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.

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 (see Figure 2). 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 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

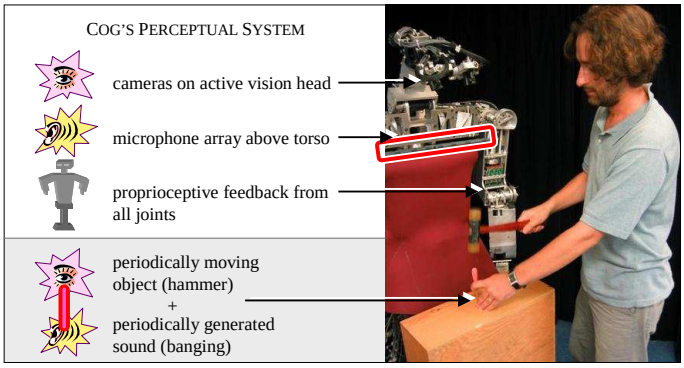


Fig. 1. The experimental platform. The humanoid robot Cog[1] is equipped with cameras in an active vision head, and a microphone array across the torso. A human demonstrates some repetitive action to the robot, such as using a hammer, while the robot watches and listens.

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 is motivated by development of cross-modal perception in infants. It relies on having the robot detect simple repeated events from multiple sensors at frequencies relevant for human interaction (Section 2). We demonstrate in Section 3 that repetitive amodal information (such as signal synchrony and timing) is useful to filter out undesirable

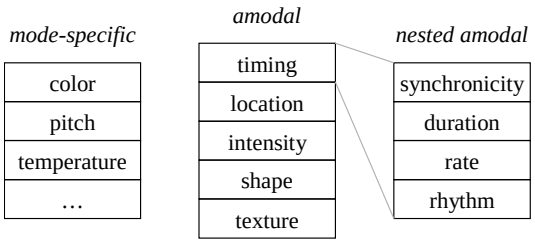


Fig. 2. Features such as color and pitch are specific to a particular sense (sight and hearing respectively). But not all features are so specific. Several are *amodal* and can manifest themselves across multiple senses. For example, smooth and rough objects can generally be distinguished both by sight and touch. Timing is a particularly productive feature, giving rise to a set of *nested* amodal features[4].

percepts as well as to associate diverse events across multiple sensor modalities. Section 4 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 5 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 6. 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. Detecting rhythmic motion

Perhaps the most direct way to detect periodicity is to use the Short-Time Fourier Transform (STFT). This transform maps a signal into a two-dimensional function of time and frequency. A STFT is applied to each input signal available,

$$I(t, f_t) = \sum_{t=0}^{N-1} i(t')h(t'-t)e^{-j2\pi f_t t'} \quad (1)$$

where h is a windowing function, and N the number of samples. Periodicity is estimated from a periodogram determined for all signals from the energy of the windowed FFTs over the spectrum of frequencies. These periodograms are then processed to determine whether they correspond to a clear unambiguous period. A periodogram is accepted if its energy is concentrated over a narrow peak.

This is a very general method, and is similar to that adopted in[5]. There are some difficulties with applying it. We must choose a time span over which to per-

form periodicity inference. The longer this span, the larger the number of signal repetitions available, and the larger the set of frequencies that can be processed, increasing the range of oscillating objects with which the robot can interact. But visual tracking performance decreases with the increase of the time window, and the assumption of constant period over this window is weakened for larger intervals, for both audio and visual signals. We attempted to address this tension by adopting a flexible compromise between the spatial and frequency based views of a signal. The periodicity detection is applied at multiple scales, with long spatial intervals (and hence small window sizes) providing more precise spatial, local information, while larger windows increase the frequency resolution. For objects oscillating during a short period of time, the movement might not appear periodic at a coarser scale, but does appear as such at a finer scale. If a strong periodicity is found, the points implicated are used as seeds for object segmentation. Otherwise the window size is halved and the procedure is repeated for each half.

