The Representation of Space and the Space of Representation: a Cognitive Science Introduction to JIGSAW

Mark W. Peters (markpeters@cse.unsw.edu.au)
School of Computer Science and Engineering
University of New South Wales, Sydney NSW 2052 Australia

Barry Drake (bdrake@cse.unsw.edu.au)
School of Computer Science and Engineering
University of New South Wales, Sydney NSW 2052 Australia

Abstract

This paper provides an introduction to JIGSAW, an algorithm for self-organising spatial mapping, capable of mapping unknown sensory systems to unknown motor systems. For computer scientists, scalability comparisons with Self Organising Kohonen Maps and Multi-Dimensional Scaling are provided. For Neuro-scientists, a model of the organisational processes resulting in retinotopicity and sensory and motor homuncularity is given. For Roboticists, a robust method of continuous self-calibration is described, and for Psychologists, a large-scale multi-variate tool of analysis is offered.

Introduction

Space is an all-pervasive abstraction, fundamental to the visual, haptic and auditory senses, as well as motor control. Spatial mapping underlies navigation, identification of the extension of, and boundaries between, objects and regions. Indeed, space is implicitly so well understood that it is used as a metaphor for aiding understanding of many non-spatial abstractions, such as social and semantic relationships. For all these reasons, spatial awareness is often taken for granted, and just assumed to be present in intelligent systems. Only once spatial awareness is absent or dysfunctional, are we faced with the problem of how to create and maintain it.

The problem we pose in this paper is how to derive spatial mappings *a priori* – with no prior concepts of distance, direction, locality, and their derivatives such as up, down, further, between, around, etc.

The significance of this problem is that once it is solved three possibilities open up. First, a disembodied and universal algorithm, placed between unknown sensory systems and unknown motor systems, can organically merge them and assume purposive control of the motor systems. Second, we may obtain an understanding of how biological systems develop spatial awareness, through processes operating at an inter-neural level. Third, as a side effect we obtain a new analytical tool, capable of organising data sources in a useful way.

Background

Within the cognitive sciences, there have been two important approaches to the generation of spatial representations. We briefly describe each.

Kohonen's Self-Organising Maps

SOMs are commonly used for dimensional reduction but can also be used for the construction of spatial representations (Ritter, 1990). To do this SOMs take as input a set of time-varying signals and produce as output a bound, sampled, convergent and topographic map of the signals. Each of the cited characteristics of the map is problematic. Being bound, it is the shape of the output format that determines the shape of the data in the representation, rather than the other way around. Being sampled, the output format determines the resolution of the mapping, particularly problematic for space-variance in the density of the data. Being convergent, the mapping is unable to cope with large reconfigurations of the input data during its learning phase. Being topographic, potentially valuable metric information is discarded.

Multidimensional Scaling

MDS takes as input a set of dissimilarity measures between nodes in a graph, and produces as output multidimensional coordinates for each node (Young & Hamer, 1987). The coordinates are those that best respect the input dissimilaritities. MDS is free of the problems we associate above with SOMs. However, it shares with SOMs a difficulty in dealing with very large sets of inputs.

Both MDS and SOMs are order $O(n^2)$ in computational complexity, which limits their effectiveness to problems involving only a few thousand inputs. Sensory systems, unfortunately, are characterized by their very large number of inputs, and are therefore difficult to model using either MDS or SOMs. This is a problem.

Research in Spatial Redundancy

The key to the problem of complexity is that spatial data is largely redundant, and a sample of inputs can be used to organize the whole set. Before addressing our methods for dealing with this, it is necessary to note that an awareness of spatial redundancy has been documented in a variety of fields.

Spatial redundancy in television signals was documented in 1952 as offering the potential for data compression (Kretzmer, 1952). Later, spatial redundancy was shown to be the characteristic that allowed lateral inhibition in the vision systems of flies to work (Srinivasan, McLaughlin and Dubs, 1982). Anisotropic spatial redundancy has been used to explain the horizontal-vertical illusion (Baddeley, 1997). In the estimation of spatial distributions of, among other things, ore deposits, fish stocks, and meteorological effects, methods have been developed for measuring samples at known points and spatially interpolating and extrapolating these measurements to derive maps (Cressie, 1993).

