International Journal of Humanoid Robotics

Exploiting Amodal Cues for Robot Perception

Artur M. Arsenio and Paul M. Fitzpatrick

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA
{*arsenio,paulfitz*}*@csail.mit.edu*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

This paper presents an approach to detecting, segmenting, and recognizing rhythmically moving objects that generate sound as they move. We show selectivity and robustness in the face of distracting motion and sounds. Our method does not require accurate sound localization, and in fact is complementary to it. The work is implemented on the humanoid robot Cog [1]. We are motivated by the fact that objects that move rhythmically are common and important for a humanoid robot. The humanoid form is often argued for so that the robot can interact well with tools designed for humans, and such tools are typically used in a repetitive manner, with sound generated by physical abrasion or collision; consider hammers, chisels, saws etc. We also work with the perception of toys designed for infants – rattles, bells etc. – which could have utility for entertainment/pet robotics. Our goal is to build the perceptual tools required for a robot to learn to use tools and toys through demonstration. We show that our approach also applies to robot perception of itself and humans, and relate our work to findings in infant development research.

Keywords: Cross-modal perception; humanoid robotics; object segmentation; object recognition; machine learning

1. Introduction

Tools are often used in a manner that is composed of some repeated motion – consider hammers, saws, brushes, files, etc. This repetition could potentially aid a robot to robustly perceive these objects and their actions. But how? We believe that a key resource in the robust perception of objects and events is the perception of *amodal* properties – that is, properties such as synchronicity and rhythm that manifest themselves across several different senses but are specific to none of them. Amodal properties are by their nature less sensitive to variation of context such as lighting or background noise which affect the individual senses of the robot. Studies of infant development suggest that the presence or absence of amodal properties has a profound impact on attention, learning, and development [2]. There is evidence that they are particularly important for unfamiliar, novel situations, which are

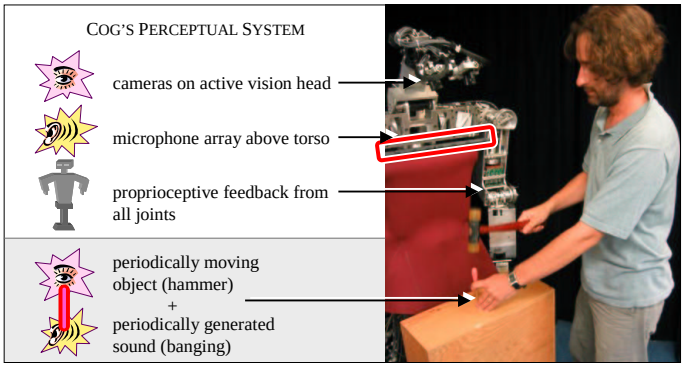


Fig. 1. The experimental platform. The humanoid robot Cog[1] is equipped with cameras in an active vision head, and a microphone array across the torso. A human demonstrates some repetitive action to the robot, such as using a hammer, while the robot watches and listens.

exactly the scenarios of deepest concern to us; it is relatively easy to build an object recognition system for a finite set of known objects, but unconstrained or changing environments are currently much harder to deal with. In previous work, synchronous movement of an object in response to prodding was used as a grouping cue for unfamiliar objects, which could then train a classical object recognition system[3]. In this work, we choose rhythmic motion as a grouping cue that works both within and across the robot's senses. The value of this cue is that it gives a great deal of redundancy, both from its multi-modal quality and its repetitive nature.

We focus on detecting amodal cues in the visual and auditory senses. The advantage of combining information across these two modalities is that they have complementary properties. Since sound waves disperse more readily than light, vision retains more spatial structure – but for the same reason it is sensitive to occlusion and the relative angle of the robot's sensors, while auditory perception is quite robust to these factors. The spatial trajectory of a moving object can be recovered quite straightforwardly from visual analysis, but not from sound. However, the trajectory in itself is not very revealing about the nature of the object. We use the trajectory to extract visual and acoustic features – patches of pixels, and sound frequency bands – that are likely to be associated with the object. Both can be used for recognition. Sound features are easier to use since they are relatively insensitive to spatial parameters such as the relative position and pose of the object and the robot.

In this paper, the humanoid robot Cog is presented with tools or toys in use (see Figure 1). The paper works through a variety of cases for processing and associating information across multiple sensory modalities. Our approach, as described in Section 2, is motivated by development of cross-modal perception in infants. It relies on having the robot detect simple repeated events from multiple sensors at frequen-

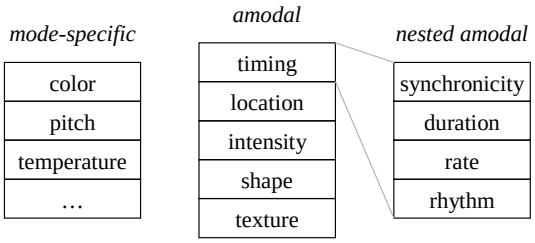


Fig. 2. Features such as color and pitch are specific to a particular sense (sight and hearing respectively). But not all features are so specific. Several are *amodal* and can manifest themselves across multiple senses. For example, smooth and rough objects can generally be distinguished both by sight and touch. Timing is a particularly productive feature, giving rise to a set of *nested* amodal features.

cies relevant for human interaction (Section 3). We demonstrate in Section 4 that repetitive amodal information (such as signal synchrony and timing) is useful to filter out undesirable percepts as well as to associate diverse events across multiple sensor modalities. Section 5 presents both acoustic and visual unimodal segmentation and recognition algorithms. Training data for building an acoustic classifier is automatically generated by the visual identification apparatus. A dynamic programming approach is then used in Section 6 to extract cross-modal features by matching patches of auditory and visual data. Such features are applied for building a cross-modal recognizer.

Amodal information, besides being useful to bind multi-modal object percepts, can also be applied to bind sounds and linguistic events to people, which is the topic of Section 7. By extending cross-modal learning to account for proprioceptive information, and integrating such data with acoustic and visual percepts, the robot identifies not only the acoustic rhythms generated by its body parts, but also its own visual appearance. This way, the robot is able to learn multiple complementary properties about objects, people and itself.

