, as shown in Fig. 1, for specific cases and describes why would an agent wish to sense a particular scene or scene element. It does so by also providing some historical perspective.

Table 2 Details of the components of the diagram of Fig. 1. For those elements where multiple Research Seeds are given, it is because each addresses a different dimension of the problem; in most cases, many open problems remain for each component

Component	Definition	Research Seeds
HOW:	Mechanical alignment	
Actuator	Active control of motors (e.g., sequence of motor actions for multi-view stereo)	Moravec (1980)
Body	Active control of robot body and body part position and pose (e.g., to move robot to a location more advantageous for current task)	Nilsson (1969)
HOW:	Priming	
Interpretation	Active adaptation of perceptual interpretation system for current task and physical environment (e.g., tune system to be more receptive to recognition of objects and events relevant to current task)	Williams et al. (1977) (spatial) Tsotsos (1977) (spatiotemporal)
Sensing	Active adaptation of sensing system (e.g., to tune sensors to be more sensitive to stimuli relevant to current task)	Bajcsy and Rosenthal (1975)
HOW:	Sensor alignment	
Optical alignment	Active control of the optical elements of a visual sensor (focal length, gain, shutter speed, white balance, etc.) (e.g., accommodation: increases optical power to maintain a clear image on an object as it draws near)	Tenenbaum (1970)
Proprioceptive alignment	Active control of non-contact, non-visual sensors, such as inertial measurement units (e.g., the choice of path along which the IMU moves to measure linear acceleration and rotational velocity)	Early twentieth century, such as rocket stabilization
Exteroceptive alignment	Active control of sensors that measure the interaction with objects and environment such as applied forces/torques, friction, and shape (e.g., the choice of contact pattern over time)	Allen and Bajcsy (1985)
WHEN:	Temporal Selection	
Instant	Active prediction of when an event is expected (e.g., predicting the object movement in a sequence)	Tsotsos et al. (1979)
Extent	Active prediction of how long an event is expected (e.g., predicting the temporal extent of movement in an image sequence)	Tsotsos (1980)
WHAT:	Scene selection	
Sensory field	Active prediction of where in a scene a stimulus relevant to current task may appear (e.g., selection of the subset of an image where a face outline can be found)	Kelly (1971)
Fixation	Active prediction of which portion of a real-world scene to view (e.g., indirect object search, where an easy search for a semantically related object might facilitate search for a target object)	Garvey (1976) (indirect search) Moravec (1980) (interest points) Aloimonos et al. (1988) (ill-posed and nonlinear problems can be well-posed and linear for an active observer) Clark and Ferrier (1988) (saliency-guided head control) Burt (1988) (foveal fixations for tracking) Ye and Tsotsos (1995) (fixation selection for visual search)
WHERE:	Viewpoint selection	
Agent pose	Active selection of agent pose most appropriate for selecting a viewpoint most useful for current task (e.g., moving an agent to a close enough position for viewing a task-related object or event)	Nilsson (1969)
Sensor pose	Active selection of the pose of a sensor most appropriate for the current task (includes convergent binocular camera systems) (e.g., pointing a camera at a target in with the best viewing angle for its recognition)	Brown (1990) (general gaze control) Coombs and Brown (1990) (binocular vergence; use of nonvisual cues in stabilizing gaze) Wilkes and Tsotsos (1992) (viewpoint behaviors for recognition)

As mentioned in the introduction, the essence of active perception is to set up a goal based on some current belief about the world and to put in motion the actions that may achieve it. In other words, active perception is purposive (Aloimonos 1990) as it has to combine perception and action in synthesizing the pieces that will achieve a goal. In order then to develop a theory, one would require some framework for organizing the set of goals, some form of behavioral calculus.

Early efforts to create such models gave rise to finite state machines and dynamical systems that would model agents navigating successfully in some environment. Notable among them are hybrid automata (Košecká and Bajcsy 1994) which later, in conjunction with Bayesian inference techniques gave rise to a set of techniques broadly referred to as Simultaneous Localization and Mapping (SLAM) (Thrun et al. 2000) which have enjoyed much success. Despite this success, however, the problem of localization and mapping in the absence of GPS, and when only visual sensing is available, still remains a challenge (Salas-Moreno et al. 2013), as we still lack notions of robust spatial representations. Navigation requires perception of course, but myopic vision is sufficient—or in other words, navigation needs "perception in the large". This changes, however, as one needs to interact with the environment: manipulation requires detail and so it needs "perception in the small". In the case of navigation, it was possible to develop the appropriate mathematical models. But if the agent is involved in actions where it needs to make decisions, it needs to identify, recognize and manipulate different objects, it needs to search for particular categories in the scene, it needs to recognize the effects and consequences of different actions, and all this while executing a sequence of motor actions, then it needs details from the perceptual system and it becomes challenging to manage this complexity. Where does one start? And which models to employ?

Referring back to Fig. 1, we are now ready to see its components in action. Imagine an active perception agent that is at a table with the task of making a Greek salad, with all the ingredients and tools present in a cluttered scene, just like in a kitchen. The agent has a plan—or knowledge—of what is needed to produce the salad. Let's say that the plan calls first for finding a knife and cutting a cucumber. Thus, the agent would need to scan the scene in order to locate a knife. This is a top down attention problem where knowledge of a model of the object is used by the attention mechanism to find the image region likely to have a knife in it. An example of a recent approach implementing such a process can be found in Teo et al. (2013, 2015). Thus the *What* elements of Fig. 1 are engaged, described further in Sect. 3. But what if the agent could not find the knife? Then, it would have to move, get another view of the scene or perhaps even affect the scene by moving objects around. Such actions could uncover the knife, which happened to be behind the bowl and thus occluded. In turn, the *Where* and *How* elements of Fig. 1 get engaged, further described in Sects. 4 and 5. Next, after an image region containing the knife has been identified, attention brings fixation to it for the purpose of segmenting and recognizing the object. Thus, the *When* element of Fig. 1 gets engaged in order to select part of the spatiotemporal data and map it to symbolic information.

The process of segmentation is as old as the field of computer vision and a large number of techniques have been developed over the years. The bulk of those techniques work, usually, with a single image and produce a complete segmentation of everything in the scene. Representatives of such algorithms can be found in Felzenszwalb and Huttenlocher (2004), Alpert et al. (2007). Algorithms of this sort perform a variety of statistical tests by comparing different parts of the image, but they are rarely real time and they need to be told in advance the number of pieces that the segmentation should contain. Clearly, such a segmentation is not appropriate for an active perception system. Ideally, one would want to segment the objects where attention is drawn (e.g. the knife, or the knife and the cucumber in our previous example). Over the past 10 years there have been a few approaches for segmenting objects driven by attention (Mishra et al. 2009, 2012; Björkman and Kragic 2010; Bjorkman and Eklundh 2006; Andreopoulos and Tsotsos 2009, Andreopoulos et al. 2011). Recently, the introduction of the Microsoft Kinect sensor has increased interest to the problem (Mishra et al. 2012b) because it makes it slightly easier, but challenges still remain. Even from a point cloud it may not be possible to perform segmentation without taking more visual information (or other information) into account. See for example a current approach on using the symmetry constraint for this task (Ecins et al. 2016).

For an active camera system, a summary of the many reasons for active control follows:

1. to move to a fixation point/plane or to track motion
- camera saccade, camera pursuit movement, binocular vergence changes

2. to see a portion of the visual field otherwise hidden due to occlusion
- manipulation; viewpoint change

3. to see a larger portion of the surrounding visual world
- exploration

4. to compensate for spatial non-uniformity of a processing mechanism
- foveation

5. to increase spatial resolution
- sensor zoom or observer motion

6. to change depth of interest
- stereo vergence

7. to focus
- adjust focal length

8. to adjust depth of field
- adjust aperture

9. to adapt to light levels
- adjust shutter exposure time

10. to disambiguate or to eliminate degenerate views
- induced motion (kinetic depth); lighting changes (photometric stereo); viewpoint change

11. to achieve a *pathognomonic*[2] *view*
- viewpoint change

12. to complete a task
- viewpoint change

3 How does an agent choose what to sense?

The selection of what in a sensory field to examine falls immediately into the realm of attentive processing. It has been a common tactic throughout the history of Robotics, Computer Vision and indeed all of Artificial Intelligence to deploy methods that limit processing due to limited computational resources, both memory and processing speed. Attention has been of interest to philosophers, psychologists and physiologists for a long time (for a history see Tsotsos et al. 2005). The number of experiments and theories is incredibly large and daunting in their complexity and mutual inconsistency. Nevertheless, the constant that remains is that humans, and many other animals, exhibit selective behavior in everyday life and scientists of many disciplines continue to be motivated by the search for its explanation. It should be noted that the vast bulk of research has focused on visual selective mechanisms with less on auditory attention and then far less for the other senses. Against this backdrop, several computer vision systems also included attentional strategies to limit processing in an attempt to both mimic human visual behavior as well as to economize on processing in the face of limited resources. The magnitude of the problem of resources was always felt in a practical sense yet it was only formally proved in 1989 (Tsotsos 1989) when the computational complexity of basic visual matching process was shown to be NP-Complete.[3] It was further proved there that the application of domain and task knowledge to guide or predict processing is a powerful tool for limiting processing cost turning the otherwise exponential complexity problem into a linear one. This section will focus on visual selection.