However, two further conceptual steps must be taken before we are able to frame methods for creating spatial representations from scratch. First, though previous research recognizes that there is often an inverse correlation between spatial distance and similarity of measurement, in general, known distances have been used to estimate measurements. The converse, using differences in measurements to estimate distances, has not been done. Second, differences in measurements were assumed to be of static 'scenes' in which the goal was to map the data rather than the system or apparatus that delivered it. In the second conceptual step, it is necessary to use differences in measurements of *time-varying signals* to produce a map that describes only the relative disposition of the signal sources, not the state of the data at any particular moment. It is simple, once signal sources have been mapped, to use their locations to map any instantaneous data that they carry.

The second step, in biological terms, is perhaps analogous to the development of sensory homuncularity in the brain, whereby maps of the surface of the skin develop, and in turn allow us to map and localize effects that are felt through the skin. The same applies to retinotopicity, which allows us to perceive visual maps of our surroundings as read on the surface of the retinae. It is not necessary that such spatial structures form in the brain (Koenderink, 1990), but it appears physically economical to organize sensory systems this way, as we shall show.

Methods

The process of deriving real-time spatial mappings can be broken down into three parallel sub-processes. First, for a given pair of signals, it is necessary to maintain some measure of difference, or similarity, depending on how we wish to express it, and based on this, an estimate of the distance between the sources of the signals. Second, we must iteratively adjust our spatial representation such that it reflects current distance estimates. Third, in order to overcome the problems of quadratic computational complexity, the overall system must dynamically optimize, i.e. intelligently restrict, its selection of data for processing.

Process 1: The *ab* Relationship

Following our previous work, we shall refer to the actual distance between two signal sources as a, and the behavioral difference of the signals as b (Peters and Drake, 2000). Our method assumes that as a increases, so does b, reflecting the common sense notion that the further apart we take our measurements, the greater the likely difference between them. This monotonic relationship is termed the ab relationship. We can express the ab relationship as

$$b = f(a) \tag{1}$$

where f is an invertible function.

Our experiments and those of others have shown that while there is variability in the ab relationship, it is approximated closely by an exponential curve (Kretzmer, 1952; Peters and Drake, 2000; Srinivasan, McLaughlin and Dubs, 1982). Hence

$$b = \mu\left(1 - e^{-\lambda a}\right) \tag{2}$$

where λ and μ are scaling constants.

In practice, each scene, image, movie, or data set has a characteristic ab relationship that is determined by a mix of scales; in other words, λ is usually a distribution of values rather than a single value. The λ distribution is capable of describing signal noise and many other effects, but for simplicity we can retain a single scaling value of λ and add a noise constant, v, to the equation:

$$b = \mu - (\mu - v)e^{-\lambda a} \tag{3}$$

Which, solved for a leads to

$$a = -\left(\frac{1}{\lambda}\right)\ln\left(\frac{\mu - b}{\mu - v}\right) \tag{4}$$

This ab relationship is illustrated in Figure 1.

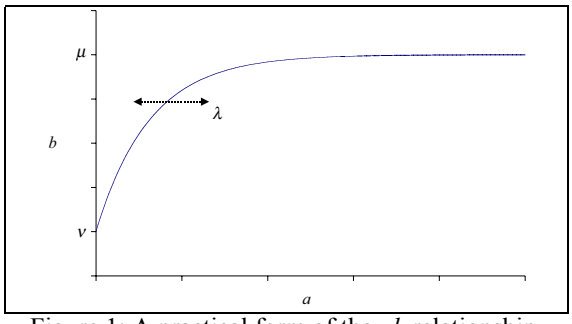


Figure 1: A practical form of the *ab* relationship.

For a given measure of the behavioral difference of two signals, we can estimate the actual distance between their sources. For b, we chose Pearson's coefficient of correlation on an exponentially decaying moving average.