2. 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. Historically, the development of perception in infants has been described using two diametrically opposed classes of theory: integration and differentiation [4]. 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 [5] – but on the other hand, there is a clear progression over

time in the kinds of relations which can be perceived (Lewkowicz[6] 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[7] (see Figure 2).

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[6]. 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[2] 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[8] 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[2]. Another experiment[9] 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

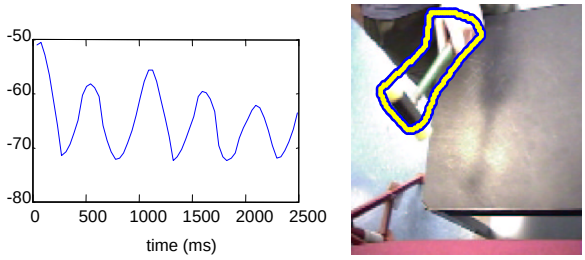


Fig. 3. When watching a person using a hammer, the robot detects and group points moving in the image with similar periodicity[11] to find the overall trajectory of the hammer and separate it out from the background. The detected trajectory is shown on the left (for clarity, just the coordinate in the direction of maximum variation is plotted), and the detected object boundary is overlaid on the image on the right.

for development to occur[10]. 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.

3. Detecting repeated 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 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[12], 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[11]. For example, Figure 3 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

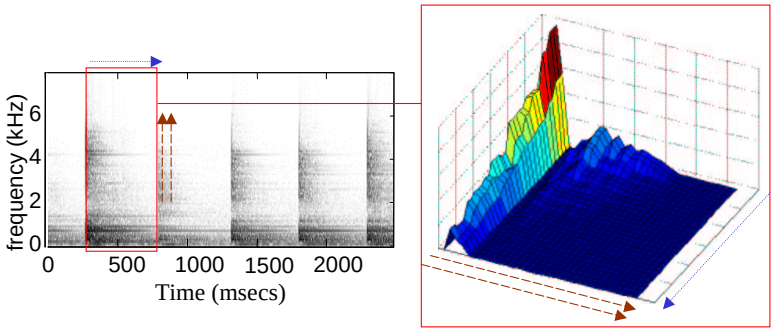


Fig. 4. Extraction of an acoustic pattern from a periodic sound (a hammer banging). The algorithm for signal segmentation is applied to each normalized frequency band. The box on the right shows one complete segmented period of the signal. Time and frequency axes are labeled with single and double arrows respectively.

or fractional *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 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 [13].

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 after-

wards. 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.

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. This method not only relaxes the assumption of constant periodicity, but is also computationally inexpensive. 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 3 shows an example of using periodicity to visually segment a hammer as a human demonstrates the periodic task of hammering, while Figure 4 shows segmentation of the sound of the hammer in the time-domain. 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 object 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. The Fourier coefficients are grouped into a set of frequency bands for the purpose of further analysis, along with the overall energy.

4. Priming for attention

Human studies have shown that attention in one of the senses can be modified by input from the other senses. For example, Bahrick[2] describes an experiment in which two movies of actions such as clapping hands are overlaid, and the sound corresponding to just one of the movies is played. Adult and infant attention is found to be directed to the matching action. In adults, there is a large reported difference between what is perceived when the sound is off (ghostly figures moving through

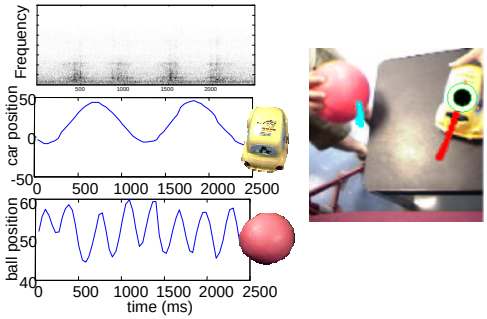


Fig. 5. The image on the right shows a car and a ball moving simultaneously, with their trajectories overlaid. The spectrogram during this event is shown on the left. Sound is only generated by the rolling car – the ball is silent. A circle is placed on the object (car) with which the sound is bound. The sound energy and the visual displacements of the objects are given.

each other) and when the sound is on (a strong sense of figure and background).

4.1. *Priming visual foreground with sound*

In this section, we consider the case of multiple objects moving in the robot's visual field, only one of which is generating sound. The robot uses the sound it hears to filter out uncorrelated moving objects and determine a candidate for cross-modal binding. This is a form of context priming, in which an external signal (the sound) directs attention towards one of a set of potential candidates.

Figure 5 shows measurements taken during an experiment with two objects moving visually, at different rates, with one - a toy car - generating a rolling sound, while the other - a ball - is moving silently. The acoustic signal is linked with the object that generated it (the car) using period matching. The movement of the ball is unrelated to the period of the sound, and so that object is rejected. In contrast, for the car there is a very definite relationship. The sound energy signal has two clear peaks per period of motion, since the sound of rolling is loudest during the two moments of high velocity motion between turning points in the car's trajectory. This is a common property of sounds generated by mechanical rubbing, so the binding algorithm takes this possibility into account by testing for the occurrence of frequencies at double the expected value.

