Reinforcing Robot Perception of Multi-Modal Events through Repetition and Redundancy and Repetition and Redundancy

Paul Fitzpatrick, Artur Arsenio and Eduardo R. Torres-Jara
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

October 11, 2004

Abstract

For a robot to be capable of development, it must be able to explore its environment and learn from its experiences. It must find (or create) opportunities to experience the unfamiliar in ways that reveal properties valid beyond the immediate context. In this paper, we develop a novel method for using the rhythm of everyday actions as a basis for identifying the characteristic appearance and sounds associated with objects, people, and the robot itself. Our approach is to identify and segment groups of signals in individual modalities (sight, hearing, and proprioception) based on their rhythmic variation, then to identify and bind causally-related groups of signals across different modalities. By including proprioception as a modality, this cross-modal binding method applies to the robot itself, and we report a series of experiments in which the robot learns about the characteristics of its own body.

1 Introduction

To robots and young infants, the world is a puzzling place, a confusion of sights and sounds. But buried in the noise there are hints of regularity. Some of this is natural; for example, objects tend to go *thud* when they fall over and hit the ground. Some is due to the child; for example, if it shakes its limbs in joy or distress, and one of them happens to pass in front of its face, it will see a fleshy blob moving in a familiar rhythm. And some of the regularity is due to the efforts of a caregiver; consider an infant's mother trying to help her child learn and develop, perhaps by tapping a toy or a part of the child's body (such as its hand) while speaking its name, or making a toy's characteristic sound (such as the *bang-bang* of a hammer).

Humans receive an enormous quantity of information from the world through their sensorial apparatus. Cues from the visual, auditory and somatosensory senses, as well as from tactile and smell senses, are processed simultaneously, and the integration of all such percepts at the brain's cerebral cortex forms our view of the world.

However, humans' sensory modalities are not independent processes. Stimuli from one sensorial modality often influences the perception of stimuli in other modalities. Auditory processing in visual brain areas of early blind subjects suggests that brain areas usually involved in vision play a role in not only auditory selective attention, but also participate in processing changes on the auditory stimulus outside the focus of attention (Alho et al., 1993). Auditory illusions can be created from visual percepts as well – one such instance is the McGurk effect (Cohen and Massaro, 1990).

But audition can also cause illusory visual motion, as described by (Churchland et al., 1994). They report an experiment in which a fixed square and a dot (to its left) are presented to the observer. Without sound stimuli, no motion is perceived for blinking of the dot. Alternate perception of a tone in the left and right ears (left ear tone coinciding with the dot presentation), creates an illusory perception of oscillatory motion of the dot (while the square creates visual occlusions).

In this paper we seek to extract useful information from repeated actions performed either by a caregiver or the robot itself. Observation of infants shows that such actions happen frequently, and from a computational perspective they are ideal learning material since they are easy to identify and offer a wealth of redundancy (important for robustness). The information we seek from repeated actions are the characteristic appearances and sounds of the object, person, or robot involved, with context-dependent information such as the visual background or unrelated sounds stripped away. This allows the robot to generalize its experience beyond its immediate context and, for example, later recognize the same object used in a different way.

We wish our system to be scalable, so that it can correlate and integrate multiple sensor modalities (currently sight, sound, and proprioception). To that end, we detect and cluster periodic signals within their individual modalities, and only then look for cross-modal relationships between such signals. This avoids a combinatorial explosion of comparisons, and means our system can be gracefully extended to deal with new sensor modalities in future (touch, smell, etc).

This paper begins by introducing previous work, our robotic platform and what it can sense in this section's remaining. We then introduce the methods we use for detecting regularity in individual modalities and the tests applied to determine when to 'bind' features in different modalities together. The remainder (and larger part) of the paper presents experiments where the robot socially interacts with a human. The robot detects regularity: in objects and people it encounters through matching acoustic beats to visual motions; in object textures, by having a person touching the texture to get its spectral energy; and in itself, associating proprioception with other sensorial data. Finally, we consider what progress is possible when repetition is not available, in the context of reaching and grasping (which are typically done once-off, rather than repeatedly).

1.1 The development of intermodal perception in infants

Infants are not born perceiving the world as an adult does; rather, their perceptual abilities develop over time. This process is of considerable interest to roboticists who seek hints on how to approach adult-level competence through incremental steps. Vision and audition interact from birth (Wertheimer, 1961). Indeed, a ten-minute-old child turns his eyes toward an auditory signal. In the animal kingdom, studies with young owls have shown that development of sound localization has strong influences from the visual senses. Inducing visual errors from prisms worn over the eyes, owls adjusted their sound localization to match the visual bias (Knudsen and Knudsen, 1985). Historically, the development of perception in infants has been described using two diametrically opposed classes of theory: integration and differentiation (Bahrick, 2003). In a theory of integration, the infant learns to process its individual senses first, and then begins to relate them to each other. In a theory of differentiation, the infant is born with unified senses, which it learns to differentiate between over time. The weight of empirical evidence supports a more nuanced position (as is usually the case with such dichotomies). On the one hand, young infants can detect certain intersensory relationships very early (Lewkowicz and Turkewitz, 1980) – but on the other hand, there is a clear progression over time in the kinds of relations which can be perceived (Lewkowicz (Lewkowicz, 2000) gives a timeline).

Time is a very basic property of events that gets encoded across the different senses but is unique to none of them. Consider a bouncing ball – the audible thud of the ball hitting the floor happens at the same time as a dramatic visual change in direction. Although the acoustic and visual aspects of the bounce may be very different in nature and hard to relate to each other, the time at which they make a gross change is comparable. The time of occurrence of an event is an *amodal* property – a property that is more or less independent of the sense with which it is perceived. Other such properties include intensity, shape, texture, and location; these contrast with properties that are relatively modality-specific such as color, pitch, and smell (Lewkowicz, 2003).

Time can manifest itself in many forms, from simple synchronicity to complex rhythms. Lewkowicz proposes that the sensitivity of infants to temporal relationships across the senses develops in a progression of more complex forms, with each new form depending on earlier ones (Lewkowicz, 2000). In particular, Lewkowicz suggests that sensitivity to *synchronicity* comes first, then to *duration*, then to *rate*, then to *rhythm*. Each step relies on the previous one initially. For example, duration is first established as the time between the synchronous beginning and the synchronous end of an event as perceived in multiple senses, and only later does duration break free of its origins to become a temporal relation in its own right that doesn't necessarily require synchronicity.