Process 2: Multidimensional Scaling

To make a map, of any dimensionality, it is efficient to imply a space via coordinates assigned to each object in the space. The objects, in this case, are point-like representations of each signal source, and are termed *locators*. Each pair of locators selected for data gathering and processing is called a *dyad*. The geometrical distance of two locators in the representation is termed g. Not stored explicitly, this measure is calculated from locator coordinates.

The a values of dyads are appropriate input to an MDS process. The role of MDS is to adjust g values until they agree with a values. We use a gradient descent form of MDS in a Euclidean model of space. In simple terms, MDS moves each locator in a dyad (i, j) towards the other by a vector, $\mathbf{h}_{i,j}$, in proportion to the distance between them, and scaled by $1 - a_{i,j} / g_{i,j}$. The scaling is negative when the locators are too close together, and positive when they are too far apart. Thus

$$\mathbf{h}_{i,j} = -k\left(\mathbf{x}_i - \mathbf{x}_j\right)\left(1 - \frac{a_{i,j}}{g_{i,j}}\right) \qquad (5)$$

where k controls the descent rate.

The MDS process is intended to be continuous, so more recent b values have greater impact on the spatial map than older b values. This allows the map a degree of adaptability, so that if the configuration of signal sources changes, the layout of locators changes too. There is a trade-off between stability and flexibility, as the mapping can be made highly reactive and volatile, or very stable and unresponsive. In practice, we set the system's adaptability so that changes in the configuration of signals sources are reflected in the disposition of locators after a few minutes.

Process 3: Dynamic Data Selection

JIGSAW's chief advantage over both MDS and SOMs is that it is of linear computation complexity: order $O(n)$. Rather than attempting to process all inter-signal statistics, only a small subset are considered. Initially, dyads are selected at random from the set of all possible pairs. However, from Figure 1, it is apparent that smaller b values are potentially more reliable as predictors of a values; as the ab curve approaches the asymptote, μ, small discrepancies in b produce larger errors in a. Consequently, spatial representations composed largely from low b values suffer less from uncertainly and ambiguity. It is obvious that if a subset of pairs is to be selected for processing, it is advantageous if the selected dyads produce a predominance of low b values. Note that this is not the only basis for preferring one pair over another, but it has immediate practical value.

If a dyad produces a high b value, in absolute terms or relative to other dyads involving a specific locator, a replacement dyad can be sought. With no knowledge of the actual distance between signal sources, it is impossible to determine, in advance, which dyads will produce low b values. However, as MDS constantly improves the fidelity of the representation, each g value approaches its corresponding a value. Thus, g provides an ever more accurate indication of a, and a low g value can be used as a criterion for selection of a new dyad.

Process Summary

It is implicit in the description of each process that all three can operate stochastically and independently. Consequently, each dyad can operate in parallel, fully isolated from information in any other dyad.

Experiments

For signals, these experiments used time-varying pixel values in image sequences. The advantage of pixels is twofold. First, they are numerous, allowing realistic tests with over 65,000 simultaneous signals to be conducted. Second, pixels have a known spatial arrangement that makes an effective reference against which to compare the results.

JIGSAW was fed constantly changing values for 65,000 pixels, but was not given information about the location of any pixel in the input sequence. The task given to JIGSAW is not to reconstruct any particular image, but to recover the relative spatial relationship of the pixels. As already pointed out, once this spatial mapping has been derived, any image or movie fed into it, as a set of values, will be reproduced correctly.

Figure 2: Test image.

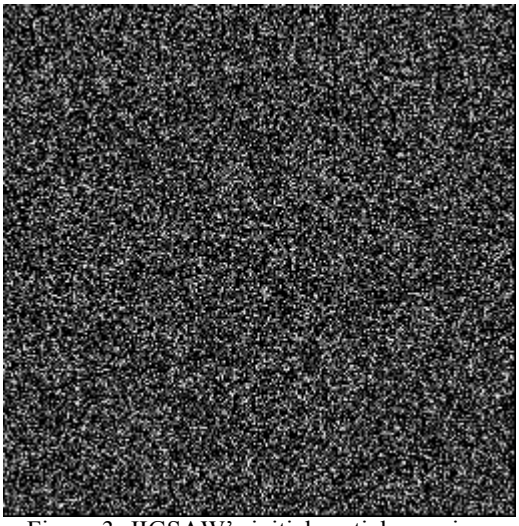


Figure 3: JIGSAW's initial spatial mapping.