The first use in a vision system was for oriented line location in a face-recognition task (Kelly 1971). Edges in reduced resolution images were found first and then mapped onto the full image to provide face outline predictions. These guided subsequent processing. Bajcsy and Rosenthal (1975) connected visual focussing with visual attention and developed an algorithm that would focus a camera system to particular image locations, motivated by human attentive visual behavior. Garvey (1976) extended these notions to include not only locations in an image but also spatial relationships among locations, or in other words, scene context. He termed this indirect search but the general idea has broader implications. One of those implications is that the computational complexity of processing would increase without an appropriate control algorithm (the number of spatial relationships among objects in a scene is represented by an exponential function). This places further importance on selective behavior.

Perhaps the first large scale vision system was VISIONS, developed by Hanson and Riseman at the University of Massachusetts, Amherst. It incorporated image pyramid representations, attentional selection and focus of components of the hypothesis model space (Williams et al. 1977). All of this was for a single, static image. The time domain was added to this repertoire of attentional methods by Tsotsos (1980). The idea was used for temporal window prediction in a motion recognition task. Positions and poses of segments of the left ventricle during the left ventricular cycle limited the region of the X-ray image sequence to search as well as the time interval during which to search. Left ventricular motion and shape knowledge was organized hierarchically and that structure was used to generalize or specialize predictions. Bajcsy and Rosenthal built upon their earlier work, in another 1980 contribution, to tie spatial focussing to conceptual focussing in that the conceptual hierarchy was linked to spatial resolution and as a recognition process progresses, moving up and down the hierarchy, it could make request of the perceptual system for images of a particular resolution at a particular location.

Interestingly, in the early days, computational limits imposed severe constraints on what practical problems could be solved. As computers became faster and memory cheaper, those constraints shifted leading to a greater range of problems that could be successfully addressed. Today's impressive successes in AI are due not in small part to such hardware advances. But the important question is: does this mean those problems can be considered as solved? The answer is 'no' and the reason was laid out in a series of papers that examined the computational complexity of basic problems in AI (see review in Tsotsos 2011). As a result, the apparent solution seen currently to problems such as object recognition is only an illusion. The problem is not solved; rather, small instances of it can be solved within some reasonable error bounds. This

---

[2] *Pathognomonic is a term borrowed from medicine, where it means a sign or symptom specifically characteristic or indicative of a particular disease or condition, but in our use, means a viewpoint that yields an image that is characteristic or indicative of a particular object.*

[3] *Also, directly relevant to active control,* Ye and Tsotsos (1996, 2001) *prove that the complexity of the Sensor Planning Problem—formulated as the optimization task of selecting the set of robot actions that maximize the probability of finding a target within cost constraints—is NP-hard.*

is a long way of saying that attentional mechanisms will have utility for a long time to come.

In computer vision, the late 1980s seemed to mark this realization. In 1987, Tsotsos showed that attention can play an important role to reduce the computational complexity of visual matching. This was quickly followed in 1988 by two works, one by Clark and Ferrier (1988) and the other by Burt (1988) who investigated ways of taking the theoretical results and using them to build functional active perception systems. Clark and Ferrier described a control system of a binocular sensor (the Harvard Head) which allows shifts in focus of attention. They are accomplished via altering of feedback gains applied to the feedback paths in the position and velocity control loops of the camera system. By altering these gains, they implemented a feature selection process motivated by the saliency maps of Koch and Ullman (1985), which was an attempt to provide a computational counterpart to the then prominent Feature Integration Theory of visual search of Treisman and Gelade (1980). This was the first realization of the saliency idea for attentive selection. Burt, on the other hand, derived a different sort of motivation from biological vision. He considered familiar aspects of eye movement control in human vision: foveation, to examine selected regions of the visual world at high resolution given a varying resolution retina; tracking, to stabilize the images of moving objects within the eye; and high-level interpretation, to anticipate where salient information will occur in a scene. The result was a high performance visual tracking system. Important and enduring elements within this work include the image pyramid and an active hypothesis-and-test mechanism.

The hypothesis-and-test idea played a role in the work of Tsotsos (1992) who examined the conditions under which image interpretation was more efficient using a passive paradigm (single image) or an active one (dynamic, interpretation guided, image sequence). He proved that even though more images are analyzed, the active approach used a *hypothesis sieve* (from an initial broadly encompassing hypothesis, evidence accumulated image by image, gradually reduces its breadth and thus amount of processing) to progressively limit processing through the image sequence so that under certain conditions, its overall complexity was far less than the blind, unguided single image case. However, active is not always more efficient than passive vision but that the constraints developed might assist an agent in choosing which to deploy. It was further shown how active perception is a subset of attentional processing in general.

The quest for methods for how to choose what to process within an image, that is, what to attend to, is now being addressed by an increasingly larger number of researchers with significant progress (see the reviews in Borji and Itti 2013 or Bylinksii et al. 2015). Nevertheless, there are still many open issues (Bruce et al. 2015). In contrast, the quest for methods that determine which image to consider, that is, which visual field to sense, is not receiving much attention at all with preference seemingly given to blanket sensing of the full environment via sensors such as LIDAR. As one counterexample, Rasouli and Tsotsos (2014) show performance improvements with the integration of active visual viewpoint methods with visual attention techniques.

4 How does an agent control how sensing occurs?

A variety of actions must precede the execution of a sensing or perceiving action. The agent must be placed appropriately within the sensory field, in other words to be mechanically aligned to its task. The sensing geometry must be set to enable the best sensing action for the agent's expectations. This can be thought of as sensor alignment, and encompasses the components internal to a sensor such as focus, light levels, etc. Finally, the agent's perception mechanism must be adapted or primed to be most receptive for interpretation of sensing results, both specific to current agent expectations as well as more general world knowledge.

Historically, we believe that the Stanford Hand-Eye project provided the earliest instance of a method for mechanical and sensor alignment. Tenenbaum, as part of this 1970 Ph.D. thesis at Stanford University, built an eye/head system with pan and tilt, focus control, and a neutral density filter control of a Vidicon camera. In 1973, WABOT-1 was created at the University of Waseda, which was the world's first full-scale anthropomorphic robot (Kato et al. 1973). It was able to communicate with a human in Japanese and measure the distance and direction of objects using external receptors such as artificial ears and eyes. WABOT-1 could perform tasks by vocal command that integrate sensing with actions of its hands and feet. This was followed at Stanford by Moravec's 1980 PhD work where he developed a TV-equipped robot, remotely controlled by a large computer. The Stanford Cart included a *slider*, a mechanical swivel that moved the television camera from side to side allowing multiple views to be obtained without moving the cart, thus to enabling depth computation. The cart moved in 1-m spurts separated by 10–15 min pauses for image processing and route planning. In 1979, the cart successfully crossed a chair-filled room without human intervention in about 5 h.

Sandini and Tagliasco (1980) demonstrated the computational benefits of a foveated camera image, borrowing the characteristics from human vision. Their model includes an explicit ability to actively scan a scene as the only way of overcoming the inherent limitations of a space-variant retina. In 1987, Krotkov as part of his Ph.D. dissertation at the University of Pennsylvania, developed the U Penn Head, with hardware similar to the previous systems, in that it had pan

and tilt, focus, vergence control (Krotkov 1988, 1989). His advance to the previous work was to present solutions to two problems that arise in the context of automatically focusing a general-purpose servo-controlled video camera on manually selected targets: (i) how to best determine the focus motor position providing the sharpest focus on an object point at an unknown distance; and (ii) how to compute the distance to a sharply focused object point. Starting in 1985, Brown and Ballard led a team that designed and built the Rochester Head, a high speed binocular camera control system capable of simulating human eye movements (Soong and Brown 1991). The system was mounted on a robotic arm that allowed it to move at 1 m/s in a 2-m radius workspace. This system led to an increased understanding of the role of behavior in vision, in particular that visual computations can be simpler when interacting in the 3D world. In 1995, Kuniyoshi et al. developed an active vision system with foveated wide-angle lenses. Pahlavan and Eklund (1992) presented the KTH-head, an anthropomorphic vision system with a focus on control of the oculomotor parameters that was inspired by human eye movements. They demonstrated an impressive, high performance, mechanical design that included 3 degrees of freedom (DOF) independently for each eye, 2 DOF for the neck and one DOF for the base. Milios et al. (1993) presented the system called TRISH in order to examine the robotic utility of yet another human eye movement, torsion (rotation about the optical axis of the eye). They found that control over torsion permits control over the slant of the vertical horopter in binocular processing and developed an algorithm for achieving this. There were several additional robot heads developed around the same time including those at the University of Oxford (Du et al. 1991), Harvard University (Clark and Ferrier 1988), University of Surrey (Pretlove and Parker 1993), Institut National Polytechnique de Grenoble (LIFIA) (Crowley et al. 1992), The NIST head TRICLOPS (Fiala et al. 1994), The University of Illinois head (Abbott and Ahuja 1992), and the University of Aalborg head (Christensen 1993), many of which are presented in detail in the volume of Christensen et al. (1993). Each of these works addressed key questions regarding which oculomotor functionalities were useful to robotic perception. However, as each group learned, the engineering effort to build and maintain such systems was very large and this cost limited the proliferation of these methodologies. However, these early developments eventually made possible the development of the now ubiquitous binocular camera heads seen in most humanoid robots.