Bahrick (Bahrick, 2004) proposes that the perception of the same property across multiple senses (intersensory redundancy) can aid in the initial learning of skills which can then be applied even without that redundancy. For example, in one experiment (Bahrick and Lickliter, 2000) infants exposed to a complex rhythm tapped out by a hammer presented both visually and acoustically can then discriminate that rhythm in either modality alone – but if the rhythm is initially presented in just one modality, it cannot be discriminated in either (for infants of a given age). The suggested explanation is that intersensory redundancy helps to direct attention towards amodal properties (in this case, rhythm) and away from mode-specific properties. In general, intersensory redundancy has a significant impact on attention, and can bias figure/ground judgements (Bahrick, 2004). Another experiment (Hernandez-Reif and Bahrick, 2001) provides evidence that an amodal relation (in this case texture, which is common to visual and tactile sensing) provides a basis for learning arbitrary relations between modality-specific properties (in this case the particular colored surface of a textured object).

Such results and theories are very relevant to robotics. For an autonomous robot to be capable of developing and adapting to its environment, it needs to be able to learn. The field of machine learning offers many powerful algorithms, but these require training data to operate. Infant development research suggests ways to acquire such training data from simple contexts, and use this experience to bootstrap to more complex contexts. We need to identify situations that enable the robot to temporarily reach beyond its current perceptual abilities, giving the opportunity for development to occur (Fitzpatrick, 2003b). An example of this in the robotic domain is the active segmentation system implemented previously on Cog, where the robot initially needed to come into physical contact with objects before it could learn about them or recognize them, since it used the contingent motion of the objects to segment them from the background, but after this familiarization period it could recognize objects without further contact. In this paper, we exploit repetition – rhythmic motion, repeated sounds – to achieve segmentation and recognition across multiple senses.

1.2 Platform and percepts

This work is implemented on the humanoid robot Cog (Brooks et al., 1999). Cog has an active vision head, two six-degree of freedom arms, a rotating torso, and a microphone array arranged along its shoulders. For this paper, we work with visual input from one of Cog's four cameras, acoustic input from the microphone array, and proprioceptive feedback from joints in the head, torso, and arms.

Figure 1 shows how the robot's perceptual state can be summarized – the icons shown here will be used throughout the paper. The robot can detect periodic events in any of the individual modalities (sight, hearing, proprioception). Any two events that occur in different modalities will be compared, and may be grouped together if there is evidence that they are causally related or *bound*. Such relations are transitive: if events A and B are bound to each other, and B and C are bound to each other, then A and C will also be bound. This is important for consistent, unified perception of events.

This kind of summarization ignores cases in which there are, for example, multiple visible objects moving periodically making different sounds. We return to this point later in the paper. We have previously demonstrated that our system can deal well with multiple-binding cases, since it performs segmentation in the individual modalities (Arsenio and Fitzpatrick, 2003). For this paper, there is no real need to consider such cases, since we don't expect the robot's caregiver to maliciously introduce distractors into its environment – but nevertheless it is an important feature of our algorithm, which we now present.

2 Detecting periodic events

We are interested in detecting conditions that repeat with some roughly constant rate, where that rate is consistent with what a human can easily produce and perceive. This is not a very well defined range, but we will consider anything above 10Hz to be too fast, and anything below 0.1Hz to be too slow. Repetitive

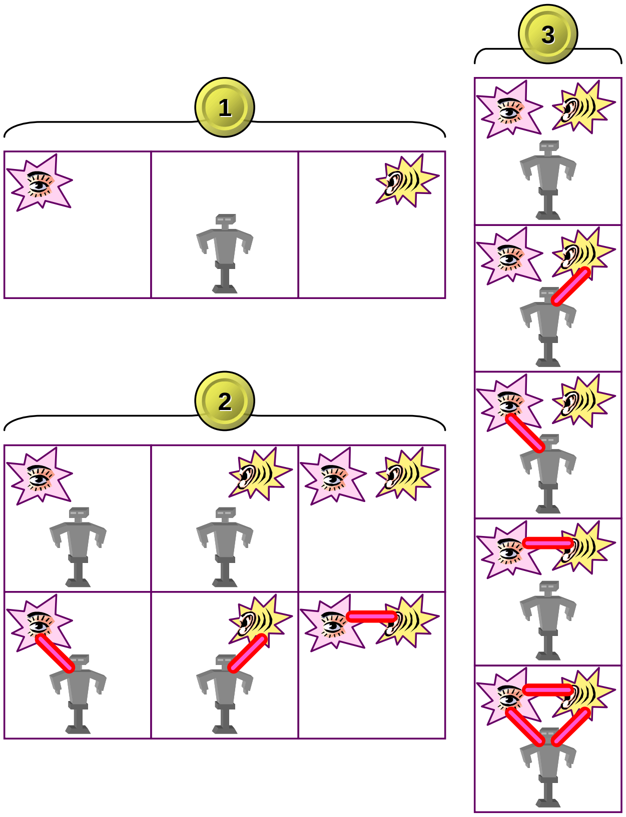

Figure 1: A summary of the possible perceptual states of our robot – the representation shown here will be used throughout the paper. Events in any one of the three modalities (sight, proprioception, or hearing) are indicated as in block **1**. When two events occur in different modalities, they may be independent (top of **2**) or bound (bottom of **2**). When events occur in three modalities, the possibilities are as shown in **3**.

signals in this range are considered to be *events* in our system. For example, waving a flag is an event, clapping is an event, walking is an event, but the vibration of a violin string is not an event (too fast), and neither is the daily rise and fall of the sun (too slow). Such a restriction is related to the idea of natural kinds (Hendriks-Jansen, 1996), where perception is based on the physical dimensions and practical interests of the observer.

To find periodicity in signals, the most obvious approach is to use some version of the Fourier transform. And indeed our experience is that use of the Short-Time Fourier Transform (STFT) demonstrates good performance when applied to the visual trajectory of periodically moving objects (Arsenio et al., 2003). For example, Figure 2 shows a hammer segmented visually by tracking and grouping periodically moving points. However, our experience also leads us to believe that this approach is not ideal for detecting periodicity of *acoustic* signals. Of course, acoustic signals have a rich structure around and above the kHz range, for which the Fourier transform and related transforms are very useful. But detecting gross repetition around the single Hz range is very different. The sound generated by a moving object can be quite complicated, since any constraints due to inertia or continuity are much weaker than for the physical trajectory of a mass moving through space. In our experiments, we find that acoustic signals may vary considerably in amplitude between repetitions, and that there is significant variability or drift in the length of the periods. These two properties combine to reduce the efficacy of Fourier analysis. This led us to the development of a more robust method for periodicity detection, which is now described. In the following discussion, the term *signal* refers to some sensor reading or derived measurement, as described at the end of this section. The term *period* is

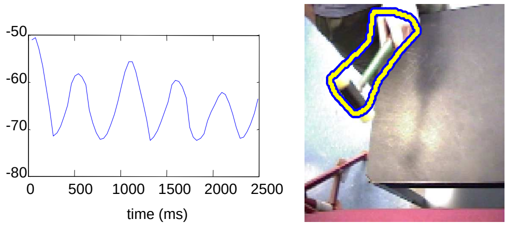

Figure 2: When watching a person using a hammer, the robot detects and group points moving in the image with similar periodicity (Arsenio et al., 2003) to find the overall trajectory of the hammer and separate it out from the background. The detected trajectory is shown on the left (for clarity, just the coordinate in the direction of maximum variation is plotted), and the detected object boundary is overlaid on the image on the right.