4.2. *Priming acoustic foreground with vision*

We now consider the case of one object moving in the robot's field of view, and one 'off-stage', with both generating sound. This is symmetric to the case already covered. Matching the correct sound to the visible object is achieved by mapping the time history of each individual coefficient band of the audio spectrogram (see Figure 6) to the visual trajectory of the object. We segment the sound of the object

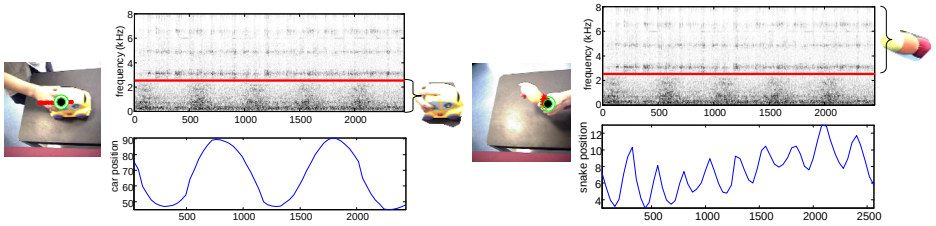


Fig. 6. The two spectrogram/trajectory pairs shown are for a shaking toy car and snake rattle. The left pair occurs with only the car visible, and the right pair occurs with only the snake visible. The line in each spectrogram represents the cutoff pitch frequency between the car and snake.

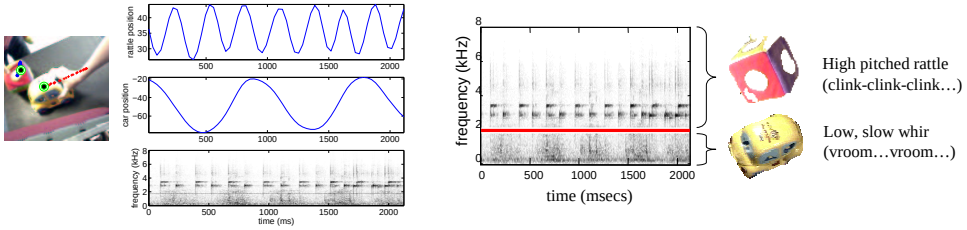


Fig. 7. The car and the cube, both moving, both making noise. The line overlaid on the spectrogram (right) shows the cutoff determined automatically between the high-pitched bell in the cube and the low-patched rolling sound of the car. A spectrogram of the car alone can be seen in Figure 5.

from the background by clustering the frequency bands with the same period (or half the period) as the visual target, and assign those bands to the object.

Within the framework being described, visual information is used to prune the range of frequency bands of the original sound. The coefficient bands of the audio visual are segmented into clusters of bands that characterize the sound of an object. For the experiment shown to the left in Figure 6, the coefficients ranging from 0 to 2.6Hz are assigned to the object. Afterwards, a band-pass filter is applied to the audio-signal to filter out the other frequencies, resulting in the clear sound of the car with the sound of the rattle removed or highly attenuated. For the experiment shown in the right part of Figure 6 the roles of the car and snake were switched. A band-pass filter between 2.6-2.8Hz is applied to the audio-signal to filter out the frequencies corresponding to the car, resulting in the snake's sound.

4.3. *Matching multiple sources*

This experiment considers two objects moving in the robot's field of view, both generating sound, as presented in Figure 7. Each frequency band is mapped to one of the visual trajectories if coherent with its periodicity. For each object, the lower and the higher coefficient band are labeled as the lower and higher cut-off

Experiment	visual period found	sound period found	bind sound, vision	candidate binds	correct binds	incorrect binds
hammer	8	8	8	8	8	0
car and ball	14	6	6	15	5	1
plane & mouse/remote	18	3	3	20	3	0
car (snake in backg'd)	5	1	1	20	1	0
snake (car in backg'd)	8	6	6	8	6	0
car & cube *car*	9	3	3	11	3	0
car & cube *cube*	10	8	8	11	8	0
car & snake *car*	8	0	0	8	0	0
car & snake *snake*	8	5	5	8	5	0

Table 1. Evaluation for four binding cases of cross-modal rhythms of increasing complexity. The simplest is when a single object (the hammer) is in view, engaged in a repetitive motion and a single repetitive sound source is also heard. This corresponds to a run of roughly 1 minute, for which binding is easy as shown by the data. The next case is when multiple moving objects are visible, but only one repeating sound is heard. Two experiments were made – a car and a ball visible and only the car generating sound, and a plane and other objects visible but only the plane generating sound. Since an object's sound is strongly affected by environment noise, highest confidence is required for this modality, which reduces the number of periodic detections, and consequently the number of bindings. The third case corresponds to two repeating sounds with different periods, and a single visible moving object (experiments for car with snake rattle in background and vice-versa). The car generates mainly low frequency sounds, but the rattle generates high frequency sounds with some weak low frequency components that cause interference with the detection of the car's sound. This is the reason for a weak percentage of bindings for the car. Finally, multiple sound and visual source can be bound together appropriately (two experiments: car and cube rattle; and car and snake rattle). Bindings occur more often for objects producing sounds with high frequency energies.

frequencies, respectively, of a band-pass filter assigned to that object. The complex sound of both the moving car-toy and the cube-rattle are thus segmented into the characteristic sound of the car and sound of the rattle through band-pass filtering. Multiple bindings are thus created for multiple oscillating objects producing distinct sounds.

It is worth stressing that the real world is full of objects making all kinds of noise. However, the system is robust to such disturbances. On the experiments presented throughout this paper, people were speaking occasionally while interacting with the robot, while other people were making everyday sounds while working. If the distracting sound occurs at the same range of frequencies as the sound of the oscillating object, then a binding might just not occur for that specific time, but occur after a few seconds when the interference noise switches to other frequencies or disappears. Table 1 shows how well the methods described for binding sounds with objects work on a series of experiments.