Touch can also be deployed in an active manner. From the beginning (since the 1980s) it was realized that vision is limited and has to be complemented by tactile perception. This has been documented by series of papers: Shape from Touch (Bajcsy 1984), Active Touch (Goldberg and Bajcsy 1984), and Feeling by Grasping (Bajcsy et al. 1984). Allen (1985) showed that tactile perception is preferred to vision when an agent must discriminate different material properties of hardness, demonstrating this on discriminating the hole on the cup handle vs. the body of the cup. Similarly, they have shown that one needs force sensing in addition to vision for exploratory mobility (Bajcsy and Sinha 1989). Thus, an agent must have sufficient control mechanisms to match sensor and sensing strategy to current task. If a robot is equipped with pressure/force sensors then it can determine material properties (hard, soft, etc.) and Lederman and Klatzky (1990) proposed exploratory procedures (EP's), which associate tactile perceptions with certain motions, such as the hardness with pressing perpendicularly on the surface, tangential motion on the surface that detects surface texture, lifting that detects weight. All these capabilities to observe geometric, material mechanical (kinematics and dynamics) properties permit the determination of surface functionality. A nice example of this is Sinha (1991) who developed a hybrid position force control scheme to guarantee stability of walking. As part of this effort he had to examine the geometry of the surface as well as its material properties, such as friction and hardness of the surface. The important result of this exploratory process is that it invokes the contextual expectations that will bring constraints on geometry (for urban straight lines, for jungle clutter, etc.) or functionality of manipulation such as cutting and piercing (Bogoni and Bajcsy 1994, 1995).

5 How does an agent determine when and where to sense?

The 'when' and 'where' of our active pentuple is next in our discussion. An agent can decide from where to sense or view a scene and to do so it must first, naturally, know where it is in the context of that environment. In other words, it must know its body position and orientation with respect to the ground/gravity. This can be represented by a 6 dimensional vector $(x, y, z, \theta_1, \theta_2, \theta_3)$ determined with respect to some arbitrary location as coordinate system origin and angles with respect to the ground plane. Given the basic coordinate system, an agent needs to establish the position and viewing angles of the head/eye as well as the kinematic chain of the arms/hands. In addition, we assume that the hands have tactile sensors so that the system can establish contact with the objects and in general external world. Vision and touch, and their kinematic relationships are established through a calibration process.

Once these forms of self-information are determined, an agent can proceed to pursue the completion of its task. With respect to vision, an agent can control the viewing angle and distance for each sensing act. After the target of a sensing action, the 'what' described earlier, is set, the agent must determine in which position to place its sensors to best per-

ceive the target and its aspects most relevant to the task at hand. The variables of interest then would be the distance to the target, r, and the angle of the camera optical axis, which can be represented using the polar angle ζ and azimuthal angle ϕ. In the context of active object recognition, Wilkes and Tsotsos (1992) showed a simple behavior-based control algorithm that, depending on the state of recognition, was able to successfully control these parameters. Dickinson et al. (1994) showed the utility of aspect graph representations for encoding the appropriate active changes of sensor viewpoint for recognition. No additional variables play a role for a non-convergent binocular camera system. However, in a convergent binocular system, vergence, version, and torsion variables are also important, as shown in Milios et al. (1993), who describe the only head involving all three. There are several binocular heads that permit vergence and version control, with perhaps the KTH head of Pahlavan and Eklundh (1992) being the best example.

Gaze control was addressed by Brown and colleagues (Brown 1990; Coombs and Brown 1990; Rimey and Brown 1991). The 1990 paper has particular interest because Brown details the many variables that need to be included in a control structure, using predictive methods, and considered different sorts of binocular eye movements including vergence. He contrasts several control strategies, paving the way for how the problem could be addressed in general. A solution to how to combine agent position and pose with sensing viewpoint was put forward in a series of papers by Ye and Tsotsos (1995, 1996, 1999, 2001) who began by examining the computational complexity of this sensor planning problem in the context of visual search in a 3D environment. They proved that it was NP-hard, and thus optimal solutions for all instances should not be expected. They then defined a strategy that successfully could find objects in unknown environments, a strategy refined and implemented on a variety of robot platforms (Tsotsos and Shubina 2007), including the PLAYBOT intelligent wheelchair (Tsotsos et al. 1998; Andreopoulos and Tsotsos 2007) and the ASIMO humanoid (Andreopoulos et al. 2011).

As implied earlier, the current goal of an agent determines which control strategy applies best. Actions such as pick up an object, insert a piece, or decompose clutter are addressed in (Tsikos 1987; Tsikos and Bajcsy 1991). In that work, they build a system composed of laser range finder (vision system) and a manipulator equipped with a selection of different grippers (two fingered hand and a suction cup) operating on a cluttered environment (postal objects, boxes, letters, tubes on a conveyor belt). The control system was modeled by a finite state machine with deterministic sensor/action connections. The goal was to have an empty conveyor belt while the system classified different objects into different pallets. The parameters to be considered were the object size in order to determine which end effector can pick them up. If the object was too big or too heavy the system had the option to push them aside. In the case that the goal is rather open ended, exploratory, the system must be equipped with some broad goals to learn about its environment. This boils down to learning about the geometrical and physical properties of the environment in order to survive or act (Weng 2003). This was executed in Bajcsy and Campos (1992) where exploratory procedures were developed that determined material properties such as hardness, brittleness, Compliance, elasticity, plasticity, viscosity, ductility, and Impact. Geometric properties were described as shape and size, while kinematic properties were determined as degrees of freedom. The exploration control was designed as to deliver the above-mentioned attributes. For vision, feedback control was guided by exhaustive search, changing the views as necessary. For haptic feedback, the guide was characterization of the surfaces for grasping purposes. The conclusion was that there is natural flow from active and dynamic perception, through exploration, leading to perceptual learning.

6 Empirical issues

## 6.1 The key difference between active and passive vision processes

Computer vision researchers have become accustomed to the use of datasets and bench-marks in their work. It is a good practical way to measure how much progress one is making. Indeed, the introduction of special datasets in stereo research or in optical flow research has acted first as a catalyst for new ideas with researchers competing for highest performance. The standard processing pipeline is shown in Fig. 2.

Having many different kinds of inputs together with ground truth information, enables almost every one with access to a personal computer to develop and experiment with new ideas. With time however, unless the datasets are replaced by new, more challenging ones, those same datasets contribute to the decline of the specific discipline. A striking example is the famous Middlebury Stereo Dataset. Researchers where researchers tailored their algorithms to

Fig. 2 The current standard processing pipeline common in computer vision

select sample from dataset → perception → record result
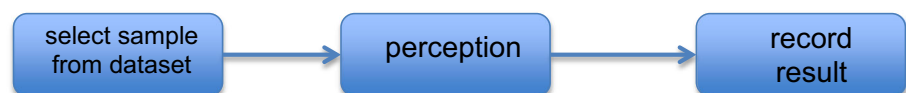
Fig. 3 Active perception processing pipeline
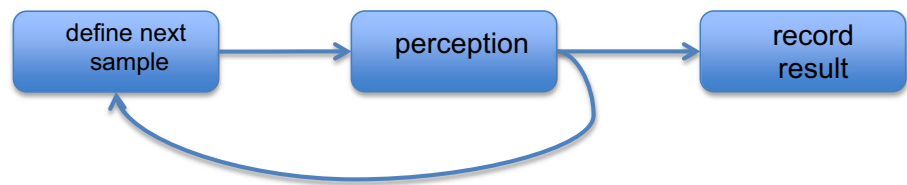
perform best for the images in the dataset. When maximum performance was reached, research clearly leveled off. There was nothing else to do with this dataset, but the problem was far from solved.

Active Perception systems attempt to deal with the perceptual-motor loop of a robot in a real environment. A hypothetical active processing pipeline is shown in Fig. 3.

Following Gibson in his famous debate with David Marr (Bruce and Green 1990), the 'agents'-literature focuses on complex behavior coming from mechanisms operating in tight coupling with a complex environment. This is in contrast to Marr's emphasis on the feed-forward computation of a representation from sensory data. On the other hand, machine learning approaches that learn both sensory perceptions and motor actions in some environment are not easy to find (but see Ackermann 2016). The reason is that it is difficult to build a statistical model of an environment when the system's perceptions are transformed into actions that affect the statistics of the input. Indeed, what could such an action-perception system learn? It would clearly have to develop a probability distribution of what happens. The hidden dependencies, redundancies and symmetries in this distribution are basically the structure of the world. But then, since the world is partly affected by the actions of the system, the shape of this distribution is action dependent—the system partly chooses what regularities exist, making the statistics difficult if not impossible to compute.

6.2 Addressing the "Entry barrier for active vision experimentation"

Modern computer vision seems highly dependent on the availability of public image datasets for algorithm evaluation and performance comparisons are done by tests on individual images within these common datasets. However, in active vision, the algorithm includes selecting and acquiring the image or images for any task. This requires specialized hardware that can be both costly and troublesome to use and maintain. The 'entry barrier' referred to in the title of this section is exactly this: it is relatively easy for anyone to download image datasets. It is far from easy to have an image acquisition platform with which to test active vision ideas. This naturally brings up the question of whether or not a simulated world might ameliorate the problem. Perhaps a simulated world might provide a testbed that can be

shared thus providing a way for algorithms to be compared. Such a testbed would need sufficient realism to be useful and needs not only good simulated world but also simulated image acquisition.[4]

This idea is not new. It was part of the 'animat' concept proposed by Terzopoulos and Rabie (1995, 1997), Rabie and Terzopoulos (2000) and further discussed in Terzopoulos (2003). Sprague et al. (2007) have also employed simulations for their work modeling human visuo-motor behaviors. Chessa et al. (2009) developed a simulator for active vision studies of stereo vision. And there are more examples. Part of the early issues with the idea had to do with how well the simulated environment matches the real world; however, modern graphics methods currently provide amazing realism as is evidenced primarily by the visual worlds available for computer games, such as Grand Theft Auto V, ARMA III, Battlefield 4 or others.[5] It may be that with only small effort, those techniques could create suitable worlds for active vision experimentation that could be appropriately evaluated against ground truth.