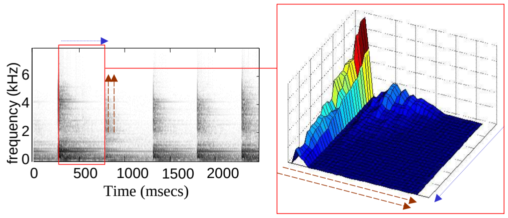

Figure 3: Extraction of an acoustic pattern from a periodic sound (a hammer banging). The algorithm for signal segmentation is applied to each normalized frequency band. The box on the right shows one complete segmented period of the signal. Time and frequency axes are labeled with single and double arrows respectively.

used strictly to describe event-scale repetition (in the Hz range), as opposed to acoustic-scale oscillation (in the kHz range).

Period estimation – For every sample of the signal, we determine how long it takes for the signal to return to the same value from the same direction (increasing or decreasing), if it ever does. For this comparison, signal values are quantizing adaptively into discrete ranges. Intervals are computed in one pass using a look-up table that, as we scan through the signal, stores the time of the last occurrence of a value/direction pair. The next step is to find the most common interval using a histogram (which requires quantization of interval values), giving us an initial estimate $p_{estimate}$ for the event period. This is essentially the approach presented in (Arsenio and Fitzpatrick, 2003). For the work presented in this paper, we extended this method to explicitly take into account the possibility of drift and variability in the period, as follows.

Clustering – The previous procedure gives us an estimate $p_{estimate}$ of the event period. We now cluster samples in rising and falling intervals of the signal, using that estimate to limit the width of our clusters but not to constrain the distance between clusters. This is a good match with real signals we see that are generated from human action, where the periodicity is rarely very precise. Clustering is performed individually for each of the quantized ranges and directions (increasing or decreasing), and then combined afterwards. Starting from the first signal sample not assigned to a cluster, our algorithm runs iteratively until all samples are assigned, creating new clusters as necessary. A signal sample extracted at time t is assigned to a cluster with center c_i if $\| c_i - t \|_2 < p_{estimate}/2$. The cluster center is the average time coordinate of the samples assigned to it, weighted according to their values.

Merging – Clusters from different quantized ranges and directions are merged into a single cluster if $\| c_i - c_j \|_2 < p_{estimate}/2$ where c_i and c_j are the cluster centers.

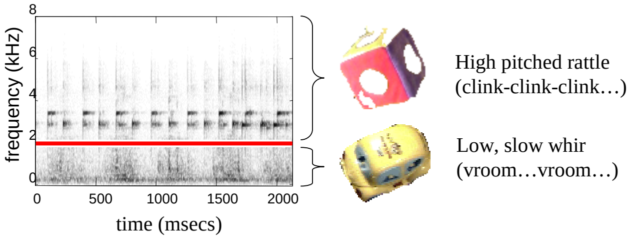

Figure 4: Results of an experiment in which the robot could see a car and a cube, and both objects were moving – the car was being pushed back and forth on a table, while the cube was being shaken (it has a rattle inside). By comparing periodicity information, the high-pitched rattle sound and the low-pitched *vroom* sound were distinguished and bound to the appropriate object, as shown on the spectrogram. The object segmentations shown were automatically determined.

Segmentation – We find the average interval between neighboring cluster centers for positive and negative derivatives, and break the signal into discrete periods based on these centers. Notice that we do not rely on an assumption of a *constant* period for segmenting the signal into repeating units. The average interval is the final estimate of the signal period.

The output of this entire process is an estimate of the period of the signal, a segmentation of the signal into repeating units, and a confidence value that reflects how periodic the signal really is. The period estimation process is applied at multiple temporal scales. If a strong periodicity is not found at the default time scale, the time window is split in two and the procedure is repeated for each half. This constitutes a flexible compromise between both the time and frequency based views of a signal: a particular movement might not appear periodic when viewed over a long time interval, but may appear as such at a finer scale.

Figure 2 shows an example of using periodicity to visual segment a hammer as a human demonstrates the periodic task of hammering, while Figure 3 shows segmentation of the sound of the hammer in the time-domain. Segmentation in the frequency-domain was demonstrated in (Arsenio and Fitzpatrick, 2003) and is illustrated in Figure 4). For these examples and all other experiments described in this paper, our system tracks moving pixels in a sequence of images from one of the robot's cameras using a multiple tracking algorithm based on a pyramidal implementation of the Lukas-Kanade algorithm. A microphone array samples the sounds around the robot at 16kHz. The Fourier transform of this signal is taken with a window size of 512 samples and a repetition rate of 31.25Hz. The Fourier coefficients are grouped into a set of frequency bands for the purpose of further analysis, along with the overall energy.

3 See The Beat

Segmented features extracted from visual and acoustic segmentations (using the method presented in last section) can serve as the basis for an object recognition system. In the visual domain, (Fitzpatrick, 2003a) used segmentations derived through physical contact as an opportunity for a robot to become familiar with the appearance of objects in its environment and grow to recognize them. (Krotkov et al., 1996) has looked at recognition of the sound generated by a single contact event. Visual and acoustic cues are both individually important for recognizing objects, and can complement each other when, for example, the robot hears an object that is outside its view, or it sees an object at rest. But when both visual and acoustic cues are present, then we can do even better by looking at the relationship between the visual motion of an object and the sound it generates. Is there a loud bang at an extreme of the physical trajectory? If so we might be looking at a hammer. Are the bangs at either extreme of the trajectory? Perhaps it is a bell. Such relational features can only be defined and factored into recognition if we can relate or *bind* visual and acoustic signals.