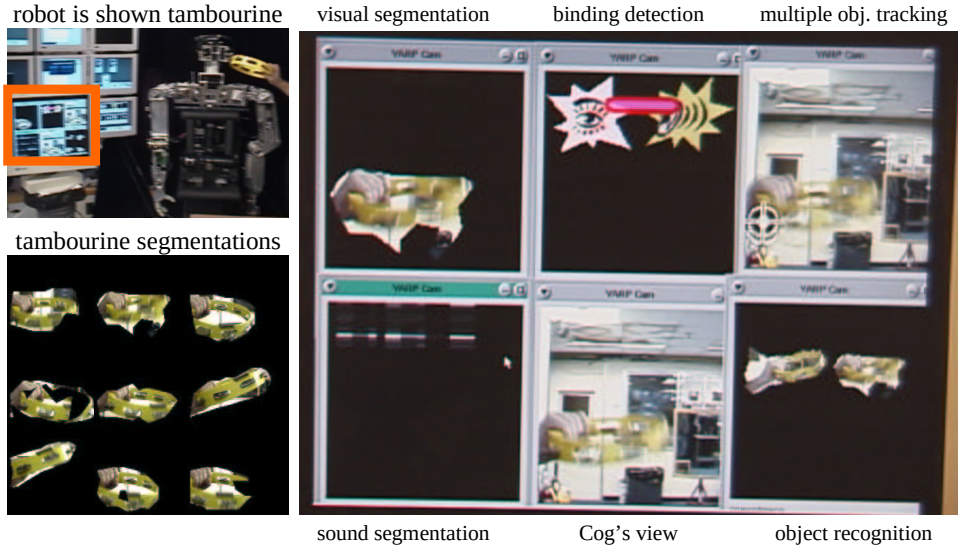


Fig. 8. Here the robot is shown a tambourine in use (top left). The robot detects that there is a periodically moving visual source, and a periodic sound source, and that the two sources are causally related and should be bound. All images in these figures are taken directly from recordings of real-time interactions. The images on the bottom left show the visual segmentations recorded for the tambourine. The background behind the tambourine, a light wall with doors and windows, is correctly removed. The panel on the right shows a real-time view of the robot's status during the experiment. 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.

5. Differentiation

Our system can extract both the acoustic signature and the visual appearance of objects independently, by detecting periodic oscillations within each sensor modality. Segmented features extracted from visual and acoustic segmentations can then serve as the basis for an object recognition system. 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 (for an approach in the visual domain see Arsenio[14] or Fitzpatrick[15], and Krotkov[16] has looked at the recognition of sound generated by a single contact event). In our system, the robot's perceptual system comprises several unimodal algorithms running in parallel to extract informative percepts within and across the senses (see the display panel in Figure 8).

5.1. *Visual segmentation and recognition*

Object segmentation is a fundamental problem in computer vision, and is particularly difficult on the unstructured, non-static, noisy, real-time, low resolution images that robots have to deal with. We approach segmentation by detecting and

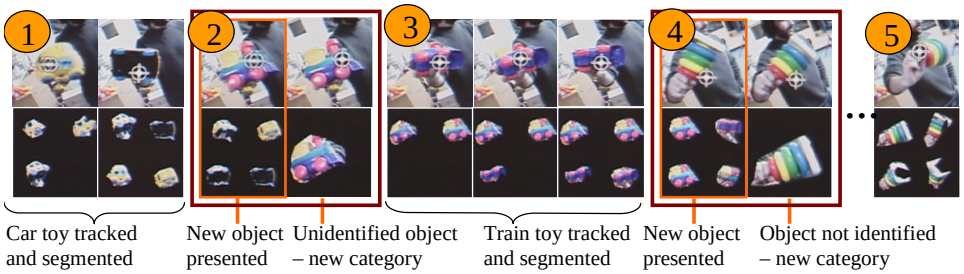


Fig. 9. Figure illustrating a sequence from an on-line experiment of several minutes on the humanoid robot Cog. (1) The robot is tracking a toy car (top row), and new template instances of it are being inserted into a database. A random set of templates from this database is shown on the bottom row. (2) A new object (a toy train) is presented. It was never seen before, so it is not recognized and a new category is created for it. (3) The toy train is tracked. (4) A new, unknown object presented, for which a new category is created on the object recognition database. (5) Templates from the new object are stored.

interpreting natural human behavior such as waving or shaking objects, clustering periodically-moving pixels in an image into a unified object (following the procedure described by Arsenio[17]). The *object templates* produced by segmentation are used as the basis for training an object recognition system, which enables object identification in several contexts and under different perspective views. The object recognition algorithm begins by clustering objects into classes according to their identity. This was implemented using color histograms; objects were classified based on the relative distribution of their color pixels. New object templates are classified according to their similarity with other object templates in an object database. A multi-target tracking algorithm (which tracks good features[18] using the Lucas-Kanade Pyramidal algorithm) was developed to keep track of object identity as it changes location and pose. An on-line experiment for object segmentation, tracking and recognition of new objects on the humanoid robot is shown in Figure 9. Arsenio[19] presents both a qualitative and quantitative analysis for recognition of previously learned objects.

5.2. *Auditory segmentation and recognition*

The repetitive nature of the sound generated by an object under periodic motion can be analyzed to extract an acoustic 'signature' for that object. We search for repetition in a set of frequency bands independently, then collect those frequency bands whose energies oscillate together with a similar period. Specifically, the acoustic signature for an object is obtained by applying the following steps:

(1) The period of repetition for each frequency band is detected using the procedure developed in Section 3.
(2) A *period histogram* is constructed to accumulate votes for frequency bands having the same estimated period (or half the period – it is common to have sounds