The STFT based-method demonstrated good performance after extensive evaluation for *visual* signals from periodic objects, but it is not appropriate for periodicity detection of *acoustic* signals. These signals may vary considerably in amplitude between periods, which – particularly when combined with variability in the length of the periods – suggests that Fourier analysis is not appropriate. This led us to the development of a more robust method for periodicity detection, which was applied to both acoustic and visual signals. We construct a histogram of the durations between successive instances of particular values of the signal, and search for the most common duration, which we take to be the signal's period. To make this process efficient, we use a hash table indexed by the signal values (combined with their derivatives) discretized relative to the minimum and maximum value of the signal. Each index of the hash table corresponds to distinct values for the signal and its derivative, and the information stored in the table corresponds to the time instants when each value/derivative pair last occurred. After building the table, a period histogram is constructed by the differences found during assignments to the table. The histogram maximum corresponds to the signal period.

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

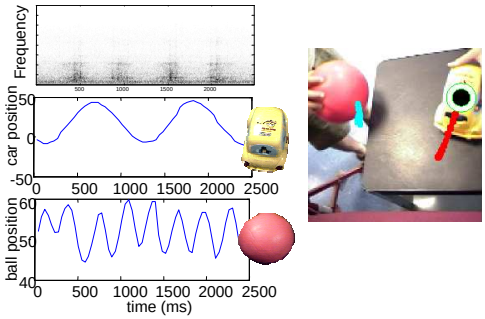


Fig. 3. The image on the right shows a car and a ball moving simultaneously, with their trajectories overlaid. The spectrogram during this event is shown on the left. Sound is only generated by the rolling car – the ball is silent. A circle is placed on the object (car) with which the sound is bound. The sound energy and the visual displacements of the objects are given.

3.1. *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 to 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 3 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.

3.2. *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 4) to the visual trajectory of the object. We segment the sound of the object 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

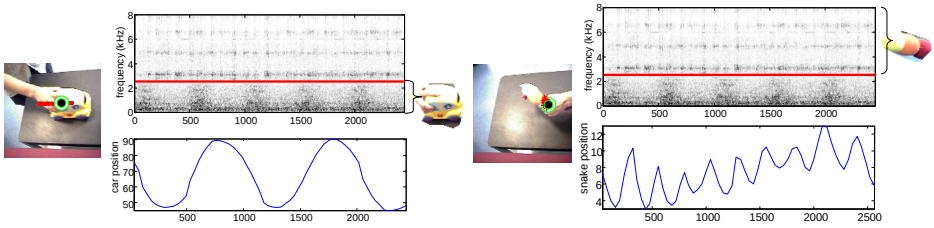


Fig. 4. The two spectrogram/trajectory pairs shown are for a shaking toy car and snake rattle. The left pair occurs with only the car visible, and the right pair occurs with only the snake visible. The line in each spectrogram represents the cutoff pitch frequency between the car and snake.

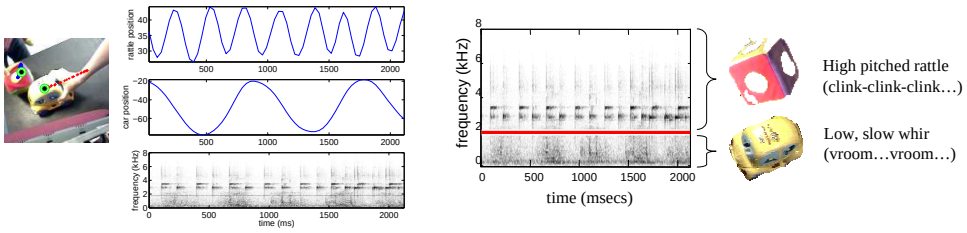


Fig. 5. The car and the cube, both moving, both making noise. The line overlaid on the spectrogram (right) shows the cutoff determined automatically between the high-pitched bell in the cube and the low-patched rolling sound of the car. A spectrogram of the car alone can be seen in Figure 3.