The visual realism of the world is not the only aspect however. An agent interacts with the world and the impact of the world on the agent and the agent on the world must also be modeled. There are a variety of ways a dynamic agent affects its world and ways the world affects the agent. A humanoid walking along a country road would raise dust or kick pebbles as its feet move, picking up an object would have proprioceptive effects on the agent's manipulator, running a hand across an object would have haptic consequences, and much more. The interaction with the visual world is just as important as the appearance of the visual world and depends strongly on agent design and configuration.

Finally, the acquisition process needs to be appropriately modelled and done so in a manner that is sufficiently flexible and customizable so that a wide variety of sensors may be considered for any particular agent design. The sensing modalities mentioned above each have unique characteristics and their control, which is a key aspect of active perception after all, must be realistically simulated.

Validation of a simulated world is of course critical and this might occur with specific side-by-side tests: choose a

---

[4] We thank an anonymous reviewer for this phrase, which describes the problem perfectly.

[5] http://www.pcworld.com/article/2840251/software-games/10-gorgeous-graphics-intensive-games-that-will-utterly-punish-your-pc.html.

real environment and sensing agent, develop a simulation of these, execute the same series of tests on each and measure and compare relevant performance properties. It could be an opportune time for a major push by the relevant funding agencies to develop an open platform that can be shared, customized, and evaluated in such a manner and then further used as a springboard to help remove the entry barrier to active perception experimentation.

6.3 Evaluating active vision systems

An active perception system consists of many components (modules, processes) that are interconnected in a network feeding each other with input. For example, an active perception system that understands human manipulation actions, has modules devoted to segmenting and tracking the hands, modules for segmenting, recognizing and monitoring objects and tools, modules for classifying movement into action descriptions, modules for goal perception, modules for prediction and modules for parsing. In addition, the attentional modules are connected to all of the above. There is a systematic way to evaluate such systems. In principle, one tests each module in isolation, then in pairs and eventually all together. Thus, one needs to find the reliability of a system based on the reliability of its components. This is a classic problem in systems engineering, mostly understood for feedforward systems but still challenging when feedback is involved. Currently, it takes the form of a question in the "cyber-physical systems" area. Indeed, an active perception system combines the cyber and physical aspects and as such, it falls within the modern science of cyberphysical systems which studies the interplay between programming language design, control and perception/knowledge (Alur 2016). Using this framework, the field could develop a methodology for evaluating active vision systems, predicting their performance as well as proving some of the system's properties.

7 Directions forward

The previous section presented possible extensions to current empirical methodologies that might be appropriate for active vision systems. Active Perception by definition needs a live environment or a random scene generator that can generate on demand given imaging geometry parameters. In principle, an active perception system can be evaluated in terms of its performance on the tasks that it is supposed to accomplish (or the performance of its components) and can be evaluated using basic techniques of systems engineering (such as, Sztipanovits et al. 2012). This represents one pointer for how research may advance in the future. However, there are several other issues that can be addressed and these follow.

7.1 Why has the robotics community failed to produce a flexible platform for researchers?

The design and implementation of robotics systems which embody the basic prerequisites for Active Perception is still hard. Why? Manipulators are typically built by mechanical engineers who appropriately equip them with position, velocity and sometimes accelerometer and force/torque sensors and of course corresponding actuators. The low level controllers (typically microcontrollers) are interfaced to a more powerful computer. The operating system on these computers then typically executes a program that, using input from the sensors, produces a control command which then is converted by the microcontroller into analog signal which in turn controls the actuation. In the old days one had to perform this programming in low level language in order to access the proper actuators. Today, this is hidden by the Robot Operating System (ROS) which enables the user to program robot activity in a higher level programming language such as C or C++ or even Python. ROS also makes it easy to access and connect the sensors (position, velocity, acceleration, force) with the actuators. To connect the vision sensors is not so straightforward but progress has been made. Both the PR2 and Baxter robots, for example, operate ROS. There is no commercially available control of a camera system, namely of the focus, aperture and field of view, though there commercially available pan and tilt controllers.

(a) **Commercial development for a complex robotic system that would include active perception is predicated on the public needs, and the market desires.**

Who are the customers? Industry uses robots in very constraint environments with limited number of tasks. The design criteria for such robotic systems are primarily accuracy, reliability and speed of performance. Active perception does not add value because those robots operate in environments where changes are predictable. The military has been investing in demonstrations that are more general. The service industry could be a prime customer for robots equipped with active perception. However, the scientific community has not yet delivered robust, real time and reliable performance, which would be adaptive to the varied environments. Currently, the most successful service robots are vacuum cleaners, clearly far from what the potential of what a robot with active perception can do. The robots operating in uncertain, dynamic environments cannot gather and process all relevant data and generate intelligent plans in real-time, so some sort of active perception is required. Why haven't results from active perception research led to more intelligent and autonomous robots? The reason is that the space of possibilities in a dynamic uncertain environment is huge and highly complex. Robots simply can't learn everything they

need to know in any reasonable time—consider the Google grasping project (Ackermann 2016). After 800,000 grasp trials in a constrained environment, the robots' abilities to pick up the objects was still improving. Not only is the space big and complex, but useful tasks (such as grasping) are also complex. It is difficult to construct models for the learning process that capture the most salient features of the tasks. This makes the construction of good reward functions for the learning process is nearly impossible. Even with "nice" reward functions, high-dimensional, non-convex spaces make learning slow. Add to this issues of convergence and generalization and it is clear that there is much work to be done

(b) **The vision community has focused on different agenda!**

The abundance of images on the web has promoted a success of the computer vision in object recognition and scene classification using convolutional neural networks, but they have not addressed the real time issue and how to deal with unexpected situations not available in the training data set. In service robotics (at home or even in hospitals, restaurants, old age homes, etc.) one will need to face up to the fact that one does not have the luxury to process all the available information (the real time constraint), hence the system has to select what to process at each instance of the time. This is the challenge of active perception, since one has to act upon the current available sensory information. Images are not the only sensory source: tactile and force sensor data and more is also important. Although some have explored low-level contact models (e.g., Trinkle et al. 2001; Kolev and Todorov 2015, and others), these are too detailed for the level of abstraction where active perception operates.

(c) **The great commercial opportunity is in the service industry**

As was mentioned earlier, the service industry seems an ideal receptor for robots with active perception. Applications such as companion robots for the elderly and infirm, hospital or nursing home robotics assistants to ameliorate the heavy burden that burns out care-givers, delivery robots, and more seem ripe. In order to make this happen, the research community will have to show that we can build robotic systems with active perception capabilities that are adaptive to varied environments and tasks, that can deal with unexpected, novel situations and that they can learn. Their performance must be robust, reliable and safe. At this point the humanoid robots focus on the control of their movements and their perceptual abilities are very limited (but see Andreopoulos et al. 2011). Robot use in hotels and hospitals is also very limited. It is unlikely that without a positive intervention, individual labs will be able to make strong progress; as

described above, there is a high entry cost to active robotics research, both hardware development and empirical work. A consortium composed of the private, industry academia and government can put together a program to create a flexible platform with a sensor suite (including vision), the operating system enabling higher level programming that will give impetus to researchers to show what is feasible. Perhaps a grand challenge such as DARPA executes from time to time is in order since it encourages competition and attracts young people to the field of robotics.

8 Conclusions

Any paper whose goal is a 're-visitation' of a research topic is bound to encounter differing viewpoints and opinions that have arisen during the history of the topic. It is a challenge to appropriately include them while not detracting from the main line of the argument. Further, it is very difficult to be certain that all the relevant landmarks of research have been included; other relevant reviews and collections include Crowley et al. (1992), Aloimonos (1993), Christensen et al. (1993), Crowley and Christensen (1995), Vernon (2008), Dahiya et al. (2010), Chen et al. (2011) and Andreopoulos and Tsotsos Andreopoulos and Tsotsos (2013a, b), among others. Here, our main argument is that despite the recent successes in robotics, artificial intelligence and computer vision, a complete artificial agent necessarily must include active perception. The reason follows directly from the definition presented in the first section: *An agent is an active perceiver if it knows why it wishes to sense, and then chooses what to perceive, and determines how, when and where to achieve that perception.* The computational generation of intelligent behavior has been the goal of all AI, vision and robotics research since its earliest days and agents that know why they behave as they do and choose their behaviors depending on their context clearly would be embodiments of this goal.

A multitude of ideas and methods for how to accomplish this have already appeared in the past, their broader utility perhaps impeded by insufficient computational power or costly hardware. We are of the opinion that those contributions are as relevant today as they were decades ago and, with the state of modern computational tools, are poised to find new life in the robotic perception systems of the next decade. To reach this conclusion we have traversed, over a period of time extending, personally for us the co-authors, at least 4 decades, and have been seen our views strengthened. Moreover, we have been affected by a variety of motivations, disciplines, approaches, and more. These cannot all be easily presented here within a single story line, but we feel that the following personal outlooks may permit the excursions that we have individually experienced to add an interesting texture to our main story line.