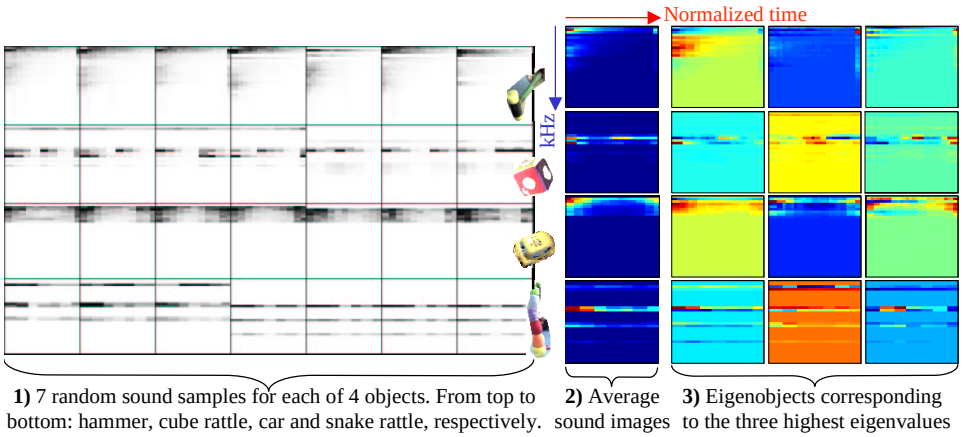


Fig. 10. Sound segmentation and recognition. Acoustic signatures for four objects are shown along the rows. (1) Seven sound segmentation samples are shown for each object, from a total of 28 (car), 49 (cube rattle), 23 (snake rattle) and 34 (hammer) samples. (2) The average acoustic signature for each object is shown. The vertical axis corresponds to the frequency bands and the horizontal axis to time normalized by the period. (3) The eigensounds corresponding to the three highest eigenvalues are shown.

that occur once per repetition, for example at one endpoint of the trajectory, or twice per repetition, for example at two instants of maximum velocity). The histogram is smoothened by adding votes for each bin of the histogram to their immediate neighbors as well.

(3) The maximum entry in the period histogram is selected as the *reference* period. All frequency bands corresponding to this maximum are collected and their responses over the reference period are stored in a database of acoustic signatures. Since the same objects can be shaken or waved at different velocities resulting in varying periodicity, it is important to normalize temporal information relative to the reference period.

A collection of annotated acoustic signatures for each object are used as input data (see Figure 10) for a sound recognition algorithm by applying the eigenobjects method, which is also widely used for face recognition[20]. This method is a modified version of Principal Component Analysis. A sound image is represented as a linear combination of base sound signatures (or *eigensounds*). Only eigensounds corresponding to the three highest eigenvalues – which represent a large portion of the sound's energy – are retained. Classification consists of projecting novel sounds to this space, determining the coefficients of this projection, computing the L_2 distance to each object's coefficients in the database, and selecting the class corresponding to the minimum distance.

Cross-modal information aids the acquisition and learning of unimodal percepts and consequent categorization in a child's early infancy. Similarly, visual data is

employed here to guide the annotation of auditory data to implement a sound recognition algorithm. Training samples for the sound recognition algorithm are classified into different categories by the visual object recognition system or from information from the visual object tracking system. This enables the system, after training, to classify sounds of unknown, not visible objects.

The system was evaluated quantitatively be randomly selecting 10% of the segmented data for validation, and the remaining data for training. This process was randomly repeated three times. It is worth noting that even samples received within a short time of each other often do not look very similar, due to noise on the segmentation process, background acoustic noise, other objects' sounds during experiments, and variability on how objects are moved and presented to the robot. For example, the car object is heard both alone and with a rattle (either visible or hidden).

The recognition rate for the three runs averaged to 82% (86.7%, 80% and 80%). Recognition rates by object category were: 67% for the car, 91.7% for the cube rattle, 77.8% for the snake rattle and 83.3% for the hammer. Most errors arise from mismatches between car and hammer sounds. Such errors could be avoided by extending our sound recognition method to use derived features such as the onset/decay rate of a sound, which is clearly distinct for the car and the hammer (the latter generates sounds with abrupt rises of energy and exponential decays, while sound energy from the toy car is much smoother). Instead, we will show that these differences can be captured by cross-modal features to correctly classify these objects.

6. Integration

Different objects have distinct acoustic-visual patterns which are a rich source of information for object recognition, if we can recover them. The relationship between object motion and the sound generated varies in an object-specific way. 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. A bell typically causes sound at either extreme of motion. All these statements are truly cross-modal in nature, and we explore here using such properties for recognition.

6.1. *Cross-Modal segmentation/recognition*

As was just described, features extracted from the visual and acoustic segmentations are what is needed to build an object recognition system. Each type of features are important for recognition when the other is absent. 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 soft relative to the visual 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. Therefore, the feature space for recognition consists of:

- ▷ Sound/Visual period ratios – the sound energy of a hammer peaks once per visual period, while the sound energy of a car peaks twice (for forward and backward movement).
- ▷ Visual/Sound peak energy ratios – the hammer upon impact creates high peaks of sound energy relative to the amplitude of the visual trajectory. Although such measure depends on the distance of the object to the robot, the energy of both acoustic and visual trajectory signals will generally decrease with depth (the sound energy disperses through the air and the visual trajectory reduces in apparent scale).

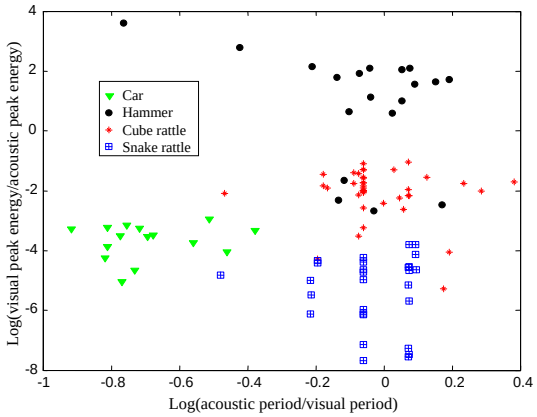
Human actions are therefore used to create associations along different sensor modalities, and objects can be recognized from the characteristics of such associations. Our approach can differentiate objects from both their visual and acoustic backgrounds by finding pixels and frequency bands (respectively) that are oscillating together. This is accomplished through dynamic programming, applied to match the sound energy to the visual trajectory signal. Formally, let $S = (S_1, \ldots, S_n)$ and $V = (V_1, \ldots, V_m)$ be sequences of sound and visual trajectory energies segmented from n and m periods of the sound and visual trajectory signals, respectively. Due to noise, n may be different to m. If the estimated sound period is half the visual one, then V corresponds to energies segmented with $2m$ half periods (given by the distance between maximum and minimum peaks). A matching path $P = (P_1, \ldots, P_l)$ defines an alignment between S and M, where $max(m, n) \leq l \leq m + n - 1$, and $P_k = (i, j)$, a match k between sound cluster j and visual cluster i. The matching constraints are imposed by:

The boundary conditions are $P_1 = (1, 1)$ and $P_l = (m, n)$.