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 4, the coefficients ranging from 0 to 2.6Hz are automatically 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 4 the roles of the car and snake were switched. A band-pass filter (in this case, 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.

3.3. *Matching multiple sources*

This experiment considers two objects moving in the robot's field of view, both generating sound, as presented in Figure 5. 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 bands are labeled as the lower and higher cut-off 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.

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.

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.

4. 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

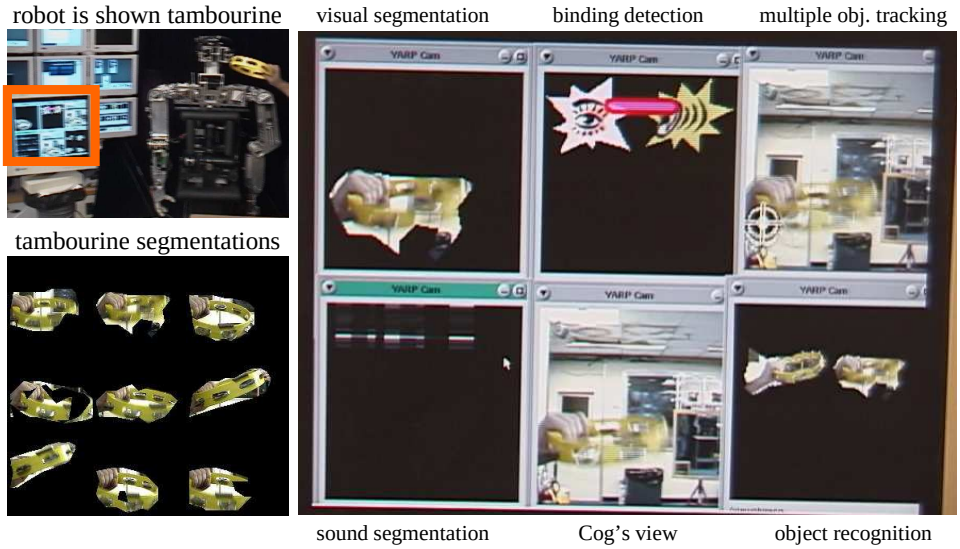


Fig. 6. Here the robot is shown a tambourine in use (top left). The robot detects that there is a periodically moving visual source, and a periodic sound source, and that the two sources are causally related and should be bound. All images in these figures are taken directly from recordings of real-time 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.

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[6] or Fitzpatrick[7], and Krotkov[8] 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 6).

4.1. *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 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[9]). The *object templates* produced by segmentation are used as the basis for training an object recognition system, which enables object

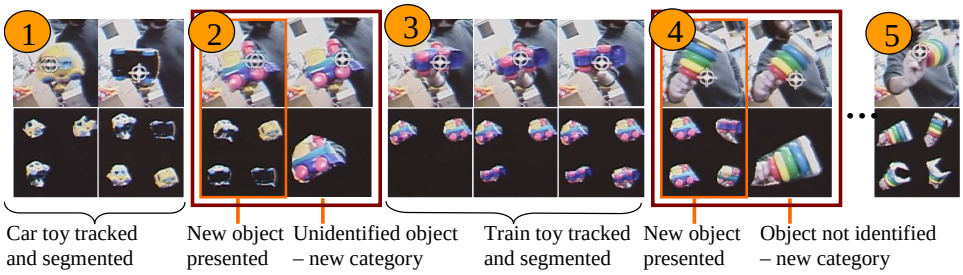


Fig. 7. Figure illustrating a sequence from an on-line experiment of several minutes on the humanoid robot Cog. (1) The robot is tracking a toy car (top row), and new template instances of it are being inserted into a database. A random set of templates from this database is shown on the bottom row. (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.

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[10] 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 7. Arsenio[11] presents both a qualitative and quantitative analysis for recognition of previously learned objects.

4.2. *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 2.
(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 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.