Several theoretical arguments support the idea of binding by temporal oscillatory signal correlations (von der Malsburg, 1995). From a practical perspective, repetitive synchronized events are ideal for learning since

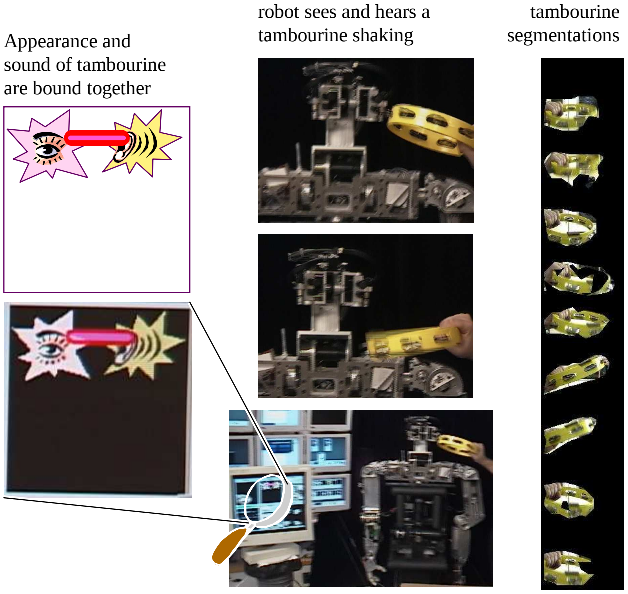
Appearance and sound of tambourine are bound together
robot sees and hears a tambourine shaking
tambourine segmentations

Figure 5: Here the robot is shown a tambourine in use. The robot detects that there is a periodically moving visual source, and a periodic sound source, and that the two sources are causally related and should be bound. All images in these figures are taken directly from recordings of real-time interactions, except for the summary box in the top-left (included since in some cases the recordings are of poor quality). The images on the far right show the visual segmentations recorded for the tambourine in the visual modality. The background behind the tambourine, a light wall with doors and windows, is correctly removed. Acoustic segmentations are generated but not shown (see Figures 3 and 4 for examples).

they provide large quantities of redundant data across multiple sensor modalities. In addition, as already mentioned, extra information is available in periodic or locally-periodic signals such as the period of the signal, and the phase relationship between signals from different senses – so for recognition purposes the whole is greater than the sum of its parts.

Therefore, a binding algorithm was developed to associate cross-modal, locally periodic signals, by which we mean signals that have locally consistent periodicity, but may experience global drift and variation in that rhythm over time. In our system, the detection of periodic cross-modal signals over an interval of seconds using the method described in the previous section is a necessary (but not sufficient) condition for a binding between these signals to take place. We now describe the extra constraints that must be met for binding to occur.

For concreteness assume that we are comparing a visual and acoustic signal. Signals are compared by matching the cluster centers determined as in the previous section. Each peak within a cluster from the visual signal is associated to a temporally close (within a maximum distance of half a visual period) peak from the acoustic signal, so that the sound peak has a positive phase lag relative to the visual peak. Binding occurs if the visual period matches the acoustic one, or if it matches half the acoustic period, within a tolerance of $60ms$. The reason for the second match is that often sound is generated at the fastest points of an object's trajectory, or the extremes of a trajectory, both of which occur twice for every single period of the trajectory. Typically there will be several redundant matches that lead to binding within a window of the sensor data for which several sound/visual peaks were detected. In (Arsenio and Fitzpatrick, 2003), we describe a more sophisticated binding method that can differentiate causally unconnected signals with periods that are similar just by coincidence, by looking for a drift in the phase between the acoustic and visual signal over time, but such nuances are less important in a benign developmental scenario supported by a caregiver.

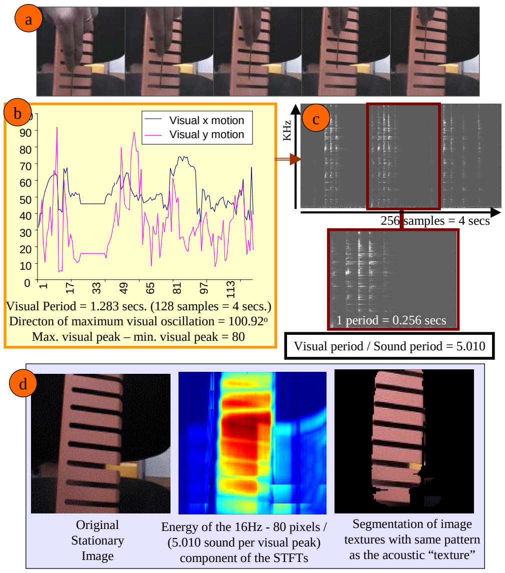

Figure 6: Matching visual/acoustic textures to visual textures. a) Sequence of images showing a human hand playing rhythmic sounds with a textured metal piece. Sound is only produced along one direction of motion. b) Horizontal and vertical visual trajectories of the human hand, during a time window of approximately 4 seconds (128 frames). The visual period is estimated as 1.28 seconds. The amplitude difference between the maximum and minimum trajectory values is 78 pixels. Maximum variation makes a 100.92^o angle with the horizontal. c) Ratio between half the visual period and the sound period is $\simeq 5$, which means that sound peaks five times from a visual minimum location to a maximum. d) Stationary image – left – is segmented using a mask – middle – computed from the $16Hz$ energy component of the STFTs applied at each point, selecting the relevant object's texture – right.

Figure 5 shows an experiment in which a person shook a tambourine in front of the robot for a while. The robot detected the periodic motion of the tambourine, the rhythmic rise and fall of the jangling bells, and bound the two signals together in real-time.

4 Touch The Beat

There is experimental evidence that an amodal relation (in this case texture, which is common to visual and tactile sensing) provides a basis for learning arbitrary relations between modality-specific properties (Hernandez-Reif and Bahrick, 2001) (in this case the particular colored surface of a textured object). This motivated a strategy to extract image textures from visual-sound patterns, i.e., by processing acoustic *textures* (the sound signatures) between visual trajectory peaks. The algorithm works by having a human probe the world by creating rhythmic sounds on a textured, roughed surface. Visual and acoustic textures are then linked as follows:

1. Hand tracking of periodic gestures using the procedure applied in the previous section to learn from educational activities, that selectively attends to the human actuator for the extraction of periodic signals from its trajectory

2. Tracking and mapping of the x and y human hand visual trajectories (horizontal and vertical directions in images, respectively) into coordinates along eigenvectors given by the Singular Value Decomposition, resulting in new axes x_1, y_1 (x_1 corresponding to the eigenvector along highest data variance). Three measures are then estimated:

- The angle β of axis x_1 relative to x
- The visual trajectory period (after trajectory smoothing to reduce noise) by detecting periodicity along the x_1 direction – using the STFT based approach
- The amplitude difference A_v between the maximum trajectory value along x_1 and the minimum value, over one visual period

3. Partition of the acoustic signal according to the visual periodicity, and sound periodic detection applied over multiple scales on such a window (Section 2). The goal is to estimate the spatial frequency F_s of the object's texture in the image (with the highest energy over the spectrum). This is done by computing the number n of acoustic periods during one (or half) visual period. The spatial frequency estimate is then given by $F_s = A_v/n$, which means that the sound peaks n times from a visual minimum location to a maximum (the human is producing sounds with n peaks of energy along the hand trajectory)

4. Spectral processing of a stationary image by applying to each image point a 1-dimensional STFT along the direction of maximum variation, with length given by the trajectory amplitude and window centered on that point, and storing for such point the energy of the F_s component of this transform. This energy image is converted to binary by a threshold given as a percentage of the maximum energy (or a lower bound, whichever is higher). The object is then segmented by applying this mask to the stationary image.

