ADAPT
IST-2001-37173
Artificial Development Approach to Presence Technologies

Deliverable Item 5.4
Initial Experiments with Multiple Sensory Modalities Integration

Delivery Date: August 30, 2005
Classification: Public
Responsible Person: Harri Valpola
Partners Contributed: ALL

Short Description: We have implemented a model which is ment for extracting coherent representations from multi-modal data. The model consists of a hierarchy of interconnected units. Initial experiments which verify that the model works in principle were reported in D5.3. This deliverable reports experiments with more realistic multi-modal data and discusses the background and motivation of the model.

Project funded by the European Community under the "Information Society Technologies" Programme (1998–2002)

1 Introduction

The feature extraction model we are developing is a hierarchy of basic units, feature maps, which were reported in D5.2. The goal of the model is to extract a coherent, behaviourally meaningful representation of the sensory input. The term “behaviourally meaningful” refers to the fact that the representation is supposed to support motor control, prediction of future rewards, etc. The level of abstraction in the representation is supposed to increase towards the higher levels in the hierarchy as explained in D5.1.

Deliverable D5.3 reported initial experiments with the model. The focus was in the basic principle: it was shown that context can guide the development of invariant features. More specifically, it was shown that translation invariant edge detectors (akin to complex cells in the visual cortex) will result when the objective is to find image features that carry similar information as the features extracted from nearby image locations. In natural images, edges tend to continuous but curved. A short piece of edge thus predicts an edge at nearby locations (due to continuity) but the precise position is somewhat uncertain (due to curvature). This uncertainty promoted the development of translation invariant features in the experiment reported in D5.3.

More generally, the idea is to develop features which can be predicted from the context. The context can include for instance other modalities or past observations. This deliverable shows that this principle is able to develop coherent representations in multi-modal data and extends D5.3 by discussing the background and motivation of the model. The technical aspects of the feature extraction model were described in D5.3 and are largely omitted here.

2 Background of the feature extraction model introduced in D5.3

2.1 Brief overview of the model structure

The feature maps roughly correspond to cortical areas such as V1. Each map consists of a set of adaptive feature detectors. They are roughly analogous to cortical microcolumns or individual cortical pyramidal cells in that each one responds the strongest for a particular combination of preferred inputs.

In the, model two types of inputs can be distinguished:

- Bottom-up inputs which (predominantly) determine the activations of the feature detectors but have relatively less influence on learning (long-term adaptation of the feature detectors).
- Top-down inputs which have only a weak influence on the activations of the features but are crucial in guiding learning.

This arrangement was, on the one hand, motivated by the need to improve upon the current feature extraction models which are usually based on unsupervised learning and, on the other hand, inspired by the structure and connectivity of the mammalian neocortex.

2.2 Semi-supervised learning

Unsupervised learning is based on adapting feature extraction (perception) to the statistical structure of the stimuli. This allows the system to represent maximal amount of information with minimum amount of resources. As such, unsupervised learning is an indispensible component in autonomously learning systems.

The underlying assumption in unsupervised learning is that all information is useful. Clearly this is not true. In practice it is often necessary to discard much (in complex problems even most) of the sensory information and retain only the information which is essential for control, decision making, etc.

Supervised learning is basically the same thing as regression or function approximation: during learning, inputs and outputs are supposed to be known and the task is to find the function mapping inputs to outputs. The purpose can, for instance, be to find a predictor or classifier.

In supervised learning, not all information is considered equally important. The learning process is able to discover the features which are relevant to the task at hand. Unfortunately, supervised learning does not seem suited to perceptual development because the target outputs are not available during learning. Instead, engineering applications often resort to a combination of initial unsupervised learning and subsequent supervised learning. The first, unsupervised stage can use unlabelled data which is typically easy to collect. After the representation has been optimised for the typical statistical structure of the data, the second, supervised stage can work with a small amount of labelled data which is typically much more difficult to collect.

This approach starts to break down when the dimensionality of the sensory input data increases and the number of labelled samples decreases. The problem is that the first, unsupervised learning phase will need to discard so much information that the essential information gets irreversibly lost. Also, the second, supervised stage often needs to be highly nonlinear, making it difficult to estimate good mappings with few samples.