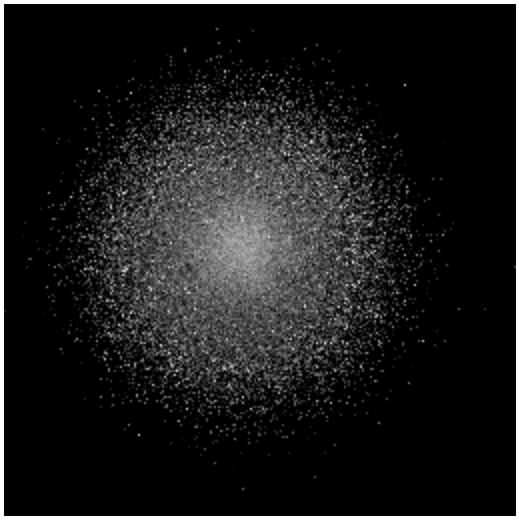
While there are several measures that indicate how closely a spatial map matches an original, they are not meaningful as indicators of how useful such a mapping might be to a robot or an organism. We have therefore chosen to illustrate the results of JIGSAW graphically, by using the test image in Figure 2 as a consistent set of pixel values. It has the same number of pixels as the input data that was fed into JIGSAW. To illustrate the fidelity of JIGSAW's mappings, each pixel is mapped to a single locator. The image will reappear perfectly only when JIGSAW has fully recovered a high fidelity mapping of the pixel's correct positions.

In each experiment, the initial spatial mapping, before any information has been received from signal sources, is completely disorganized. When the pixel values of the test image are fed to each locator, the image is unrecognizable in the mapping (see Figure 3).

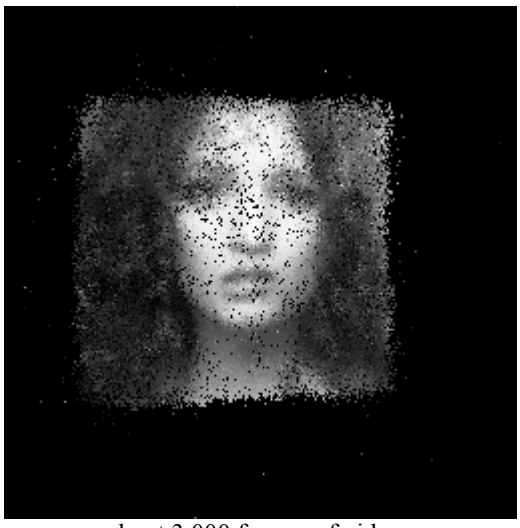
The first experiment (see Figure 4) was based on simple input images generated on the fly. In Figure 4a, JIGSAW's initially random placement of locators has begun to assume some degree of spatial structure. The structure progressively improves to near-perfect (Figure 4c). NB: the mappings in Figures 4b and 4c have been manually rotated to have the familiar orientation.

The appearance of black gaps in the mapping in Figure 4c is a quantizing effect due to the difficulty of representing a stochastic arrangement of locators in a perfect orthogonal grid. A magnified section of the mapping reveals that the fidelity of the results is much higher than might be imagined (see Figure 5).

The second experiment (see Figure 6) uses a 25-frame-per-second movie of carelessly filmed outdoor scenes.



a: at 1,000 frames of video



b: at 3,000 frames of video

c: at 322,000 frames of video

Figure 4: Experiment 1 spatial mapping.