All these steps are demonstrated by the experiment in Figure 6. It shows a human hand playing rhythmic sounds with a textured metal piece (Figure 6-a), producing sound chiefly along the vertical direction of motion. The x and y visual trajectories of the human hand are tracked during a period of approximately 4 seconds (128 frames). The axis x_1 is at an angle of $\beta = 100.92^o$ with the x axis for the experiment shown. Periodic detection along x_1 (after smoothing to reduce noise) estimates a visual period of 1.28 seconds. The visual trajectory's amplitude difference A_v is 78 pixels over one visual period (Figure 6-b). Sound periodic detection is applied on the visually segmented acoustic signal over 2 scales of temporal resolution. For this experiment, the ratio between half the visual period and the sound period is $n \simeq 5$ (Figure 6-c).

The sound presents therefore 5 peaks of energy along the hand trajectory, which corresponds to a frequency of $F_s = 16Hz$. The stationary image in Figure 6-d is processed by selecting 16Hz components of the STFTs, resulting an energy image – middle – which masks the texture which produced such sound. Children toys like xylophones have a similar structure to the metal piece, which motivated this experiment. The algorithm extracted two such templates out of a 2 minute run. This corresponds to about 20% of the total number of templates which were theoretically possible to extract over this time interval. But once again, a conservative approach led to algorithm robustness to errors. It is also worth stressing however that this approach could also be applied by replacing sound with proprioceptive or tactile sensing, and the human action by robotic manipulation.

5 Feel The Beat

So far we have considered only external events that do not involve the robot. In this section we turn to the robot's perception of its own body. Cog treats proprioceptive feedback from its joints as just another sensory modality in which periodic events may occur. These events can be bound to the visual appearance of its moving body part – assuming it is visible – and the sound that the part makes, if any (in fact Cog's arms are quite noisy, making an audible "whirr-whirr" when they move back and forth).

Figure 7 shows a basic binding experiment, in which a person moved Cog's arm while it is out of the robot's view. The sound of the arm and the robot's proprioceptive sense of the arm moving are bound together. This is an important step, since in the busy lab Cog inhabits, people walk into view all the time, and there are frequent loud noises from the neighboring machine shop. So cross-modal rhythm is an important cue for filtering out extraneous noise and events of lesser interest.

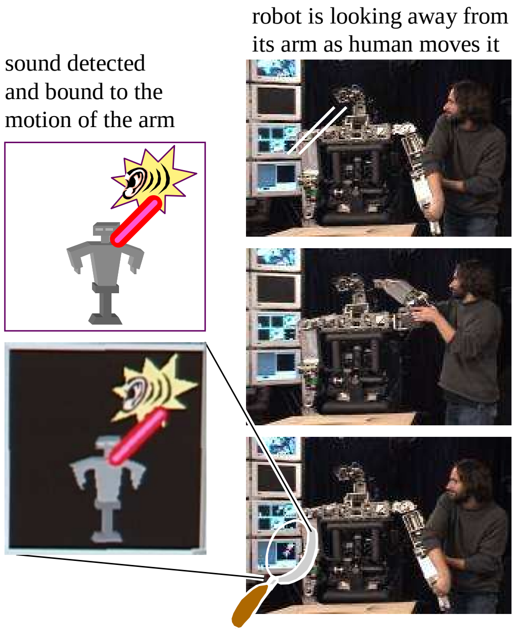

Figure 7: In this experiment, a person grabs Cog's arm and shakes it back and forth while the robot is looking away. The sound of the arm is detected, and found to be causally related to the proprioceptive feedback from the moving joints, and so the robot's internal sense of its arm moving is bound to the external sound of that motion.

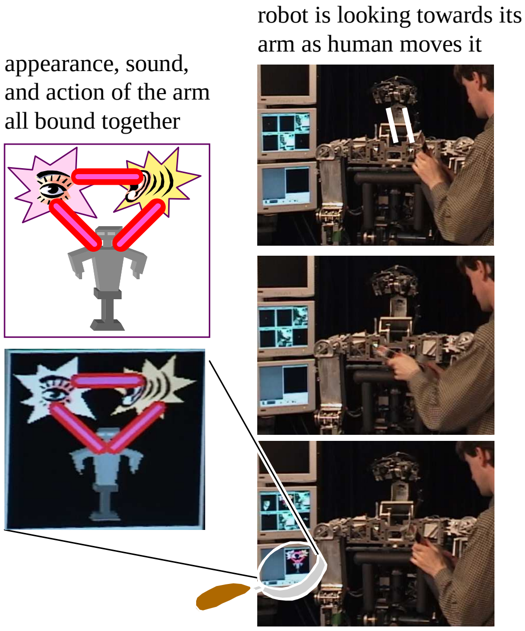

Figure 8: In this experiment, a person shakes Cog's arm in front of its face. What the robot hears and sees has the same rhythm as its own motion, so the robot's internal sense of its arm moving is bound to the sound of that motion and the appearance of the arm.