Temporal continuity satisfies $P_{k+1} \in \{(i + 1, j + 1), (i + 1, j), (i, j + 1)\}$. This restricts steps to adjacent elements of P.

The function cost $c_{i,j}$ is given by the square difference between V_i and S_j periods. The best matching path W can be found efficiently using dynamic programming, by incrementally building an $m \times n$ table caching the optimum cost at each table cell, together with the link corresponding to that optimum. The binding W will then result by tracing back through these links, as in the Viterbi algorithm.

Figure 11 shows cross-modal features for a set of four objects. It would be hard to cluster automatically such data into groups for classification. But as in the sound recognition algorithm, training data is automatically annotated by visual recognition and tracking. After training, objects can be categorized from cross-modal cues alone. The system was evaluated quantitatively by selecting randomly 10% of the data for validation, and the remaining data for training. This process was randomly repeated fifteen times. The recognition rate averaged over all these runs were, by object category: 100% for both the car and the snake rattle, 86.7% for



Confusion matrix	car	cube	snake	hammer
car	**30**	0	0	0
cube	0	**52**	7	1
snake	0	0	**45**	0
hammer	0	5	0	**25**

Fig. 11. Object recognition from cross-modal clues. The feature space consists of period and peak energy ratios. The confusion matrix for a four-class recognition experiment is shown. The period ratio is enough to separate well the cluster of the car object from all the others. Similarly, the snake rattle is very distinct, since it requires large visual trajectories for producing soft sounds. Errors for categorizing a hammer originated exclusively from erroneous matches with the cube rattle, because hammering is characterized by high energy ratios, and very soft bangs are hard to identify correctly. The cube rattle generates higher energy ratios than the snake rattle. False cube rattle recognitions resulted mostly from samples with low energy ratios being mistaken for the snake rattle.

the cube rattle, and 83% for the hammer. The overall recognition rate was 82.1%. Such results demonstrate the potential for recognition using cross-modal cues.

6.2. *Cross-modal enhancement of detection*

There is evidence that, for humans, simple visual periodicity can aid the detection of acoustic periodicity. If a repeating segment of noise is played, the repetition can be detected for much longer periods if a light is flashing in synchrony with some point in the period[21]. More generally, there is evidence that the cues used to detect periodicity can be quite subtle and adaptive[22], suggesting there is a lot of potential for progress in replicating this ability beyond the ideas already described. We believe that cross-modal priming can be used to refine detection, both for detecting signals that would otherwise be missed, and ignoring signals that would otherwise distract.

Much of the noise in the results of the previous section were symptomatic of a general problem: the sound generated by a periodically moving object can be much more complex and ambiguous than its visual trajectory. The extrema of an

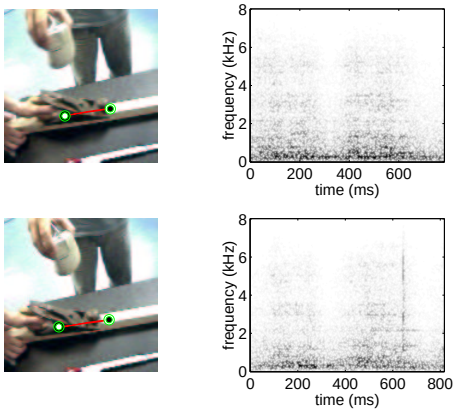


Fig. 12. An experiment in which a plane is being pushed across wood while a mouse is shaken in the background. Shown are the highest quality acoustic matches for this sound (right) and the object with which they correspond (left). Matches against the mouse are much lower and below threshold.

approximately repeating trajectory can be found with ease, and used to segment out single periods of oscillation within an object's movement. Single periods of the sound signal can be harder to find, since there is more ambiguity – for example, some objects make noise only at one point in a trajectory (such as a hammer), others make noise at the two extrema (some kinds of bell), others make noise during two times of high velocity between the extrema (such as a saw), and so on. For cases where periodicity detection is difficult using sound, it makes sense to define the period of an action in the visual domain based on its trajectory, and match against this period in the sound domain – instead of detecting the period independently in each domain. We have developed an approach, where for each object moving visually, fragments of the sound are taken for periods of that object, aligned, and compared. If the fragments are consistent, with sound and vision in phase with each other, then the visual trajectory and the sound are bound. This is a more stringent test than just matching periods, yet avoids the problem of determining a period reliably from sound information. Figure 12 shows results for an experiment where two objects are moving, a mouse and a plane. Only the plane is generating sound. The sound is a rough noise with silence at the two extrema of the plane's motion, and hence appears to have a frequency of double that of the trajectory. By coincidence, this is close to the frequency of oscillation of the mouse, so simple period matching is difficult. But by using the simple visual period to segment the acoustic data, small differences can be amplified as the sound and vision of a near match drift around in phase while a true match stays exactly in phase.