Ruzena Bajcsy I believe that there is a natural flow from Active Perception through Exploration to Perceptual Learning. In spite of a great deal of progress in sensing, computational power and memory, the goals set out in the original Active Perception paradigm are valid today. We are interested in conceptualizing the perceptual process of an organism/cyber-physical system that has the top-level task of functioning/surviving in an unknown environment. All the recognition algorithms depend on the context and it is context that determines the control of sensing, reasoning and action components of any system. To conceptualize this perceptual process four necessary ingredients have emerged for either artificial or biological organisms. First, the sensory apparatus and processing of the system must be active (controllable/adjustable) and flexible. Second, the system must have exploratory capabilities/strategies. It must also be able to evaluate at each step of exploration if it gains new information. Third, The system must be selective in its data acquisition in order to be able to act upon the perceptual information in real time. Fourth, The system must be able to learn. The learning process depends very much on the assumptions/models of what is innate and what is learned. The theory of the cyber-physical system that I have been pursuing predicts that an agent can explore and learn about its environment modulo its available sensors, kinematics and dynamic of its manipulators/end effectors, its degrees of freedom in mobility and exploratory strategies /attribute extractors. It can describe its world with an alphabet of set of perceptual and actionable primitives.

I take a great deal of inspiration from what is known about infant vision, and from psychological studies on perceptual exploration (Gibson 1950, 1979; Maturana and Varela 1987) and learning (Piaget 1962; Meltzoff and Moore 1989). Infant vision especially seems inspirational. At birth visual structures are fully present but not fully developed. Newborns can detect changes in brightness, distinguish between stationary and kinetic objects, and follow moving objects in their visual fields. During the first 2 months, due to growth and maturation their acuity improves, they can focus and their light sensitivity improves due to pupil growth. Stereo comes into play at about 3 months and motion parallax capabilities occur around 5 month of age. Monocular depth cues come about much later, perhaps at the 4–5 month mark. The control of head movement is also gradual. Given the above facts, one can state that at best our visuo-artificial systems correspond to an infant's capabilities at about 2–5 months. We have much to do!

Another goal is to understand how to build systems that interact with other humans. Collaboration between agents has been a challenging problem in the robotic community for many years. Perception of intent and agent affordances, the communication between agents, and the coordinated action are all open research problems. This challenge has an added layer of complexity when humans are added to the interaction the notorious *human-in-the-loop*. One of the ways we are exploring these human cyberphysical systems, is by using active perception to not only perceive the surroundings of the agents, but to examine the agents themselves. Collaborative teams of agents with different affordances offers methods of completing tasks that are far richer than homogeneous teams. By sensing the abilities of each agent, this diversity can be utilized to its furthest extent.

The affordances of the human agent are often treaded as being identical for all people. While this simplifies the modelling of the collaboration, it results in a poor tradeoff between system efficiency and safety. For instance, a device that helps an individual walk, should not impede a firefighter, but should also prevent an elderly individual from falling. While the need for this functionality is clear, the design and control of these systems requires knowledge of the individual's affordances - the system is not just human-in-the loop, but *individual*-in-the-loop.

Everyone is different, and we all change over time due to age, illness and treatment. By accounting for these changes we can better adapt a system to the interaction with the individual, while making rigorous statements about the abilities of the team, and the safety of the agents involved (Bestick et al. 2015).

Yiannis Aloimonos My own views have been shaped by basic questions regarding the overall organization of an intelligent system with perception, a fundamental requirement for building intelligent autonomous robots. What kind of information should a visual system derive from the images? Should this information be expressed in some kind of internal language? Should the information be in a single general purpose form leaving it to other modules to change it to fit their needs, or can a visual system directly produce forms of information suited to other specific modules? Is part of the vision system's function to control processing in other subsystems? Is it possible to draw a sharp boundary between visual processing and other kinds of processing? How is the interaction between Vision and Cognition? The basic thesis I have been developing, along the ideas of Aristotle, Varela, Gibson and others, is that action lies at the foundation of cognition. Every component of the intelligent system is designed so that it serves the action space, which in turn serves that component (feedback loops).

The Active Vision Revolution of the '80's and 90's developed such models for the early parts of the vision system, the eye. But if the agent is involved in actions where it needs to make decisions as well as recognize and manipulate different objects, then it becomes challenging to develop the appropriate mathematical models. Where does one start? And which models to employ?

The discovery of mirror neurons (Gallese et al. 1996) solved the dilemma, for me and others. It became clear that the system responsible for generating actions and the system responsible for interpreting actions that are visually observed, are basically the same at a high level. Thus, if it is hard to figure out how to organize the perception-action cycles of an active perception system, it may be easier to understand humans performing actions that we visually observe. In other words, if we could develop systems that understand human activity by observing it, at the same time we are developing the high level architecture of an active robotic agent that can perform the same actions. This was a powerful idea that led to a gargantuan development of computational approaches to interpreting human action from visual information and amounted to a shift of some "active vision" research to "action vision" research in my own work.

Consider again the kitchen scenario from before—the making of the Greek salad. But now consider the dual problem: an active perception system is watching a human making a salad. Its goal is to understand what it observes. Understanding has of course many layers, one needs more understanding of the action in order to replicate it than in order to name it. The active perception system watching the action would still need to engage the modules of Fig. 1 just as before (Yang et al. 2013). In addition, it would have to perform selection in space-time, i.e. to segment the video that it receives into meaningful chunks that can map to symbols. For example, the part of the video with the hand moving towards the knife will be segmented when the hand reaches the knife and will become the symbol REACH. The next video chunk will be GRASP. The next one will be MOVE KNIFE TO CUCUMBER'S LOCATION. Those symbols obey rules that in general form a grammar (tree like rules) or some graphical model (graphs, random fields, conditional random fields, etc.). Such knowledge allows a linguistic treatment of action, since one would be able using such models to create a parsing of the scene and a subsequent semantic analysis (Wörgötter et al. 2012; Summers-Stay et al. 2013; Pastra and Aloimonos 2012; Manikonda et al. 1999; Dantam and Stilman 2013; Fainekos et al. 2005; Aksoy et al. 2011; Worch et al. 2015; Yang et al. 2015; Coates et al. 2008; Maitin-Shepard et al. 2010).

Considering then an active perception system that is able to perform thousands of actions involving thousands of objects, how would we call the actions and objects involved? Would we name them action-1, action-2,..., action-m and similarly object-1, object-2, and so on? Or would we use their actual names, such as grasp, cut, move, screw, turn, press,…, cucumber, tomato, knife, and so on? The second choice allows the possibility of structuring the knowledge of an active perception system in terms of language, allowing for better communication with humans as well as for using natural language processing techniques in the integration of

the system. This avenue contributes to the grounding of the meaning of language and is an important research direction. (Teo et al. 2012; Kollar et al. 2010; Tellex and Roy 2009; Siskind 2001; Yu et al. 2011; Yang et al. 2014; Zampogiannis et al. 2015; Yang et al. 2015).

John Tsotsos My early years in the field, when we collectively believed that we could use computation to understand human visual and cognitive abilities, have played a large role in shaping my viewpoint today. Over the years I find myself more and more influenced by our ever-growing knowledge about the human visual system and human behavior. Perhaps more is known about the visual system of the brain than any other component, and more of the brain's cortical neurons are devoted to visual processing than any other task. Each discovery about human vision can be regarded as a hint or clue that might be helpful in developing a functioning artificial vision system. But there are far more experimental discoveries than useful clues. And how one translates these hints into real systems is also very important. At what level of abstraction should it apply? What mathematical formulation or computational construct best models the hint? How is the result evaluated with respect to its computational performance? What degree of faithfulness to those neurobiological and behavioral observations is most useful? It is important to understand that not all experimental observations about vision are useful and generally we do not have a way of knowing which. Any choice of neurobiological or behavioral hints to use in an artificial vision system constrains the space of possible systems that can result. A large aspect of the art of creating artificial vision is to select the right subset and to determine the best way to translate those hints into enabling elements.

The hallmark of human vision is its generality. The same brain and same visual system allow one to play tennis, drive a car, perform surgery, view photo albums, read a book, gaze into your loved one's eyes, go online shopping, solve 1000-piece jigsaw puzzles, find your lost keys, chase after your young daughter when she appears in danger, and so much more. The reality is that incredible as the AI successes so far have been, it is humbling to acknowledge how far there is still to go. The successes have all been uni-taskers (they have a single function)—the human visual system is a multi-tasker, and the tasks one can teach that system seem unbounded. And it is an infeasible solution to simply create a brain that includes a large set of uni-taskers. So how to move forward? We need to consider a broader set of vision problems. We can slowly move towards that goal by remembering that people—and their visual systems—move. Thus, the constraints that can be applied to vision system development can be expanded by including the constraints that self-motion imposes, such as the spatial, as well as temporal, correspondence between successive images; there is a cost in time and energy involved in moving the eyes (or body), and this constrains how often

one may be willing to do so; there must be an innate understanding of visual perspective and geometry in order to build an internal representation of what we have seen in previous views and where it is; and more. If we wish to fulfill the dream of humanoid robotic companions for our elderly and infirm, or household assistants, such constraints are central.

But with this possibility a new problem arises: the amount of sensory data to be processed grows rapidly and computational power begins to be strained. My solution is to seek a general purpose vision system that can be tuned to different functions depending on the task required of it and input it views (Tsotsos 2011). This, in fact forms the basis for my definition of attention: *Attention is a set of mechanisms that tune and control the search processes inherent in perception and cognition, with the major types of mechanisms being Selection, Suppression, and Restriction.* One aspect of this tuning is to select which portion of the input to process at any time (select the visual field, from which viewpoint it is sensed, etc.). The vision system can also select the manner in which this input is processed at any time, by priming the system for its current expectations, suppressing irrelevant computations thus enhancing the relevant ones, improving responses to task-relevant image characteristics, sharpening decision processes, and more. This is the essence of attentional behavior and active perception represents an important subset of the full range of attentional behaviors observed (Tsotsos 1992). Thus, I arrive my current perspective on the problem. It's all about control! A passive sensing strategy, no matter how much data is collected, gives up control over the quality and specific characteristics of what is sensed and at what time and for which purpose. Passive sensing reduces or eliminates the utility of any form of predictive reasoning strategy (hypothesize-and-test, verification of inductive inferences including Bayesian, etc.). And it's too early in the robot intelligence story to believe such reasoning strategies are ultimately unnecessary.