In between unsupervised and supervised learning there is semisupervised learning which operates with two data sets just as supervised learning but does not assume target outputs to be known exactly. Rather, the learning process has to identify which part of the given second data set can act as the target. There is a parallel between the different learning approaches and classical linear estimation techniques:

- Unsupervised learning corresponds to principal component analysis (PCA) which reduces the dimensionality of one data set (A) such that the original data can be reconstructed from the resulting features as well as possible (in terms of mean square error).
- Supervised learning corresponds to linear regression where one data set (A) is used to predict the values of another data set (B).
- Semi-supervised learning corresponds to canonical correlation analysis (CCA) which tries to reduce the dimensionality of two data sets (A and B) such that the extracted features would be as similar to each other as possible.

CCA uses exactly two data sets and is symmetric with respect to them. However, it is possible to assume that one of the data sets corresponds to the sensory input whose features are supposed to be learned.

With linear methods, multistage approaches do not offer significant benefits over one-stage approaches. This is because two consequtive linear mappings can be combined into one linear mapping. With nonlinearities the situation is different. The method we have developed can be considered to be an extension of CCA. Each processing stage takes bottom-up inputs A and extracts moderately nonlinear features which 1) retain information from the inputs A and 2) have correspondance with a context B. The context plays the role of supervisory signal but, like in CCA and other semisupervised approaches, it does not need to give explicitly the target outputs.

The combined effect of the weak nonlinearities can be a very strong nonlinearity but less samples should be needed than in strict separation to unsupervised and supervised learning stage. This is because each stage now implements only a relatively weak nonlinearity which is easier to learn.

The exact nonlinearity that is used should not be critical. This is analogous to the fact that in conventional multi-layer feedforward networks used in supervised learning, several different types on nonlinearities can be used. The nonlinearity could be a simple positivity constraint: negative activations were truncated to zero (this is basically what was used in D5.3). It should also be possible to combine nonlinearity with temporal filtering which would provide features which contain information about the dynamics of the inputs.

2.3 Biological relevance

Our implementation of semisupervised learning uses two stages. In the first, unsupervised stage, the bottom-up inputs are expanded nonlinearly, decorrelated and normalised by an adaptive feature extraction system. This stage can be associated with the initial bottom-up feature extraction by cortical layer 4.

The second stage uses modulatory contextual inputs to bias learning. In Hebbian learning, strong inputs usually have strong influence in learning but due to decorrelation and normalisation at the first stage, bottom-up inputs don't favour any particular representation. Even a weak top-down modulation decides how the feature detectors develop.

In the cortex, this second stage can be associated with pyramidal neurons which receive bottom-up inputs preferentially on their basal dendrites and top-down inputs on their apical dendrites. The input to the basal dendrites is the driving input for the pyramidal neurons and can activate them without any further inputs. The input to the apical dendrites cannot normally activate pyramidal neurons but instead can strongly amplify the activity triggered by the basal-dendrite input.

2.4 Topology

The term cortical *map* refers to topological representations found in various cortical areas, such as retinotopical maps in V1 or tonotopic maps in A1. While topology is likely to be important in large cortical areas (because topological arrangement minimizes cabling costs), at this stage our models corresponds to such a small brain area that topology does not seem to play an important role. For simplicity, we have restricted to non-topological feature maps, but D5.2 outlined possible mechanisms for developing topological maps. They can be expected to be useful if larger models are implemented.

3 Experimental results with climate data

The model is supposed to be generally applicable to different modalities (although it is clear that it is always possible to optimise it to a particular modality). We have applied the model to a climate data set. Although this research was not part of ADAPT, it has successfully used methods developed in ADAPT.

The application is interesting and of practical value because humans don't have an innate "climate sense". The climate phenomena are complex spatiotemporal patterns that are usually visualised by showing "weather maps". However, these maps can only show a small part of the complex 3D patterns and different variables at a time. Developing a "climate sense" for a machine therefore has the potential for discovering novel "climate objects". This is in stark contrast to modalities such as audition, vision and touch where humans excel.

As an example, we have concentrated on three different "climate modalities": temperature, pressure and precipitation (rain fall). In the climate system, these variables are related, but not in a simple manner. The relation may include temporally and spatially separated locations. The information in these different modalities is interlapping but in "different coordinate systems" in much the same way as for instance speech can be perceived both in visual and auditory modalities.

We used a so-called reanalysis data set which has virtual measurement points over the whole globe on a grid with 2.5° spacing. Exactly these measurements were never really made but they represent climatologists' best estimate based on the available measurements.