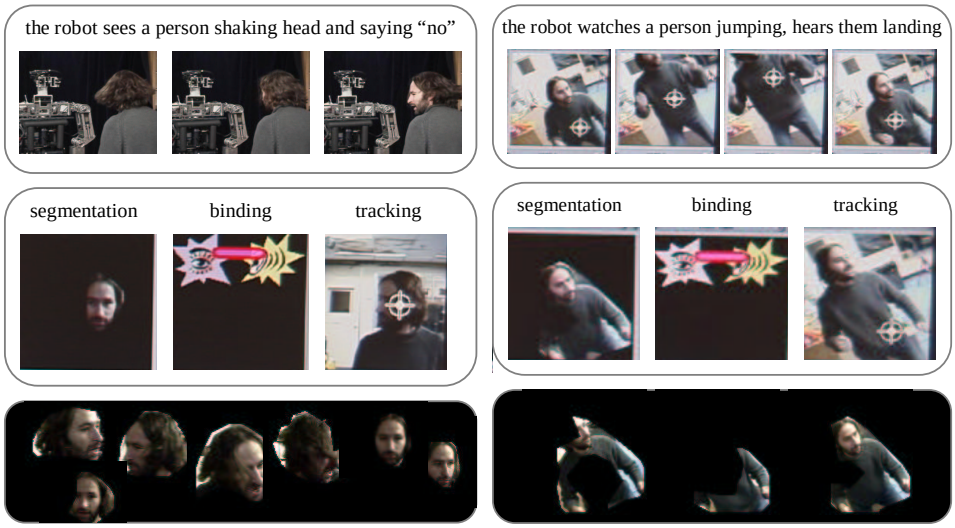


Fig. 13. (left) In this experiment, the robot sees people shaking their head. In the top row, the person says "no, no, no" in time with his head-shake. The middle row shows the recorded state of the robot during this event – it binds the visually tracked face with the sound spoken. Recorded segmentations for these experiments are shown on the lower row. (right) Result for one human actor jumping up and down like crazy in front of the robot. The thud as he hit the floor was correctly bound with segmentations of his body (bottom row).

7. Beyond objects: detecting the self and others

The cross-modal binding method we developed for object perception also applies to perceiving people. Humans often use body motion and repetition to reinforce their actions and speech, especially with young infants. If we do the same in our interactions with Cog, then it can use those cues to link visual input with corresponding sounds. For example, Figure 13 shows a person shaking their head while saying "no! no! no!" in time to his head motion. The figure shows that the robot extracts a good segmentation of the shaking head, and links it with the sound signal. Such actions appear to be understood by human infants at around 10-12 months.

Sometimes a person's motion causes sound, just as an ordinary object's motion might. Figure 13 shows a person jumping up and down in front of Cog. Every time he lands on the floor, there is a loud bang, whose periodicity matches that of the tracked visual motion. We expect that there are many situations like this that the robot can extract information from, despite the fact that those situations were not considered during the design of the binding algorithms. The images in these figures are taken from online experiments – no offline processing is done.

So far we have considered only external events that do not involve the robot. Now 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

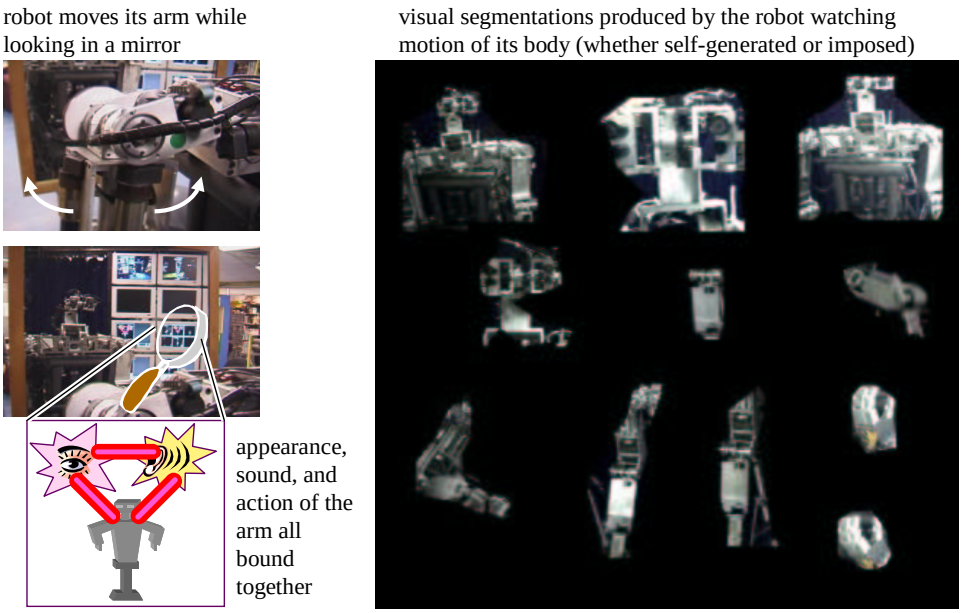


Fig. 14. In the experiment shown to the left, Cog is looking at itself in a mirror, while shaking its arm back and forth. The reflected image of its arm is bound to the robot's sense of its own motion, and the sound of the motion. This binding is identical in kind to the binding that occurs if the robot 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.

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).

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[23]. Self-recognition in a mirror is also the focus of extensive study in biology. Work on self-recognition in mirrors for chimpanzees[24] 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[25,26]. 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[27]. Such behavior requires the integration of visual cues from the mirror with proprioceptive cues from the child's body. As shown in Figure 14, 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[28]). 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[29,25]. 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.

8. Conclusions and Discussion

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).

Most of us have had the experience of feeling a tool become an extension of ourselves as we use it (see[30] 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 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.

Other work in robotics has taken advantage of cross-modal cues for word learning [31]. 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[21] – 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[22], suggesting there is a lot of potential for progress in replicating this ability beyond the ideas already described.

Although the potential for expanding this work is vast, from a practical per-

spective complex levels of functionality have already been accomplished. Consider Figure 8, 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, the relationship between data from different senses detected and stored, and objects tracked to be further used by statistical learning processes for object location. 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.

A lot about the world can be communicated to a humanoid robot through human demonstration. The robot's learning process will be facilitated by sending it repetitive information through this communication channel. If more than one communication channel is available, such as the visual and auditory channels, both sources of information can be correlated for extracting richer pieces of information. We demonstrated in this paper a specific way to take advantage of correlating multiple perceptual channels at an early stage, rather than just by analyzing them separately - the whole is truly greater than the sum of the parts.

Acknowledgements

This work was funded by DARPA DABT 63-00-C-10102 ("Natural Tasking of Robots Based on Human Interaction Cues"), and by NTT under the NTT/MIT Collaboration Agreement. Arsenio was supported by Portuguese grant PRAXIS XXI BD/15851/98.