References

Abbott, A. L., & Ahuja, N. (1992, November). University of Illinois active vision system. In *Applications in optical science and engineering* (pp. 757–768). International Society for Optics and Photonics.

Ackermann, E. (2016). How google wants to solve robotic grasping by letting robots learn for themselves. IEEE Spectrum, March 28, http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/google-large-scale-robotic-grasping-project

Aksoy, E., Abramov, A., Dörr, J., Ning, K., Dellen, B., & Wörgötter, F. (2011). Learning the semantics of object-action relations by observation. *The International Journal of Robotics Research*, *30*, 1229–1249.

Allen, P. & Bajcsy, R. (1985). Two sensors are better than one: example of integration of vision and touch. *Proceedings of 3rd ISRR*, France, October.

Allen, P. K. (1985). *Object recognition using vision and touch*. PhD dissertation: University of Pennsylvania.

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, *1*(4), 333–356.

Aloimonos, J. (1990). Purposive and qualitative active vision. *Proceedings of 10th IEEE International Conference on Pattern Recognition*, vol. 1, pp. 346–360.

Aloimonos, Y. (Ed.). (2013). *Active perception*. Milton Park: Psychology Press.

Alpert, S., Galun, M., Basri, R., & Brandt, A. (2007). Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE Transaction on PAMI*, *34*(2), 315–327.

Alur, R. (2016). *Principles of cyber-physical systems*. Cambridge: MIT Press.

Andreopoulos, A., & Tsotsos, J. K. (2007). A framework for door localization and door opening using a robotic wheelchair for people living with mobility impairments. In *Robotics: Science and systems, Workshop: Robot manipulation: Sensing and adapting to the real world, Atlanta*.

Andreopoulos, A., & Tsotsos, J. K. (2009). A theory of active object localization. In *IEEE 12th international conference on computer vision*, pp. 903–910.

Andreopoulos, A., Hasler, S., Wersing, H., Janssen, H., Tsotsos, J. K., & Körner, E. (2011). Active 3D object localization using a humanoid robot. *IEEE Transactions on Robotics*, *27*(1), 47–64.

Andreopoulos, A., & Tsotsos, J. K. (2013a). 50 Years of object recognition: Directions forward. *Computer Vision and Image Understanding*, *117*, 827–891.

Andreopoulos, A., & Tsotsos, J. K. (2013b). A computational learning theory of active object recognition under uncertainty. *International Journal of Computer Vision*, *101*(1), 95–142.

Bajcsy, R. (1984). Shape from Touch. In G. N. Saridis (Ed.), *Advances in Automation and Robotics*. Stamford: JAI Press.

Bajcsy, R. (1985). Active perception vs passive perception. *Proceedings of 3rd IEEE workshop on computer vision: Representation and Control*, October 13–16, Bellaire, MI. (Washington DC: IEEE Computer Society Press), pp 55–62.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*(8), 966–1005.

Bajcsy, R., & Campos, M. (1992). Active and exploratory Perception. *CVGIP: Image Understanding*, *56*(1), 31–40.

Bajcsy, R., & Rosenthal, D. A. (1980). Visual and conceptual focus of attention. In S. Tanimoto & A. Klinger (Eds.), *Structured computer vision* (pp. 133–149). London, NY: Academic Press.

Bajcsy, R. K., & Rosenthal, D. A. (1975). Visual focussing and defocussing-an essential part of the pattern recognition process. Pattern recognition and data structures: In *Proceedings on IEEE Conference on Computer Graphics*.

Bajcsy, R., McCarthy, M. J., & Trinkle, J. C. (1984). *Feeling by Grasping. Proceedings of the IEEE international conference on robotics*, Atlanta

Bajcsy, R. & Sinha, P. R. (1989). Exploration of surfaces for robot mobility In *Proceedings of the Fourth international conference on CAD/CAM robotics and factories of the future*, pp. 397–404, vol. III, Tata McGraw-Hill, New Delhi, India.

Ballard, D. (1991). Animate vision. *Artificial Intelligence*, *48*(1), 57–86.

Barrow, H., & Popplestone, R. (1971). Relational descriptions in picture processing. In B. Meltzer & D. Michie (Eds.), *Machine intelligence 6* (pp. 377–396). Edinburgh: Edinburgh University Press.

Bestick, A.M., Burden, S.A., et al. (2015). *Personalized kinematics for human-robot collaborative manipulation*, IROS

Bogoni, L., & Bajcsy, R. (1994). Functionality investigation using a discrete event system approach. *Journal of Robotics and Autonomous Systems*, *13*(3), 173–196.

Bogoni, L., & Bajcsy, R. (1995). Interactive recognition and representation of functionality. *Computer Vision and Image Understanding*, *62*(2), 194–214.

Bjorkman, M., & Eklundh, J. O. (2006). Vision in the real world: Finding, attending and recognizing objects. *International Journal of Imaging Systems and Technology*, *16*(5), 189–208.

Björkman, M. & Kragic, D. (2010). Active 3D scene segmentation and detection of unknown objects. In *Proceedings of IEEE international conference on robotics and automation*, pp. 3114–3120.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

Brown, C. (1990). Prediction and cooperation in gaze control. *Biological Cybernetics*, *63*(1), 61–70.

Bruce, V., & Green, P. (1990). *Visual perception: Physiology, psychology, and ecology* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bruce, N., Wloka, C., Frosst, N., Rahman, S., & Tsotsos, J. K. (2015). On computational modeling of visual saliency: Understanding what's right and what's left, Special issue on computational models of visual attention. *Vision Research*, *116*(Part B), 95–112.

Burt, P. J. (1988). Attention mechanisms for vision in a dynamic world. In Proceedings on 9th international conference on pattern recognition, pp. 977–987.

Bylinksii, Z., DeGennaro, E., Rajalingham, R., Ruda, H., Jiang, J., & Tsotsos, J. K. (2015). Towards the quantitative evaluation of computational attention models, Special issue on computational models of visual attention. *Vision Research*, *116*, 258–268.

Williams, T., Lowrance, J., Hanson, A., & Riseman, E. (1977). II. Model-building in the visions system 1. In IJCAI-77: 5th International joint conference on artificial intelligence-1977: Proceedings of the conference, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, August 22–25, 1977 (Vol. 2, p. 644).

Chen, S., Li, Y. F., Wang, W., & Zhang, J. (Eds.). (2008). *Active sensor planning for multiview vision tasks* (Vol. 1). Heidelberg: Springer.

Chen, S., Li, Y., & Kwok, N. M. (2011). Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research*, *30*(11), 1343–1377.

Chessa, M., Solari, F., & Sabatini, S.P. (2009). A virtual reality simulator for active stereo vision systems. In *VISAPP (2)* (pp. 444–449).

Christensen, H. I. (1993). A low-cost robot camera head. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(01), 69–87.

Christensen, H. I., Bowyer, K. W., & Bunke, H. (Eds.). (1993). *Active robot vision: Camera heads, model based navigation and reactive control* (Vol. 7). Singapore: World Scientific Publishers.

Clark, J. J., & Ferrier, N. J. (1988). Modal control of an attentive vision system. In *Proceeding on international conference on computer vision*, pp. 514–523.

Coates, A., Abbeel, P., & Ng, A. Y. (2008). Learning for control from multiple demonstrations. In *Proceedings of the 25th international conference on machine learning* (pp. 144–151). New York: ACM.

Coombs, D. J., & Brown, C. M. (1990). Intelligent gaze control in binocular vision. In *Proceedings of 5th IEEE international symposium on intelligent control* (pp. 239–245). IEEE.

Crowley, J.L., Krotkov, E., & Brown, C. (1992). Active computer vision: A tutorial. In *IEEE international conference on robotics and automation*, Nice, France, May 11.

Crowley, J. L. & Christensen, H. I. (1995). Integration and control of active visual processes. *Proceedings of IROS 95*, Pittsburgh, August.

Crowley, J. L., Bobet, P. & Mesrabi, M. (1992). Gaze control for a binocular camera head. In *Computer Vision—ECCV'92* (pp. 588–596). Springer: Berlin.

Dahiya, R. S., Metta, G., Valle, M., & Sandini, G. (2010). Tactile sensing–from humans to humanoids. *IEEE Transactions on Robotics*, *26*(1), 1–20.

Dantam, N., & Stilman, M. (2013). The motion grammar: Analysis of a linguistic method for robot control. *Transactions on Robotics*, *29*, 704–718.

Dickinson, S., Christensen, H., Tsotsos, J. K. & Olofsson, G. (1994). Active object recognition integrating attention and viewpoint control. In: *Proceedings on European conference on computer vision*, pp 2–14.

Du, F., Brady, M., & Murray, D. (1991). Gaze control for a two-eyed robot head. In: *Proceedings on BMVC91* (pp. 193–201). London: Springer.