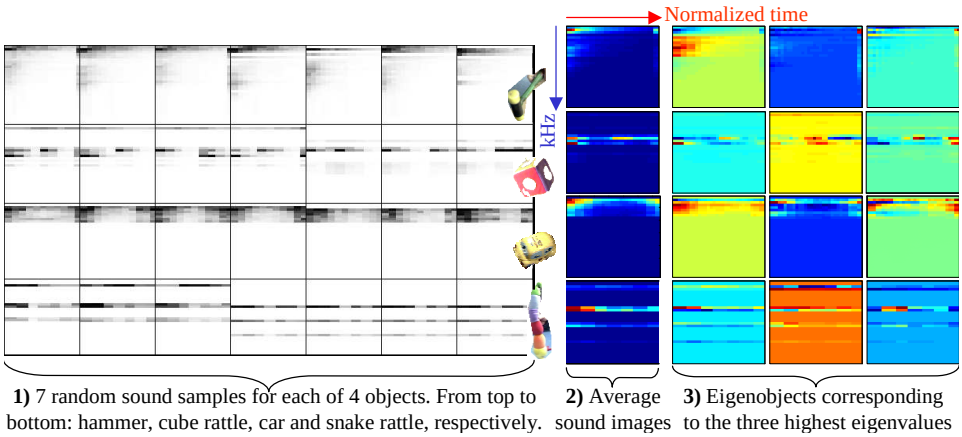


Fig. 8. Sound segmentation and recognition. Acoustic signatures for four objects are shown along the rows. (1) Seven sound segmentation samples are shown for each object, from a total of 28 (car), 49 (cube rattle), 23 (snake rattle) and 34 (hammer) samples. (2) The average acoustic signature for each 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.

(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 8) for a sound recognition algorithm by applying the eigenobjects method, which is also widely used for face recognition [12]. 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 by ranadomly 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.

5. 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.

5.1. *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).

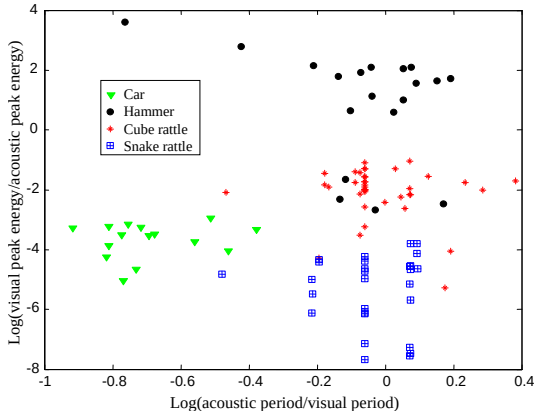
Human actions are therefore used to create associations along different sensor modalities, and objects can be recognized from the characteristics of such associations. Our approach can differentiate objects from both their visual and acoustic backgrounds by finding pixels and frequency bands (respectively) 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 9 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 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.



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. 9. 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.

5.2. *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[13]. More generally, there is evidence that the cues used to detect periodicity can be quite subtle and adaptive[14], 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 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,

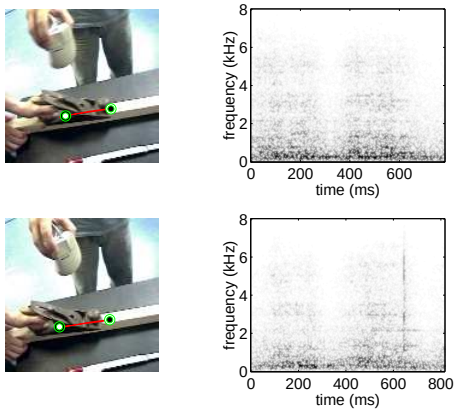


Fig. 10. An experiment in which a plane is being pushed across wood while a mouse is shaken in the background. Shown are the highest quality acoustic matches for this sound (right) and the object with which they correspond (left). Matches against the mouse are much lower and below threshold.

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 10 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.