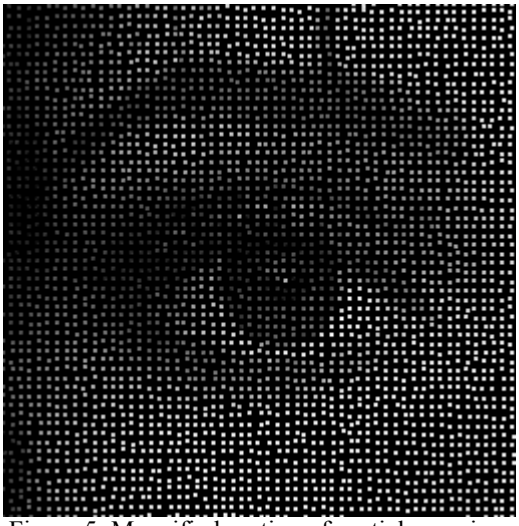


Figure 5: Magnified section of spatial mapping.

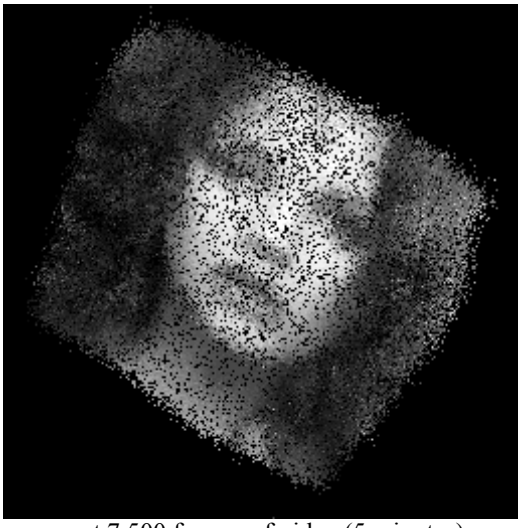
In this experiment, the gross spatial mapping is already evident after 5 minutes of video (see Figure 6a), and the subsequent refinement of detail illustrates JIGSAW continuous dynamic data selection – low *b* values giving low *a* values, producing visibly more accurate estimates of actual representations of the distances between signal sources. The slight irregularity of the overall mapping may be explained by statistically non-stationary anisotropy in the video data. A less reactive setting for JIGSAW might minimize such effects but at the cost of speed of organization and adaptability to change. In this experiment the orientation of the mapping has been left unchanged.

Discussion

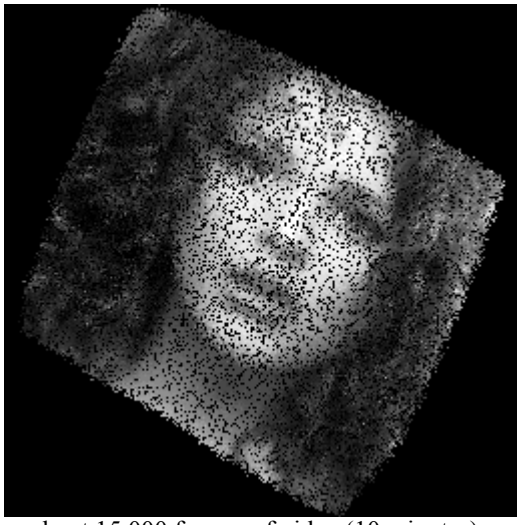
The experiments documented in this paper have been chosen to illustrate the fidelity and speed of spatial organization achieved by JIGSAW with a large set of signals.

JIGSAW is a method for generating spatial mappings. Its key contribution to any intelligent system is that it takes essentially non-spatial input and autonomously generates the prerequisites for spatial awareness. JIGSAW is domain-independent and therefore has application in several areas.

In robotics, JIGSAW has two main functions. First, it will fuse images from multiple cameras, adjusting the calibration during run-time if necessary. It can spatially arrange the input of one camera with respect to another camera, even when the two fields of view do not overlap. Second, by a simple adaptation, JIGSAW can be made to calibrate motor maps to such composite spatial representations, by treating each combination of pan and tilt position as a locator. It will perform the sensory and motor calibration *simultaneously*, starting from scratch. In order for motor commands to be represented as locators, they need to be given a virtual



a: at 7,500 frames of video (5 minutes)



b: at 15,000 frames of video (10 minutes)

c: at 22,500 frames of video (15 minutes)

Figure 6: Experiment 2 spatial mapping.

signal. This signal is the value found at the fovea (a privileged point which may be selected arbitrarily) after the command has been executed. When compared with signal values recorded just before the motor command was executed, this provides enough information to iteratively position the motor command at the point in the spatial representation that will be foveated after the next execution of that command.