Ecins, C. F. & Aloimonos, Y. (2016). Cluttered Scene segmentation using the symmetry constraint. *IEEE international conference on robotics and automation.*

Fainekos, G. E., Kress-Gazit, H., & Pappas, G. J. (2005). Hybrid controllers for path planning: A temporal logic approach. In *Proceedings of the Forty-fourth IEEE conference on decision and control* (pp. 4885–4890). Philadelphia: IEEE.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*(2), 167–181.

Fiala, J. C., Lumia, R., Roberts, K. J., & Wavering, A. J. (1994). TRI-CLOPS: A tool for studying active vision. *International Journal of Computer Vision*, *12*(2–3), 231–250.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.

Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics*, *55*(1), 5–15.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(2), 593–609.

Garvey, T. D. (1976) *Perceptual strategies for purposive vision*. Doctoral Dissertation, Stanford University Stanford, CA.

Gibson, J. J. (1950). *The perception of the visual world*. Oxford: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton, Mifflin and Company.

Goldberg, K., & Bajcsy, R. (1984). Active touch and robot perception. *Computation and Brain Theory*, *7*(2), 199–214.

Kato, I., Ohteru, S., Kobayashi, H., Shirai, K., & Uchiyama, A. (1973). Information-power machine with senses and limbs. *Proc* (pp. 12–24). Udine, Italy: CISM-IFToMM sympoisum on theory and practice of robots and manipulators.

Kelly, M. (1971). Edge detection in pictures by computer using planning. *Machine Intelligence*, *6*, 397–409.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kolev, S. & Todorov, E. (2015). Physically consistent state estimation and system identification for contacts. In *IEEE/RAS international conference on humanoid robots*.

Kollar, T., Tellex, S., Roy, D., & Roy, N. (2010). Toward understanding natural language directions. In *Proceedings of human robot interaction conference*. Osaka, Japan.

Košecká, J., & Bajcsy, R. (1994). Discrete event systems for autonomous mobile agents. *Robotics and Autonomous Systems*, *12*(3), 187–198.

Krotkov, E. P. (1988). Focusing. *International Journal of Computer Vision*, *1*(3), 223–237.

Krotkov, E. P. (1987). *Exploratory visual sensing with a Agile Camera System*, Ph.D. theses, TR-87-29, UPENN

Krotkov, E. P. (1989). *Active computer vision by cooperative focus and stereo*. Berlin: Springer.

Kuniyoshi, Y., Kita, N., Sugimoto, K., Nakamura, S. & Suehiro, T. (1995). A foveated wide angle lens for active vision. *Proceedings on ICRA*, pp 2982–2988

Lederman, S. J., & Klatzky, R. L. (1990). Haptic classification of common objects: Knowledge-driven exploration. *Cognitive Psychology*, *22*(4), 421–459.

Manikonda, V., Krishnaprasad, P. S., & Hendler, J. (1999). *Languages, behaviors, hybrid architectures, and motion control*. New York: Springer.

Maturana, H. R., & Varela, F. J. (1987). *The tree of knowledge: The biological roots of human understanding*. New Science Library/Shambhala Publications.

Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Journal of Developmental Psychology*, *25*(6), 954.

Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., & Abbeel, P. (2010). Cloth grasp point detection based on multiple-view geometriccues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2308–2315). IEEE.

Milios, E., Jenkin, M., & Tsotsos, J. K. (1993). Design and performance of TRISH, a binocular robot head with torsional eye movements, Special issue on active robot vision: Camera heads. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(1), 51–68.

Mishra, A., Aloimonos, Y. & Fah, C. L. (2009). Active segmentation with fixation. *IEEE 12th international conference on computer vision*, pp 468–475.

Mishra, A. K., Aloimonos, Y., Cheong, L. F., & Kassim, A. (2012a). Active visual segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 639–653.

Mishra, A.K., Shrivastava, A., & Aloimonos, Y. (2012b). Segmenting "simple" objects using RGB-D. *IEEE Proceedings on international conference on robotics and automation*, pp 4406–4413.

Moravec, H. P. (1980). *Obstacle avoidance and navigation in the real world by a seeing robot rover* (No. STAN-CS-80-813). Stanford University CA, Dept of Computer Science.

Nilsson, N. J. (1969). *A mobile automaton: An application of artificial intelligence techniques*. Menlo Park, CA: Artificial Intelligence Center, SRI International.

Pahlavan, K., & Eklundh, J.-O. (1992). A head-eye system: Analysis and deisgn. *CVGIP: Image Understanding*, *56*(1), 41–56.

Pastore, N. (1971). *Selective history of theories of visual perception* (pp. 1650–1950). Oxford: Oxford University Press.

Pastra, K., & Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical Transactions of the Royal Society: Biological Sciences*, *367*, 103–117.

Piaget, J. (1962). *Play, dreams, and immitation on childhood*. New York: W.W. Norton.

Poletti, M., & Rucci, M. (2013). Active vision: Adapting how to look. *Current Biology*, *23*(17), R718–R720.

Pretlove, J. R. G., & Parker, G. A. (1993). The surrey attentive robot vision system. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(01), 89–107.

Rabie, T. F., & Terzopoulos, D. (2000). Active perception in virtual humans. *Vision Interface (VI)*, pp 16–22.

Rasouli, A., & Tsotsos, J. K. (2014). Visual saliency improves autonomous visual search. In *2014 Canadian Conference on Computer and Robot Vision (CRV)* (pp. 111–118). IEEE.

Rimey, R. D., & Brown, C. M. (1991). Controlling eye movements with hidden Markov models. *International Journal of Computer Vision*, *7*(1), 47–65.

Salas-Moreno, R., Newcombe, R., Strasdat, H., Kelly, P., & Davison, A. (2013). Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1352–1359.

Sandini, G., & Tagliasco, V. (1980). An anthropomorphic retina-like structure for scene analysis. *Computer Graphics and Image Processing*, *14*(4), 365–372.

Sinha, P. R. (1991). *Robotic exploration of surfaces and its application to legged locomotion*, Ph.D. dissertation, Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, *15*, 31.

Soatto, S. (2013). Actionable information in vision. In *Machine learning for computer vision* (pp. 17–48). Berlin: Springer.

Soong, J. & Brown, C. M. (1991). *Inverse kinematics and gaze stabilization for the Rochester robot head*, TR 394, Computer Science Dept., U. Rochester.

Sprague, N., Ballard, D., & Robinson, A. (2007). Modeling embodied visual behaviors. *ACM Transactions on Applied Perception (TAP)*, *4*(2), 11.

Sztipanovits, J., Koutsoukos, X., Karsai, G., Kottenstette, N., Antsaklis, P., Gupta, V., et al. (2012). Toward a science of cyber–physical system integration.*Proceedings of the IEEE, 100*(1).

Summers-Stay, D., Teo, C., Yang, Y., Fermüller, C., & Aloimonos, Y. (2013). Using a minimal action grammar for activity understanding in the real world. *Proceedings of the 2013 IEEE/RSJ international conference on intelligent robots and systems* (pp. 4104–4111). Vilamoura, Portugal.

Tellex, S. & Roy, D. (2009). Grounding spatial prepositions for video search. *Proceedings of the eleventh international conference on multimodal interfaces (ICMI-2009)*. Cambridge, MA.

Tenenbaum, J. M. (1970). *Accommodation in computer vision*, Ph. D. Thesis, Computer Science Department Report, No. CS182, Stanford University, Stanford, CA.

Teo, C., Yang, Y., Daume, H., Fermüller, C., & Aloimonos, Y. (2012). Towards a Watson that sees: Language-guided action recognition for robots. *Proceedings of the 2012 IEEE international conference on robotics and automation* (pp. 374–381). Saint Paul, MN.

Teo, C. L., Fermüller, C., & Aloimonos, Y. (2015). A Gestaltist approach to contour-based object recognition: Combining bottom-up and top-down cues. *The International Journal of Robotics Research*, 0278364914558493.

Teo, C.L., Myers, A., Fermuller, C. & Aloimonos, Y. (2013). Embedding high-level information into low level vision: Efficient object search in clutter. In *Proceedings on IEEE international conference on robotics and automation*, pp. 126–132.

Thrun, S., Burgard, W., & Fox, D. (2000). *Probabilistic robotics*. Cambridge: MIT Press.

Terzopoulos, D., & Rabie, T. F. (1995). Animat vision: Active vision in artificial animals. In *Proceedings on fifth international conference on computer vision, 1995* (pp. 801–808). IEEE.

Terzopoulos, D., & Rabie, T. (1997). Animat vision: Active vision in artificial animals. *Videre: Journal of Computer Vision Research*, *1*(1), 2–19.

Terzopoulos, D. (2003). Perceptive agents and systems in virtual reality. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 1–3.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Trinkle, J. C., Tzitzouris, J. A., & Pang, J. S. (2001). Dynamic multi-rigid-body systems with concurrent distributed contacts: Theory and examples. *Philosophical Transactions: Mathematical, Physical, and Engineering Sciences*, *359*(1789), 2575–2593.

Tsikos, C. J. (1987). Segmentation of 3D scenes using multimodal Interaction between machine vision and programmable mechanical Scene Manipulation, Ph. D dissertation, Department of Computer and information Science, University of Pennsylvania, December.

Tsikos, C. J. & Bajcsy, R. (1991). Segmentation via manipulation. *IEEE Transaction on Robotics and Automation, 7*(3).

Tsotsos, J. K. (1977). Knowledge-base driven analysis of cinecardioangiograms. In *Proceedings of the 5th international joint conference on Artificial intelligence*. Vol. 2 (pp. 699–699). San Mateo: Morgan Kaufmann Publishers Inc.