The grid has more than 10,000 measurement points. So far we have only used one altitude although the data set includes 17 of them. From each point, we used three (out of the six available) different daily average measurements: temperature, pressure and precipitation. This means that there are over 30,000 measurements each day. The data set covers more than fifty years which is more than 20,000 days.

The model we used was the simplest possible implementation of the basic principle:

- Three separate principal component analyses were performed, one for each modality. This extracts linear features reflecting the most prominent statistical structure of individual modalities. Learning is purely unsupervised.
- Coherent spatio-temporal patterns are extracted from the three modalities using temporal filtering and PCA. This corresponds to finding those features from individual modalities which can be predicted from all modalities and past and future inputs.

The process resembles canonical correlation analysis (CCA) which is defined for two data sets. Our approach includes more than two modalities and temporal information. The result is still similar to CCA in the sense that each modality develops the same set of features. This means that the resulting representation is basically amodal.

The above process relies on PCA which is a linear method. This means that the features are normally not meaningful individually. Each underlying "climate object" is reflected in more than one feature. It is, however, relatively easy to rotate the feature space such that climate objects become aligned with coordinate axis. In other words, each climate object can be identified with one or few rotated feature components.

Rotation with methods resembling independent component analysis recovered the following climate objects:

- Annual oscillation: seasonal variation.
- Long-term changes: continuous trend is the strongest of these.
- Mid-term objects: El Niño and other similar phenomena.

Everybody knows about seasonal variation and most of the other found phenomena, too, are well known to climatologists. However, we did find a few interesting phenomena that are apparently unknown or at least not widely known among climatologists. As climatology is not really a central topic of ADAPT, I will not go into much detail but I will raise the following points:

- The representation is coherent and amodal: each one of the above climate objects is somehow apparent in every modality but the temporal and spatial patterns are quite different.
- The amount of data in these experiments is quite large but the bio-inspired approach developed in ADAPT can handle it easily.
- The model was now completely linear and some climate parameters (such as wind speed and direction) were omitted.
- Even these preliminary results were interesting for climatologists and our research has therefore opened interesting new possibilities for data analysis. Obvious future research directions include using more data, nonlinear feature expansion and, perhaps, a multilayered model.

4 Discussion

It is clear that the structure of the neocortex is far more complex than the simple model we are studying. Nevertheless, it seems that the kind of semisupervised learning envisioned here is consistent with the physiology of the neocortex. At least the model gives falsifiable hypothesis about the structure and function of the neocortex. For instance, it is known that primates have a cortical area which responds to visual features of hands. Since activity in this area is triggered by vision, we would hypothesise that the bottom-up inputs, terminating on layer 4, would derive from visual areas. Sensitivity to hand features would suggest that the top-down inputs, preferentially terminating on the apical dendrites on layer 1, are derived from motor and/or proprioceptive hand areas.

5 References

As the purpose of this deliverable is rather to introduce the articles written within ADAPT than to be a stand-alone publication, I will only list pointers to relevant publications that accompany this deliverable.

5.1 Articles which acknowledge ADAPT

H. Valpola. Development of representations, categories and concepts—a hypothesis. In Proceedings of the 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA 2005), Espoo, Finland, 2005.

J. Särelä and H. Valpola. Denoising source separation: a novel approach to ICA and feature extraction using denoising and Hebbian learning. In AI 2005 special session on correlation learning, Victoria, British-Columbia, Canada, pp. 45–56, 2005.

J. Särelä and H. Valpola. Denoising source separation. Journal of Machine Learning Research, 6:233–272, 2005.

H. Valpola and J. Särelä. Accurate, fast and stable denoising source separation algorithms. In Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004), Granada, Spain, pp. 65–72, 2004.

H. Valpola. Behaviourally meaningful representations from normalisation and context-guided denoising. AI Lab technical report, University of Zurich, 2004.

5.2 Others (climate-data analysis)

A. Ilin and H. Valpola. Frequency-Based Separation of Climate Signals. Accepted to the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005), Porto, Portugal.

A. Ilin, H. Valpola and E. Oja. Semiblind source separation of climate data detects El Niño as the component with the highest interannual variability. In Proceedings of the International Joint Conference on Neural Networks (IJCNN 2005), Montréal, Québec, Canada, pp. 1722–1727, 2005.