On the Integration of Grounding Language and Learning Objects

Chen Yu and Dana H. Ballard
Department of Computer Science
University of Rochester
Rochester, NY, 14620
{yu,dana}@cs.rochester.edu

Abstract

This paper presents a multimodal learning system that can ground spoken names of objects in their physical referents and learn to recognize those objects simultaneously from naturally co-occurring multisensory input. There are two technical problems involved: (1) the correspondence problem in symbol grounding – how to associate words (symbols) with their perceptually grounded meanings from multiple co-occurrences between words and objects in the physical environment. (2) object learning – how to recognize and categorize visual objects. We argue that those two problems can be fundamentally simplified by considering them in a general system and incorporating the spatio-temporal and cross-modal constraints of multimodal data. The system collects egocentric data including image sequences as well as speech while users perform natural tasks. It is able to automatically infer the meanings of object names from vision, and categorize objects based on teaching signals potentially encoded in speech. The experimental results reported in this paper reveal the effectiveness of using multimodal data and integrating heterogeneous techniques in machine learning, natural language processing and computer vision.

Introduction

Intelligent machines, such as humanoid robots, are expected to be situated in users' everyday environments, communicate with users using natural language and provide services. To achieve this goal, machines must have the same perceptual and cognitive abilities as humans, such as visual object recognition and language understanding. However, sensory perception and knowledge acquisition of machines are quite different from those of human counterparts. The first difference is about *what* to learn. Machine learning is most often disembodied and focuses on manipulating symbols intelligently, but humans develop and learn based on their sensorimotor experiences with the physical environment. To mimic human perceptual abilities and ultimately build artificial embodied systems, a challenge in machine intelligence is how to establish a correspondence between internal symbolic representations and external sensory data, which is termed as the symbol grounding problem by Harnad (1990).

The second difference lies in *how* to learn. Researchers have tried different approaches to building intelligent machines (see a review in Weng *et al.*,2001). A knowledge-based system uses logic rules to represent the knowledge acquired from human experts and perform inference. A learning-based approach applies probabilistic models and then uses "spoon-fed" human-labeled sensory data to train the parameters of those models. In this way, the machine acquires some basic perceptual abilities, such as speech recognition (converting speech to text) and object recognition (converting visual signals into pre-defined labels). A relatively new approach termed *autonomous development* is proposed by several scientists in different fields (Weng *et al.* 2001). In this approach, a brain-like artificial embodied system develops based on real-time interactions with the environments by using multiple sensors and effectors. Although this approach shares many computational techniques with the learning-based one, these two are fundamentally different. In autonomous development, the machine continuously develops and learns to simulate the lifelong process of development and learning in humans (Thrun & Mitchell 1995). Thus, it does not need human involvement for segmenting and labeling data during the learning process. In contrast, most learning-based systems apply one-time training on labeled data and are not able to acquire any new capabilities after the training phase.

We report here on the first steps toward autonomous development and learning while focusing on grounding spoken language in sensory perception. People use words to refer to objects in the physical environment. To communicate with users and provide helps, a computational system also needs to know how to categorize instances of objects (visual object recognition) as well as map objects to linguistic labels. Despite intensive research in computer vision, the first problem is still an open issue due to practical reasons, such as the variations in lighting conditions in the environment. The second problem, termed *reference uncertainty* (Quine 1960), deals with finding the correspondence between words and perceptually grounded meanings from multiple co-occurrences of words and objects in a natural environment. To tackle these two problems, we develop a multimodal learning system that is situated in users' everyday environments wherein users introduce several objects to machines using natural language just like teaching their chil-

dren or familiarizing newcomers with the environment. The system collects user-centric multisensory input consisting of visual and speech data, and automatically learns to build a word-world mapping in an unsupervised manner without human involvement. In addition, the system also learns to categorize objects based on the corresponding linguistic labels in speech.