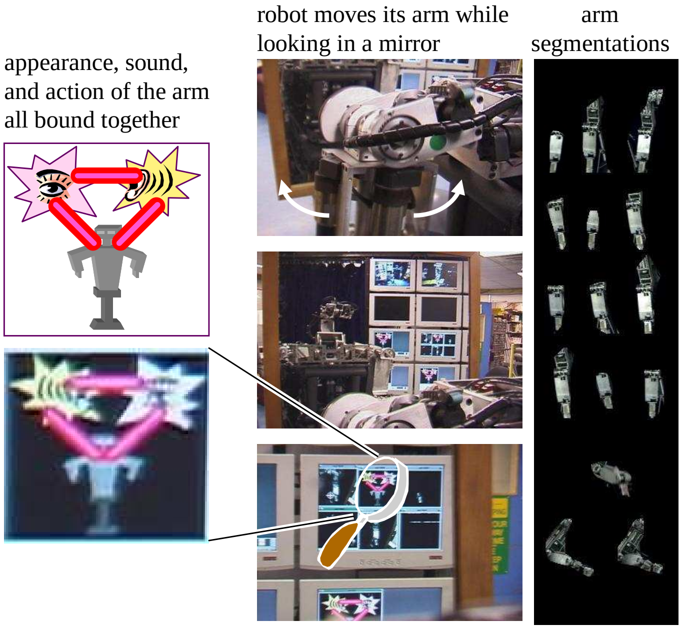
appearance, sound, and action of the arm all bound together
robot moves its arm while looking in a mirror
arm segmentations

Figure 9: In this experiment, Cog is looking at itself in a mirror, while shaking its arm back and forth (the views on the right are taken by a camera behind the robot's left shoulder, looking out with the robot towards the mirror). The reflected image of its arm is bound to the robot's sense of its own motion, and the sound of the motion. This binding is identical in kind to the binding that occurs if the robot sees and hears its own arm moving directly without a mirror. However, the appearance of the arm is from a quite different perspective than Cog's own view of its arm.

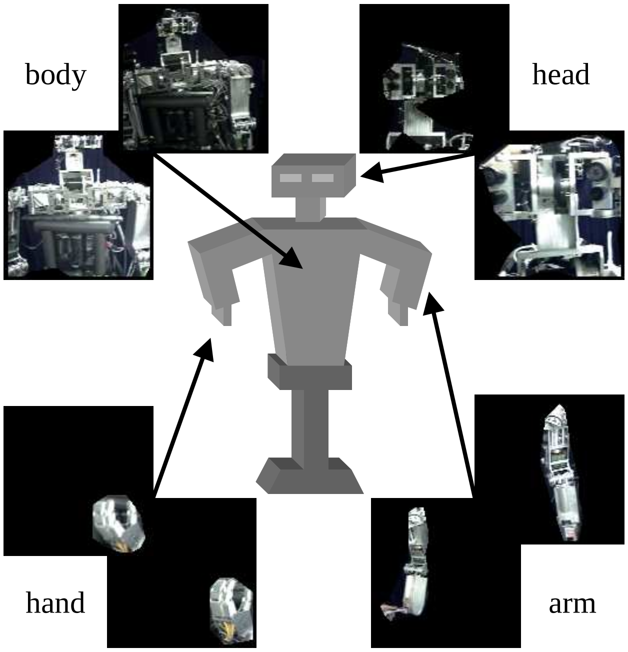

Figure 10: Cog can be shown different parts of its body simply by letting it see that part (in a mirror if necessary) and then shaking it, such as its (right) hand or (left) flipper. Notice that this works for the head, even though shaking the head also affects the cameras.

In Figure 8, the situation is similar, with a person moving the robot's arm, but the robot is now looking at the arm. In this case we see our first example of a binding that spans three modalities: sight, hearing, and proprioception. The same is true in Figure 9, where Cog shakes its own arm while watching it in a mirror. This idea is related to work in (Metta and Fitzpatrick, 2003), where Cog located its arm by shaking it.

An important milestone in child development is reached when the child recognizes itself as an individual, and identifies its mirror image as belonging to itself (Rochat and Striano, 2002). Self-recognition in a mirror is also the focus of extensive study in biology. Work on self-recognition in mirrors for chimpanzees (Gallup et al., 2002) suggests that animals other than humans can also achieve such competency, although the interpretation of such results requires care and remains controversial. Self-recognition is related to the notion of a theory-of-mind, where intents are assigned to other actors, perhaps by mapping them onto oneself, a topic of great interest in robotics (Kozima and Yano, 2001; Scassellati, 2001). Proprioceptive feedback provides very useful reference signals to identify appearances of the robot's body in different modalities. That is why we extended our binding algorithm to include proprioceptive data.

Children between 12 and 18 months of age become interested in and attracted to their reflection (American Academy Of Pediatrics, 1998). Such behavior requires the integration of visual cues from the mirror with proprioceptive cues from the child's body. As shown in Figure 10, the binding algorithm was used not only to identify the robot's own acoustic rhythms, but also to identify visually the robot's mirror image (an important milestone in the development of a child's theory of mind (Baron-Cohen, 1995)). It is important to stress that we are dealing with the low-level *perceptual* challenges of a theory of mind approach, rather than the high-level *inferences* and mappings involved. Correlations of the kind we are making available could form a grounding for a theory of mind and body-mapping, but are not of themselves part of a theory of mind – for example, they are completely unrelated to the intent of the robot or the people around it, and intent is key to understanding others in terms of the self (Kozima and Zlatev, 2000; Kozima and Yano, 2001). Our hope is that the perceptual and cognitive research will ultimately merge and give a truly intentional robot that understands others in terms of its own goals and body image – an image which could develop incrementally using cross-modal correlations of the kind explored in this paper.

6 Beyond Periodicity

In all our work so far, we have benefited from two sources of redundancy: multimodal cues and repetition. This redundancy provides robustness to noise and error. But we cannot always expect to have repetition. What can we do without it to combat noise? We can improve the accuracy of our sensors, we can improve our algorithms, and we can *find more cues*.

It is quite clear that human perception makes use of large numbers of redundant cues, even within a single sensory modality. Consider depth perception. Many cues play a role in the human perception of depth, including occlusions, stereopsis, motion parallax, shadows (attached or cast), interreflections, perspective effects, etc. The need to understand human perception has grown in urgency as researchers in computer graphics and virtual reality strive for increasing realism. The different cues can reinforce each other, with their individual effectiveness varying with circumstance. For example, stereopsis has proven to be an important depth cue, followed closely by shadows and interreflections (Hubona et al., 1999; Hu et al., 2000, 2002; Hubona et al., 2004).

Compared to human perception, robots are very much empoverished. We have begun work to expand the range of cues available for robot perception, starting with perception of its own arm. In this paper, we have already shown how the robot's sense of its own motion can be bound to the visual appearance and sounds associated with the arm. This has the advantage of not requiring strong prior models of the arm, and essentially allowing self discovery. One difficulty with this approach is that a shadow of the robot's arm can occasionally get confused with the arm itself, since it moves in a correlated manner. We have now begun to exploit this as an additional, valuable cue, to reinforce the robot's perception of its arm (Fitzpatrick and Torres-Jara, 2004). For this work we have moved to a new robot called *Coco*. We are inspired by speculation that the cast shadow of the hand may be integrated into the body schema, in a manner somewhat analogous to the extension of the schema to include tools (Pavani and Castiello, 2004).