In psychological analysis, MDS has had a major role in producing simple spatial representations showing the proximity of phenomenal experiences. The fact that the computational complexity of MDS remains quadratic at best means that JIGSAW has a potential role in analyzing data sets that are simply too large for standard MDS algorithms. It can be used to render spatial representations of multi-variate data to any dimensionality – JIGSAW has successfully organised data from over a quarter of a million signals.

JIGSAW offers an interesting alternative model of how neurons connect, one that clearly predicts the homuncularity and other spatial mappings of the brain. The key characteristic of the JIGSAW model is that the relative position of locators is used to represent the similarity of signals being received. Relative to the Hebbian paradigm, the proximity of two locators in a JIGSAW mapping is an indication of how closely two neurons fire together. While in the brain neural nuclei may not *physically* migrate towards each other, their processes often do, and in this way the neurons can be thought of as *logically* migrating towards each other. Being enclosed in a small space, the brain is constrained by certain physical limitations to the extent of connectivity between neurons. These limitations are somewhat ameliorated if it is possible to arrange neurons such that those that are most highly interconnected are physically close together (Mitchison, 1991). In this way, less space is given over to connecting processes, or white matter.

While some aspects of JIGSAW are not biologically plausible (the ability for two locators to occupy the same position, or the ability for one locator to be at a great distance from all others), its information processing certainly is. All processing is local. Each locator is connected to a small number of others (20 to 50). Any of its pairings can be broken if the two signals behave quite differently, and might then be replaced with a new connection, biased towards other locators in the immediate vicinity.

In summary, it should be emphasized that with JIGSAW there now exists a method for forming spatial representations that are independent of any apparatus-imposed spatial organization. The implication of this is that long-standing problems such as data fusion, shape recognition in space-variant systems, and symbol-grounding, can be approached in a new way.

In the case of data fusion, a common goal is to achieve a universal spatial mapping based on multiple local given spatial mappings (such as pixel position within a single camera). The key problem is that generally these local mappings are partially incompatible. JIGSAW reformulates the problem by discarding any assumed spatial mapping and deriving its own from signal correlation.

Similarly, space-variant sub-sampling distorts the shape of visible objects, particularly while they move through the field of view. This complicates the work of algorithms that depends on shape recognition. When JIGSAW preprocesses space-variant data the space-variance is removed and shape is rendered consistently.

JIGSAW deconstructs the sensorium of an agent, and rebuilds spatial mappings based on the most fundamental of sensory phenomena. All it requires to do this is many small independent processors with the simplicity of a single neuron – limited memory in the form of habituation, only a single level of activation, and the ability to transform correlation with the activity of its neighbors into a tangible relationship, spatial proximity. The correlative structure of the outside world therefore dictates the internal structure of the sensory system. This enables whatever derivate symbols or meta-representations the system reifies to be grounded in the world.

References

Baddeley, R. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science, 21*(3), 351-372.

Cressie, N. A. C. (1993). *Statistics for spatial data* (revised ed.). New York: Wiley.

Koenderink, J. J. (1990). The brain is a geometry engine. *Psychological Research*, 52, 122-127.

Kretzmer, E. R. (1952). Statistics of Television Signals. *The Bell System Technical Journal, 31*(4), 751-763.

Mitchison, G. (1991). Neuronal branching patterns and the economy of cortical wiring. *Proceedings of the Royal Society of London, Series B*, *245*, 151-158.

Peters, M. W., & Drake, B. (2000). *Jigsaw: the unsupervised construction of spatial representations* (Tech. Rep. UNSW-CSE-TR-0007). Sydney, Australia. University of New South Wales, School of Computer Science and Engineering.

Ritter, H. (1990). Self-organizing maps for internal representations. *Psychological Research*, 52, 128-136.

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, Series B, 216*, 427-459.

Young, F. W., & Hamer, R. M. (Eds.). (1987). *Multidimentional Scaling: History, Theory, and Applications*. Hillsdale, New Jersey: Lawence Erlbaum Associates.