Related Work

The symbol grounding problem has been studied in the context of language acquisition in a number of recent works. The algorithm in Siskind (1995) was based on cross-situational learning and can successfully recognize various event types described by predicate logic. Regier (1996) proposed a connectionist system encoding domain-motivated structure, which was able to learn the meanings of prepositions (e.g., above, in, on and through). Steels & Vogt (1997) reported the experiments in which autonomous visually grounded agents bootstrap meanings and language through adaptive language games. They argued that language is an autonomous evolving adaptive system maintained by a group of distributed agents without central control and a lexicon may adapt to cope with new meanings that arise. Cohen *et al.* (2002) demonstrated that a robot can learn the denotational meanings of words from sensory data. Roy & Pentland (2002) used the correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. The learning algorithm was based on cross-modal mutual information to discover words and their visual associations.

A relatively new topic termed anchoring (Coradeschi & Saffiotti 2003a) concerns grounding object names in sensory data and maintaining the word-world mapping to take into account the dynamic changes of sensory input. Different from the symbol grounding problem that aims at considering all kinds of symbols, anchoring focuses on a practical problem of connecting object names with visual objects in a physical environment. The recent progresses in anchoring can be found in the special issue of Robotics and Autonomous Systems on perceptual anchoring (Coradeschi & Saffiotti 2003b).

System Overview

Figure 1 shows an overview of our system. The multisensory input consists of image sequences, each of which contains several objects in the scene, and speech that contains spoken names of objects. To achieve the goal of grounding object names and learning to categorize visual objects simultaneously, the system consists of the following three modules:

- **Natural language processing** first converts acoustic signals into transcripts, and then extracts a list of object name candidates using syntactic and semantic constraints.
- **Visual processing** detects scene changes in the video captured from a head-mounted camera. It then finds the objects from the scene. Next, visual features are extracted from objects and those perceptual representations are labeled with temporally co-occurring object name candidates to form many-to-many word-meaning pairs.
- **Multimodal integration** is the crucial step in which information from different modalities is integrated to discover word-object correspondences. In addition, linguistic information provides teaching signals to guide the grouping of visual features in their space. As a result, the system obtains not only a mapping of words to their grounded meanings extracted from visual perception, but also a better categorization of visual objects compared with unimodal clustering purely based on visual features.

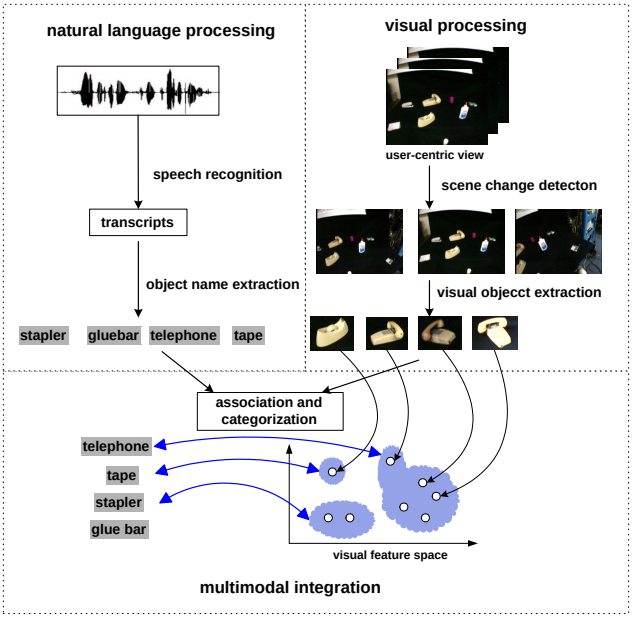

Figure 1: **Overview of the system.** Speech processing provides object name candidates based on speech recognition and natural language processing. Visual processing obtains a set of visual objects. The system associates object names with their visually grounded meanings and uses these linguistic labels to guide the categorization of objects in the visual feature space.

The following sections provide technical descriptions of the design and implementation of those subsystems in detail. The overall system is evaluated using the data that is collected when users perform natural tasks in everyday scenarios. The experimental setup is described and results are discussed.