References

1. R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, The Cog project: Building a humanoid robot, *Lect. Notes in Comp. Sci.* **1562** (1999) 52–87.
2. L. E. Bahrick, The development of perception in a multimodal environment, in G. Bremner, A. Slater (eds.), Theories of infant development (Blackwell Publishing, Malden, MA, 2004) 90–120.
3. P. Fitzpatrick, Object lesson: discovering and learning to recognize objects, in Proceedings of the Third International Conference on Humanoid Robots, Karlsruhe, Germany (2003) .
4. L. E. Bahrick, Development of intermodal perception, in L. Nadel (ed.), Encyclopedia of Cognitive Science (Nature Publishing Group, London, 2003), vol. 2 614–617.
5. D. J. Lewkowicz, G. Turkewitz, Cross-modal equivalence in early infancy: Auditory- visual intensity matching, *Developmental Psychology* **16** (1980) 597–607.
6. D. J. Lewkowicz, The development of intersensory temporal perception: an epigenetic systems/limitations view, *Psych. Bull.* **126** (2000) 281–308.

7. D. J. Lewkowicz, Learning and discrimination of audiovisual events in human infants: The hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues, *Developmental Psychology* **39** (2003) (5) 795–804.
8. L. E. Bahrick, R. Lickliter, Intersensory redundancy guides attentional selectivity and perceptual learning in infancy, *Developmental Psychology* **36** (2000) 190–201.
9. M. Hernandez-Reif, L. E. Bahrick, The development of visual-tactual perception of objects: Amodal relations provide the basis for learning arbitrary relations, *Infancy* **2** (2001) (1) 51–72.
10. P. Fitzpatrick, Perception and perspective in robotics, in Proceedings of the 25th Annual Conference of the Cognitive Science Society (Boston, 2003) .
11. A. Arsenio, P. Fitzpatrick, C. C. Kemp, G. Metta, The whole world in your hand: Active and interactive segmentation, in Third International Workshop on Epigenetic Robotics (2003) .
12. H. Hendriks-Jansen, Catching Ourselves in the Act (MIT Press, Cambridge, Massachusetts, 1996).
13. A. Arsenio, P. Fitzpatrick, Exploiting cross-modal rhythm for robot perception of objects, in Proceedings of the Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems (Singapore, 2003) .
14. A. Arsenio, Embodied vision - perceiving objects from actions, *IEEE International Workshop on Human-Robot Interactive Communication* (2003).
15. P. Fitzpatrick, From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot, Ph.D. thesis, MIT, Cambridge, MA, 2003.
16. E. Krotkov, R. Klatzky, N. Zumel, Robotic perception of material: Experiments with shape-invariant acoustic measures of material type (O. Khatib and K. Salisbury, editors, Experimental Robotics IV. Springer-Verlag, 1996).
17. A. M. Arsenio, An embodied approach to perceptual grouping (2004) Accepted to the IEEE CVPR Workshop on Perceptual Organization in Computer Vision.
18. J. Shi, C. Tomasi, Good features to track, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1994) 593 – 600.
19. A. M. Arsenio, Map building from human-computer interactions (2004) Accepted to the IEEE CVPR Workshop on Real-time Vision for Human Computer Interaction.
20. M. Turk, A. Pentland, Face recognition using eigenfaces, in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (1991) .
21. J. A. Bashford, B. S. Brubaker, R. M. Warren, Cross-modal enhancement of repetition detection for very long period recycling frozen noise, *Journal of the Acoustical Soc. of Am.* **93** (1993) (4) 2315.
22. C. Kaernbach, Temporal and spectral basis of the features perceived in repeated noise, *Journal of the Acoustical Soc. of Am.* **94** (1993) (1) 91–97.
23. P. Rochat, T. Striano, Who's in the mirror? self-other discrimination in specular

images by four- and nine-month-old infants, *Child Development* **73** (2002) (1) 35–46.
24. G. Gallup, J. R. Anderson, D. J. Shillito, The mirror test, in M. Bekoff, C. Allen, G. M. Burghardt (eds.), The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition (Bradford Books, 2002) 325–33.
25. H. Kozima, H. Yano, A robot that learns to communicate with human caregivers, in Proceedings of the First International Workshop on Epigenetic Robotics (2001) .
26. B. Scassellati, Foundations for a Theory of Mind for a Humanoid Robot, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, 2001.
27. American Academy Of Pediatrics, Caring for Your Baby and Young Child: Birth to Age 5 (Bantham, 1998).
28. S. Baron-Cohen, Mindblindness: An Essay on Autism and Theory of Mind (MIT Press, 1995).
29. H. Kozima, J. Zlatev, An epigenetic approach to human-robot communication, in IEEE International Workshop on Robot and Human Communication (ROMAN00) (Osaka, Japan, 2000) .
30. A. Stoytchev, Computational model for an extendable robot body schema, Tech. Rep. GIT-CC-03-44, Georgia Institute of Technology, College of Computing, 2003.
31. D. Roy, A. Pentland, Learning words from sights and sounds: A computational model, *Cognitive Science* **26** (2002) (1) 113–146.

Artur M. Arsenio received his M.S. degree in Electrical and Computer Engineering from the Technical University of Lisbon, in 1998. He is currently a Fulbrighter pursuing a Ph.D. degree (expected completion in May 2004) at the MIT Computer Science and Artificial Intelligence Laboratory. He was an assistant professor at New University of Lisbon from 1997 to 1998. He was the recipient of the Rice-Cullimore Award from the American Society of Mechanical Engineers in 1999.

Paul M. Fitzpatrick is currently a Postdoctoral Lecturer at the MIT Computer Science and Artificial Intelligence Laboratory. He receive his M.Eng. in Computer Engineering from the University of Limerick, Ireland, and a Ph.D. in Computer Science from MIT in June 2003 for work addressing developmental approaches to machine perception for a humanoid robot.