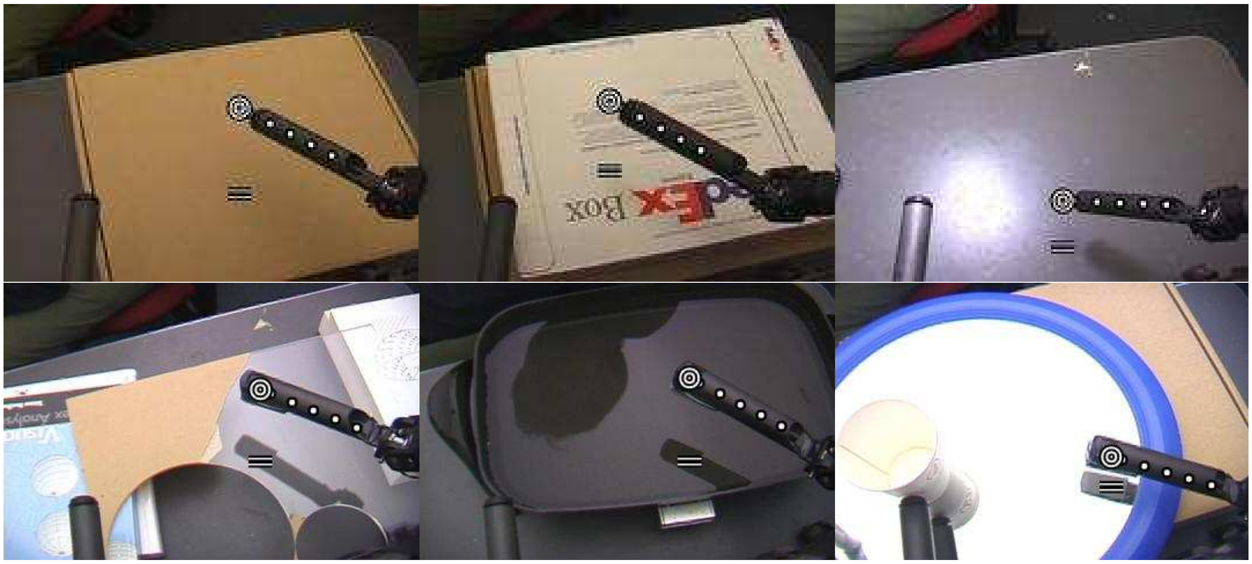

Figure 11: In our work on reaching, we detect the endpoint of the arm (circle and white dots) and the shadow it throws (bar) using motion cues (Fitzpatrick and Torres-Jara, 2004) Changes in the distance to the shadow (or reflection) of the robot arm are interpreted as changes in the distance to the surface.

As the arm approaches a surface, the shadow that it casts rushes to meet it. This gives an indication of the arm's proximity to the surface. We track this approach and estimate a "time-to-contact" to predict when touch will occur. This is analogous to the time-to-contact quantity derived from optic flow for navigation (Lee, 1980; Gibson, 1986). We use experiments in which touching occurs (detected by a tactile sensor on the endpoint) to train this prediction. With the shadow cue available, the robot can direct its arm to touch a target using information from a *single* camera. This is done by driving the arm to simultaneously reduce the visual error between the endpoint and the target and the visual error between the endpoint's shadow and the target. When both these errors are zero, the endpoint is at the target in 3D space. There are advantages to using shadows instead of conventional stereo vision. Shadows and stereopsis have somewhat complementary properties; stereo is at is best when depth changes are sharp, while shadows are easiest to track when depth changes are relatively smooth. The error in stereo measurements grows with distance from the camera, while the error in shadow measurements grows with distance from the surface. Shadows (and reflections) are detectable even in the absence of texture, or with reflective surfaces, situations that can confuse stereo. We believe that combining stereopsis and shadows could lead to a more robust system for manipulation.

In this case, we have looked at redundancy *within* a sensory modality (vision) rather than just between modalities. Even for shadows, we are just at the beginning of using the redundancy available; we use the coarse movement of the shadow, and ignore its sharpness, size, etc. Now that our robots can have reasonable computational resources available for perception, the time is ripe to start investigating what cues we can add, at the same time as we improve the performance of algorithms for existing cues.

7 Discussion and conclusions

Most of us have had the experience of feeling a tool become an extension of ourselves as we use it (see (Stoytchev, 2003) for a literature review). Many of us have played with mirror-based games that distort or invert our view of our own arm, and found that we stop thinking of our own arm and quickly adopt the new distorted arm as our own. About the only form of distortion that can break this sense of ownership is a delay between our movement and the proxy-arm's movement. Such experiences argue for a sense of self that is very robust to every kind of transformation except latencies. Our work is an effort to build a perceptual

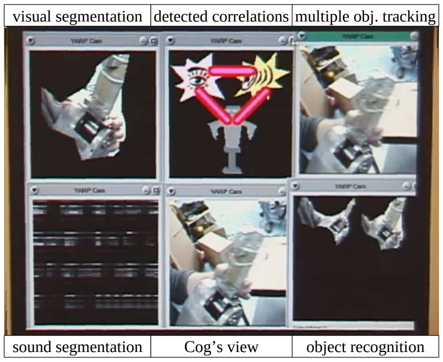

Figure 12: This figure shows a real-time view of the robot's status during the experiment in Figure 8. The robot is continually collecting visual and auditory segmentations, and checking for cross-model events. It also compares the current view with its database and performs object recognition to correlate with past experience (bottom right).

system which, from the ground up, focuses on timing just as much as content. This is powerful because timing is truly cross-modal, and leaves its mark on all the robot's senses, no matter how they are processed and transformed.

We are motivated by evidence from human perception that strongly suggests that timing information can transfer between the senses in profound ways. For example, experiments show that if a short fragment of white noise is recorded and played repeatedly, a listener will be able to hear its periodicity. But as the fragment is made longer, at some point this ability is lost. But the repetition can be heard for far longer fragments if a light is flashed in synchrony with it (Bashford et al., 1993) – flashing the light actually changes how the noise sounds. More generally, there is evidence that the cues used to detect periodicity can be quite subtle and adaptive (Kaernbach, 1993), suggesting there is a lot of potential for progress in replicating this ability beyond the ideas already described.

Although there is much to do, from a practical perspective a lot has already been accomplished. Consider Figure 12, which shows a partial snapshot of the robot's state during one of the experiments described in the paper. The robot's experience of an event is rich, with many visual and acoustic segmentations generated as the event continues, relevant prior segmentations recalled using object recognition, and the relationship between data from different senses detected and stored. We believe that this kind of experience will form one important part of a perceptual toolbox for autonomous development, where many very good ideas have been hampered by the difficulty of robust perception.