Natural Language Processing

Object name candidates are extracted from speech using lexical and grammatical analysis shown in Figure 2. First, the "Dragon Naturally Speaking" software is utilized for speech recognition. Given a spoken utterance consisting of several spoken words, the speech recognizer converts the continuous wave pattern into a series of recognized words by considering phonetic likelihoods and grammars. In practice, the recognition rate is above 90% in our experiments. We then extract the nouns from transcripts, which will be used as object name candidates. To do so, we first use the Link Grammar Parser (Sleator & Temperley 1993), a syntactic parser,

to assign an input sentence with a syntactic structure consisting of a set of labeled links. The parser also labels the syntactic categories of the words in the sentence if they belong to nouns, verbs, or adjectives in the dictionary used by the parser. In this way, we discover a list of nouns from each sentence. Next, we apply two filters to get object name candidates as follows:

- The Oxford text archive text710 dictionary (Mitton 1992), which includes more than 70,000 words, is applied to spot proper nouns. A word is selected to be a name candidate if it is annotated as a countable noun or both countable and uncountable nouns in the dictionary.
- The WordNet (Fellbaum 1998), an electronic lexical database, is used to obtain underlying lexical concepts of candidate nouns. The system selects the nouns that belong to "physical object" or "entity".

The final list of object name candidates contains the words that satisfy both constraints.

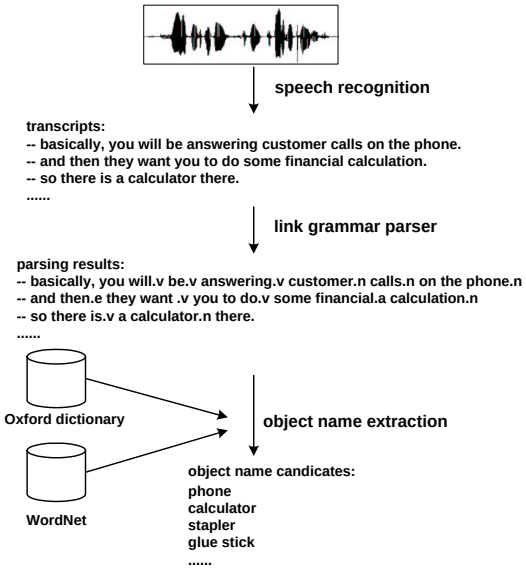

Figure 2: The acoustic signals are first converted into text. We then use syntactic and semantic constraints to obtain a list of object name candidates.

Object Detection and Description

Image sequences are captured from a head-mounted camera to get a dynamic view of a speaker. We use a position sensor to track the motion of the speaker's head. At every time point that the head is stable, we capture a snapshot of the scene and label it with temporally co-occurring spoken utterances. One of the most challenges in computer vision is the segmentation of objects in a natural scene. This issue is easily handled in our experiment in which images have a simple uniform background. Figure 3 shows a sample of snapshot and the result of object segmentation using the method described in (Comanicu & Meer 2002). Next, each extracted object is represented by features including color, shape and texture properties. Based on the work of Mel (1997), we construct visual features of objects which are large in number, invariant to different viewpoints, and driven by multiple visual cues. Specifically, 64-dimensional color features are extracted by a color indexing method (Swain & Ballard 1991), and 48-dimensional shape features are represented by calculating histograms of local shape properties (Schiele & Crowley 2000). The Gabor filters with three scales and five orientations are applied to the segmented image. It is assumed that local texture regions are spatially homogeneous, and the mean and the standard deviation of the magnitude of the transform coefficients are used to represent an object in a 48-dimensional texture feature vector. The feature representations consisting of a total of 160 dimensions are formed by combining color, shape and texture features, which provide fundamental advantages for fast, inexpensive recognition. Most pattern recognition algorithms, however, do not work efficiently in high dimensional spaces because of the inherent sparsity of the data. This problem has been traditionally referred to as the dimensionality curse. In our system, we reduced the 160-dimensional feature vectors into 15-dimensional vectors by principle component analysis (PCA), which transforms the data in a lower dimensional subspace that contains as much of the variance in the data as possible.

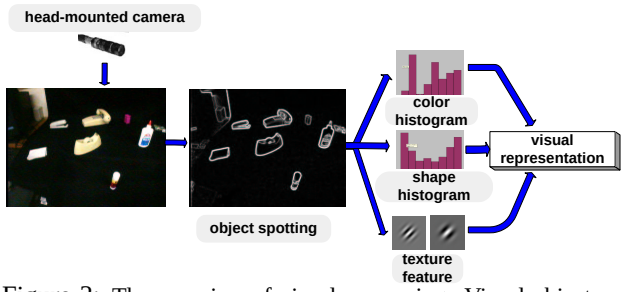

Figure 3: The overview of visual processing. Visual objects are segmented from the background scene and then multiple features are extracted to form perceptual representations.