Tsotsos, J.K., Mylopoulos, J., Cowey, H. D., & Zucker, S. W. (1979). ALVEN: A study on motion understanding by computer. In *Proceedings of the 6th international joint conference on Artificial intelligence* (Vol. 2, pp. 890–892). San Mateo: Morgan Kaufmann Publishers Inc.

Tsotsos, J. K. (1980). *a framework for visual motion understanding*, Ph.D. Dissertation, Department of Computer Science, CSRG TR-114, University of Toronto, May.

Tsotsos, J. K. (1987). A "Complexity Level" Analysis of Vision. *Proceedings of the 1st international conference on computer vision*, pp. 346 – 55, London, UK.

Tsotsos, J. K. (1989). The complexity of perceptual search tasks. In *IJCAI* (Vol. 89, pp. 1571–1577).

Tsotsos, J. K. (1992). On the relative complexity of passive vs active visual search. *International Journal of Computer Vision*, *7*(2), 127–141.

Tsotsos, J. K., Verghese, G., Dickinson, S., Jenkin, M., Jepson, A., Milios, E., et al. (1998). Playbot a visually-guided robot for physically disabled children. *Image and Vision Computing*, *16*(4), 275–292.

Tsotsos, J. K., Itti, L., & Rees, G. (2005). A brief and selective history of attention. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention*. Amsterdam: Elsevier Press, pp. xxiii - xxxii.

Tsotsos, J. K., & Shubina, K. (2007). Attention and visual search: Active robotic vision systems that search. *Proceedings of The 5th international conference on computer vision systems*, pp. 21–24.

Tsotsos, J. K. (2011). *A computational perspective on visual attention*. Cambridge, MA: MIT Press.

Yang, Y., Fermüller, C., & Aloimonos, Y. (2013). Detection of manipulation action consequences (MAC). Proceedings of the 2013 IEEE conference on computer vision and pattern recognition (pp. 2563–2570). Portland, OR: IEEE.

Yang, Y., Fermuller, C., Aloimonos, Y., & Aksoy, E. E. (2015). *Learning the semantics of manipulation action the 53rd annual meeting of the association for computational linguistics*. Beijing: ACL.

Yang, Y., Guha, A., Fermuller, C., & Aloimonos, Y. (2014). A cognitive system for understanding human manipulation actions. *Advances in Cognitive Sysytems*, *3*, 67–86.

Yang, Y., Li, Y., Fermuller, C. & Aloimonos, Y. (2015). Robot learning manipulation action plans by "Watching" unconstrained videos from the world wide web, the twenty-ninth AAAI conference on artificial intelligence.

Ye Y., & Tsotsos, J. K. (1995). Where to look next in 3D object search. In Proceedings of international symposium on computer vision, pp 539–544.

Ye, Y., & Tsotsos, J. K. (1996). 3D Sensor planning: Its formulation and complexity. In Kautz, H. & Selman, B. (Eds.), Proceedings on 4th international symposium on artificial intelligence and mathematics. Fort Lauderdale, FL.

Ye, Y., & Tsotsos, J. K. (1999). Sensor planning for object search. *Computer Vision and Image Understanding*, *73*(2), 145–168.

Ye, Y., & Tsotsos, J. K. (2001). A complexity-level analysis of the sensor planning task for object search. *Computational Intelligence*, *17*(4), 605–620.

Yu, X., Fermüller, C., Teo, C. L., Yang, Y. & Aloimonos, Y. (2011). Active scene recognition with vision and language. *IEEE International Conference on Computer Vision (ICCV)*, pp 810–817.

Vernon, D. (2008). Cognitive vision: The case for embodied perception. *Image and Vision Computing*, *26*(1), 127–140.

Weng, J. (2003). Developmental Robots. *International Journal of Humanoid Robots*, *1*(2), 109–128.

Wade, N. J., & Wade, N. (2000). *A natural history of vision*. Cambridge: MIT Press.

Wilkes, D., & Tsotsos, J. K. (1992). Active object recognition. In Proceedings of computer vision and pattern recognition, pp. 136–141.

Wörgötter, F., Aksoy, E. E., Kruger, N., Piater, J., Ude, A., & Tamosiunaite, M. (2012). A simple ontology of manipulation actions based on hand-object relations. *IEEE Transactions on Autonomous Mental Development*, *1*, 117–134.

Worch, J.-H., Balint-Benczedi, F., and Beetz, M. (2015). Perception for Everyday Human Robot Interaction,KI - Kunstliche Intelligenz, pp.1-7,Springer Berlin-Heidelberg.

Zampogiannis, K., Yang, Y., Fermuller, C., & Aloimonos, Y. (2015). Learning the spatial semantics of manipulation actions through preposition grounding. In *2015 IEEE international conference on robotics and automation (ICRA)* (pp. 1389–1396). IEEE.

Ruzena Bajcsy received the Master's and Ph.D. degrees in electrical engineering from Slovak Technical University, Bratislava, Slovak Republic, in 1957 and 1967, respectively, and the Ph.D. in computer science from Stanford University, Stanford, CA, in 1972. From 1972 until 2001 she was a professor of Computer and Information Science department at the University of Pennsylvania, Philadelphia. Here she established in 1978 an interdisciplinary robotics laboratory, named GRASP lab .GRASP stands for General Robotics and Active Perception. As the director of the GRASP lab she fostered interdisciplinary research activities and attracted faculty from Electrical and mechanical Engineering as well as from Psychology/Cognitive Science and of course Computer Science. For 28 years while at UPENN she worked on Robotics research, including Computer Vision, tactile perception and in general problem of system identification. In addition she also worked on medical imaging, and developed with her students digital anatomy atlas coupled with elastic matching algorithms that enabled to automatically identify anatomic structures of the brain,first in X-ray tomography, later MRI, and positron image tomography. This technology is now a standard in medical practice. In 2001 she accepted the position to become the founding director of the Center for Information Technology Research in the Interest of Science (CITRIS) and Professor-

ship of Electrical Engineering and Computer Sciences at the University of California, Berkeley. CITRIS is a multi-campus organization comprised of 4 campuses: UC Berkeley, UC Davis, UC Santa Cruz and UC Merced. As part of her activities in CITRIS, she together with the University California Center for Humanities, played a founding role in establishing program of Digital Humanities. Prior to joining Berkeley, she headed the Computer and Information Science and Engineering Directorate at the National Science Foundation between 1999 and 2001. Dr. Bajcsy is a member of the National Academy of Engineering and the National Academy of Science Institute of Medicine as well as a Fellow of the Association for Computing Machinery (ACM) and the American Association for Artificial Intelligence. In 2001, she received the ACM/Association for the Advancement of Artificial Intelligence Allen Newell Award, and was named as one of the 50 most important women in science in the November 2002 issue of Discover Magazine. She is the recipient of the Benjamin Franklin Medal for Computer and Cognitive Sciences (2009) and the IEEE Robotics and Automation Award (2013) for her contributions in the field of robotics and automation. Her current research is in the use of robotic technology, namely measuring and extracting noninvasively kinematic and dynamic parameters of individual in order to assess the physical movement capabilities or limitations. If there are limitations, her students have designed assistive devices that can compensate for the lack of kinematic agility and /or physical strength.

Yiannis Aloimonos (Ph.D 1987, University of Rochester, NY) is Professor of Computational Vision at the Department of Computer Science in the University of Maryland at College Park and the Director of the Computer Vision Laboratory in the Institute for Advanced Computer Studies (UMIACS). He is affiliated with the Department of Electrical and Computer Engineering (ECE), the Institute for Systems Research (ISR) and is a founding member of the Maryland Center for Robotics and the NACS (Neural and Cognitive Science) Program. His work is devoted to Active Perception and Visual Motion Analysis and the study of vision as a dynamic process that serves control (motor system) and reasoning (cognitive system). In the past five years he has been working on coupling attention and segmentation and developing algorithms that segment and recognize objects and actions from video data. With funding from the European Union, under the project POETICON (www.poeticon.eu) he has been developing attention systems for cognitive robotics that integrate perception with symbolic knowledge (language). With funding from the National Science Foundation, under the project ROBOTS WITH VISION THAT FIND OBJECTS he has been developing algorithms for task based attention and search for specific and generic objects. With funding from the NIH he has been developing HAL, a Human Activity Language Tool that can be used to represent human actions.

John Tsotsos is Distinguished Research Professor of Vision Science and Director of the Centre for Innovation in Computing at Lassonde at York University. He received his doctorate in Computer Science from the University of Toronto and then joined the University of Toronto on faculty in Computer Science and in Medicine. In 1980 he founded the highly respected Computer Vision Group at the University of Toronto, which he led for 20 years. He moved to York University in 2000 as Director of the Centre for Vision Research. He has been a Fellow of the Canadian Institute for Advanced Research (1985-95) and currently holds the Canada Research Chair in Computational Vision (2003-2017). He has received many awards and honours including the 2006 Canadian Image Processing and Pattern Recognition Society Award for Research Excellence and Service, the 1st President's Research Excellence Award by York University on the occasion of the University's 50th anniversary in 2009, and the 2011 Geoffrey J. Burton Memorial Lectureship. He was elected as Fellow of the Royal Society of Canada in 2010, in the Division of Mathematical and Physical Sciences of the Academy of Science, and was awarded its 2015 Sir John William Dawson Medal for sustained excellence in multidisciplinary research, the first computer scientist to be so honoured. His current research focuses on a comprehensive theory of visual attention in humans. A practical outlet for this theory forms a second focus, embodying elements of the theory into the vision systems of mobile robots.