6. 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 interac-

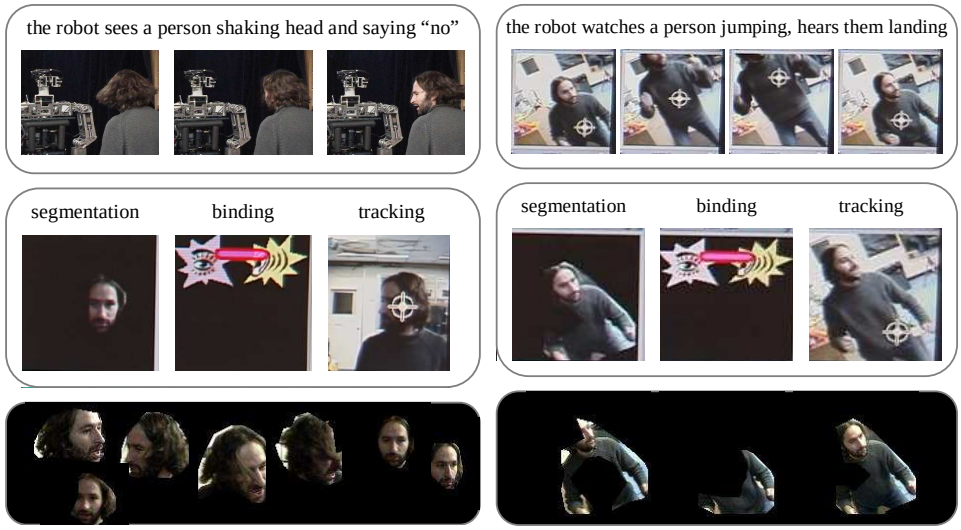


Fig. 11. (left) In this experiment, the robot sees people shaking their head. In the top row, the person says "no, no, no" in time with his head-shake. The middle row shows the recorded state of the robot during this event – it binds the visually tracked face with the sound spoken. Recorded segmentations 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).

tions with Cog, then it can use those cues to link visual input with corresponding sounds. For example, Figure 11 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 11 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. But we can also deal with 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).

7. Conclusions and Discussion

We described techniques to detect periodicity of signals, identifying their strengths and limitations. Through the detection of visual periodic events, we were able to localize an object in the visual field and extract information concerning its trajectory over time, as well as to segment a visual representation of an object from an image. In addition, sound segmentation - the identification of the frequency bands that best characterize an object - was also possible from just acoustic information. A cross-modal strategy to period detection proved necessary and advantageous, being more robust to disturbances either in sound or vision, and providing a better characterization of objects. We discussed how to reliably bind the visual appearance of objects to the sound that they generate, and to achieve selectivity: a visual distractor was filtered out using sound, and sound was also used to prime the visual field. In addition, we argued that the cross-modal strategy is well suited for integration with object recognition strategies for searching visually for tools and toys and finding/recognizing them whenever their sound is perceived by the robot.

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).

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. 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.
5. R. Polana, R. C. Nelson, Detection and recognition of periodic, non-rigid motion, *International Journal of Computer Vision* **23** (1997) (3) 261–282.
6. A. Arsenio, Embodied vision - perceiving objects from actions, *IEEE International Workshop on Human-Robot Interactive Communication* (2003).
7. P. Fitzpatrick, From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot, Ph.D. thesis, MIT, Cambridge, MA, 2003.
8. 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).
9. A. M. Arsenio, An embodied approach to perceptual grouping (2004) Accepted to the IEEE CVPR Workshop on Perceptual Organization in Computer Vision.
10. 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.
11. A. M. Arsenio, Map building from human-computer interactions (2004) Accepted to the IEEE CVPR Workshop on Real-time Vision for Human Computer Interaction.
12. M. Turk, A. Pentland, Face recognition using eigenfaces, in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (1991) .
13. 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.
14. 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.

Artur M. Arsenio received his M.S. degree in Electrical and Computer Engineering from the Technical University of Lisbon in 1998, and his Ph.D. from the MIT Computer Science and Artificial Intelligence Laboratory in 2004. 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 received 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.