Generative Correspondence Model

Now we have object name candidates from the audio stream and perceptual representations of objects in the visual stream, which form many-to-many co-occurring word-meaning pairs shown in Figure 4. Machine learning techniques have been widely applied to process multimodal data. Most works focus on either the correspondence problem or the clustering problem. The correspondence problem deals with building a map between items from acoustic and visual data. In Duygulu *et al.* (2002), images were segmented into regions that are then classified into region types. A mapping between region types and the keywords in captions supplied with the images was then learned using a method based on machine translation. Wachsmuth & Sagerer (2002) used Bayesian networks to relate verbal and visual descriptions of the same object. Satoh *et al.* (1997) developed a mathematical description of co-occurrence measurement to associate names with faces that appear in televised news reports.

The multimodal clustering problem studies the influence of the data in one modality on another one based on the assumption that multisensory input is perfectly synchronized in time. de Sa & Ballard (1998) related this idea mathematically to the optimal goal of minimizing the number of misclassifications in each modality and applied it to derive

an algorithm for two piecewise linear classifiers, in which one uses the output of the other as supervisory signals. Hofmann & Puzicha (1998) proposed several statistical models for analyzing the data that are joint occurrences of pairs of abstract objects from two finite sets.

Different from previous studies, we consider the correspondence and the clustering problems simultaneously. To build explicit correspondences between visual objects and words, we take a view of the data in term of a generative process. It first generates a latent variable, and then visual objects are generated based on the latent variable. Finally, words are generated conditioned on visual objects. This is because verbal descriptions are produced based on visual objects in the scene.

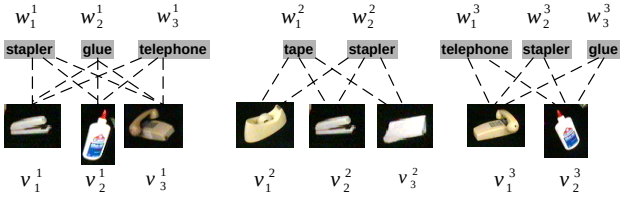

Figure 4: **Word learning.** The object name candidates in speech and co-occurring perceptual representations of objects are temporally associated to build possible lexical items.

Formally, we collect multiple pairs of co-occurring speech and visual data that can be represented by a set $S = \{S_1, S_2, ..., S_L\}$. Each $S_i (1 \leq i \leq L)$ consists of several linguistic items and visual features $\{w_1^i, w_2^i, ..., w_{m_i}^i, v_1^i, v_2^i, ..., v_{n_i}^i\}$, where $w_{m_i}^i$ represents a spoken word and $v_{n_i}^i$ is a visual feature vector extracted from an object's appearance. Given the data S, we need to maximize the log-likelihood expressed as follows:

$$\log P(S) = \sum_{i=1}^{L} \log P(S_i)$$

where

$$\begin{aligned} logP(S_i) &= logP(w_1^i, w_2^i, ..., w_{m_i}^i, v_1^i, v_2^i, ..., v_{n_i}^i) \\ &\approx \log \prod_{n=1}^{n_i} p(v_n^i) \prod_{m=1}^{m_i} \sum_{n=1}^{n_i} p(w_m^i | v_n^i) \end{aligned}$$

Two assumptions are made here. The first is that each individual object in a scene is generated independently. The second one claims that the influence of each object on a word is independent. Thus, the generative probability of a word given a visual scene equals to the sum of the condition probabilities of the word given each individual object in the scene. Now let us assume that the visual data is generated by a mixture model that consists of K components. Specifically, a visual feature is generated by first selecting a mixture component according to the mixture weights $p(\alpha)$ (class prior probabilities), then having this selected mixture component to generate the visual feature according to its own parameters with the distribution $p(v_n^i|\alpha)$. In addition, we assume that v_n^i and w_m^i are independent given the latent label α. Based on that, the log-likelihood can be expressed as:

$$\log \prod_{n=1}^{n_i} p(v_n^i) + \log \prod_{m=1}^{m_i} \sum_{n=1}^{n_i} \sum_{\alpha=1}^{K} p(w_m^i|\alpha) p(\alpha|v_n^i) \quad (1)$$

To overcome the difficulties in maximizing a log of a sum, we introduce a set of latent variables and use the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin 1977) that will iteratively increase the likelihood and make it converge to a local maximum.