Another ongoing line of research we are pursuing is truly cross-modal object recognition. A hammer causes sound after striking an object. A toy truck causes sound while moving rapidly with wheels spinning; it is quiet when changing direction – therefore, the car's acoustic frequency is twice as much as the frequency of its visual trajectory. A bell typically causes sound at either extreme of motion. All these statements are truly cross-modal in nature, and with our system we can begin to use such properties for recognition.

Acknowledgements

Project funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement. Arsenio was supported by Portuguese grant PRAXIS XXI BD/15851/98.

References

Alho, K., Kujala, T., Paavilainen, P., Summala, H., and Naatanen, R. (1993). Auditory processing in visual brain areas of the early blind: evidence from event-related potentials. *Electroencephalogr Clin Neurophysiol*, 86(6):418–27.

American Academy Of Pediatrics (1998). *Caring for Your Baby and Young Child: Birth to Age 5*. Bantham.

Arsenio, A. and Fitzpatrick, P. (2003). Exploiting cross-modal rhythm for robot perception of objects. In *Proceedings of the Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems*, Singapore.

Arsenio, A., Fitzpatrick, P., Kemp, C. C., and Metta, G. (2003). The whole world in your hand: Active and interactive segmentation. Proceedings of the Third International Workshop on Epigenetic Robotics.

Bahrick, L. E. (2003). Development of intermodal perception. In Nadel, L., editor, *Encyclopedia of Cognitive Science*, volume 2, pages 614–617. Nature Publishing Group, London.

Bahrick, L. E. (2004). The development of perception in a multimodal environment. In Bremner, G. and Slater, A., editors, *Theories of infant development*, pages 90–120. Blackwell Publishing, Malden, MA.

Bahrick, L. E. and Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36:190–201.

Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.

Bashford, J. A., Brubaker, B. S., and Warren, R. M. (1993). Cross-modal enhancement of repetition detection for very long period recycling frozen noise. *Journal of the Acoustical Soc. of Am.*, 93(4):2315.

Brooks, R. A., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M. M. (1999). The Cog project: Building a humanoid robot. In Nehaniv, C. L., editor, *Computation for Metaphors, Analogy and Agents*, volume 1562 of *Springer Lecture Notes in Artificial Intelligence*, pages 52–87. Springer-Verlag.

Churchland, P., Ramachandran, V., and Sejnowski, T. (1994). *A Critique of Pure Vision, in C. Koch and J. Davis eds, 'Large-Scale Neuronal Theories of the Brain'*. MIT Press.

Cohen, M. and Massaro, D. (1990). Synthesis of visible speech. *Behaviour Research Methods, Intruments and Computers*, 22(2):pp. 260–263.

Fitzpatrick, P. (2003a). Object lesson: discovering and learning to recognize objects. Proceedings of the Third International Conference on Humanoid Robots, Karlsruhe, Germany.

Fitzpatrick, P. (2003b). Perception and perspective in robotics. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, Boston.

Fitzpatrick, P. and Torres-Jara, E. (2004). The power of the dark side: using cast shadows for visually-guided reaching. In *Accepted to Humanoids 2004*.

Gallup, G., Anderson, J. R., and Shillito, D. J. (2002). The mirror test. In Bekoff, M., Allen, C., and Burghardt, G. M., editors, *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*, pages 325–33. Bradford Books.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Hendriks-Jansen, H. (1996). *Catching Ourselves in the Act*. MIT Press, Cambridge, Massachusetts.

Hernandez-Reif, M. and Bahrick, L. E. (2001). The development of visual-tactual perception of objects: Amodal relations provide the basis for learning arbitrary relations. *Infancy*, 2(1):51–72.

Hu, H. H., Gooch, A. A., Creem-Regehr, S. H., and Thompson, W. B. (2002). Visual cues for perceiving distances from objects to surfaces. *Presence: Teleoperators and Virtual Environments*, 11(6):652–664.

Hu, H. H., Gooch, A. A., Thompson, W. B., and Smits, B. E. (2000). Visual cues for imminent object contact in realistic virtual environments. In *Proceedings of the 11th IEEE Visualization Conference*, Salt Lake City, Utah.

Hubona, G. S., Shirah, G., and Jennings, D. (2004). The effects of cast shadows and stereopsis on performing computer-generated spatial tasks. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 34(4):forthcoming.

Hubona, G. S., Wheeler, P. N., Shirah, G. W., and Brandt, M. (1999). The relative contributions of stereo, lighting, and background scenes in promoting 3D depth visualization. *ACM Transactions on Computer-Human Interaction*, 6(3):214–242.

Kaernbach, C. (1993). Temporal and spectral basis of the features perceived in repeated noise. *Journal of the Acoustical Soc. of Am.*, 94(1):91–97.

Knudsen, E. I. and Knudsen, P. F. (1985). Vision guides the adjustment of auditory localization in young barn owls. *Science*, 230:545–548.

Kozima, H. and Yano, H. (2001). A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*.

Kozima, H. and Zlatev, J. (2000). An epigenetic approach to human-robot communication. In *IEEE International Workshop on Robot and Human Communication (ROMAN00)*, Osaka, Japan.

Krotkov, E., Klatzky, R., and Zumel, N. (1996). Robotic perception of material: Experiments with shape-invariant acoustic measures of material type. In *Experimental Robotics IV*. Springer-Verlag.

Lee, D. N. (1980). The optic flow field: the foundation of vision. *Philosophical Transactions of the Royal Society of London B*, 290(1038):169–179.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psych. Bull.*, 126:281–308.

Lewkowicz, D. J. (2003). Learning and discrimination of audiovisual events in human infants: The hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. *Developmental Psychology*, 39(5):795–804.

Lewkowicz, D. J. and Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory- visual intensity matching. *Developmental Psychology*, 16:597–607.

Metta, G. and Fitzpatrick, P. (2003). Early integration of vision and manipulation. *Adaptive Behavior*, 11(2):109–128.

Pavani, F. and Castiello, U. (2004). Binding personal and extrapersonal space through body shadows. *Nature Neuroscience*, 7(1):14–15.

Rochat, P. and Striano, T. (2002). Who's in the mirror? self-other discrimination in specular images by four- and nine-month-old infants. *Child Development*, 73(1):35–46.

Scassellati, B. (2001). *Foundations for a Theory of Mind for a Humanoid Robot*. PhD thesis, MIT Department of Electrical Engineering and Computer Science.

Stoytchev, A. (2003). Computational model for an extendable robot body schema. Technical report, Georgia Institute of Technology, College of Computing. GIT-CC-03-44.

von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5:520–526.

Wertheimer, M. (1961). Psychomotor coordination of auditory and visual space at birth. *Science*, 134:1692.