Let $R_{n\alpha}^i$ be an indicator variable to represent the unknown class α from which the observation v_n^i is generated. A set of latent labels form a Boolean matrix $\{R_{n\alpha}^i\}_{n_i \times K}$ where $\sum_{\alpha=1}^{K} R_{n\alpha}^i = 1$, which indicates that only one item in each row is 1 and all others are 0. In this way, latent variables partition the visual data into K clusters and can treated as additional unobserved data. We also need to introduce the other set of latent labels $\{Z^1, Z^2, ..., Z^L\}$, each of which is a Boolean matrix $\{Z_{nm}^i\}_{n_i \times m_i}$ indicating whether a word is generated by a specific visual feature in ith data pair with the constraint $\sum_{n=1}^{n_i} Z_{nm}^i = 1$. By treating R and Z as additional unobserved data, the complete data of the second item in Equation 1 is given by:

$$\sum_{m=1}^{m_i} \sum_{n=1}^{n_i} Z_{nm}^i \sum_{\alpha=1}^{K} R_{n\alpha}^i \log(p(w_m^i|\alpha) p(v_n^i|\alpha) p(\alpha))$$

As a result of introducing new latent variables, we decouple the parameters we want to estimate. In the E-step, the expected values of the posterior probabilities of $R_{n\alpha}^i$ and Z_{nm}^i can be estimated as follows:

$$R_{n\alpha}^i = \frac{p(v_n^i|\alpha)p(\alpha)}{\sum_{k=1}^{K} p(v_n^i|k)p(k)} \quad Z_{nm}^i = \frac{p(v_n^i|w_m^i)}{\sum_{\gamma=1}^{n_i} p(v_\gamma^i|w_m^i)} \quad (2)$$

In the M-step, we calculate the derivative of Equation 1 using the estimated hidden variables and adding the normalization constraints by Lagrange multipliers. The class prior probability becomes as follow:

$$p(\alpha) = \frac{\sum_{i=1}^{L} \sum_{n=1}^{n_i} R_{n\alpha}^i (\sum_{m=1}^{m_i} Z_{nm}^i + 1)}{\sum_{k=1}^{K} \sum_{i=1}^{L} \sum_{n=1}^{n_i} R_{nk}^i (\sum_{m=1}^{m_i} Z_{nm}^i + 1)} \quad (3)$$

For $p(v_n^i|\alpha)$, we use multidimensional Gaussian distributions over a number of visual features. We also assume the independence of the features and then enforce a block diagonal structure for the covariance matrix to capture the most important dependencies. Then the estimates of new parameters are as follows:

$$\begin{aligned} m_\alpha &= \frac{\sum_{i=1}^{L} \sum_{n=1}^{n_i} R_{n\alpha}^i (\sum_{m=1}^{m_i} Z_{nm}^i + 1) v_n^i}{\sum_{i=1}^{L} \sum_{n=1}^{n_i} R_{n\alpha}^i (\sum_{m=1}^{m_i} Z_{nm}^i + 1)} \quad (4) \\ \sigma_\alpha^2 &= \frac{\sum_{i=1}^{L} \sum_{n=1}^{n_i} R_{n\alpha}^i (\sum_{m=1}^{m_i} Z_{nm}^i + 1)(v_n^i - m_\alpha)^2}{\sum_{i=1}^{L} \sum_{n=1}^{n_i} R_{n\alpha}^i (\sum_{m=1}^{m_i} Z_{nm}^i + 1)} \end{aligned}$$

For linguistic data, $p(\alpha|w_m^i)$ is given by:

$$p(w_m^i|\alpha) = \frac{\sum_{i=1}^{L} \sum_{j=1}^{m_i} \sum_{n=1}^{n_i} Z_{nm}^i R_{n\alpha}^i \delta(w_m^i, w_j^i)}{\sum_{k=1}^{K} \sum_{i=1}^{L} \sum_{j=1}^{m_i} \sum_{n=1}^{n_i} Z_{nm}^i R_{nk}^i \delta(w_m^i, w_j^i)} \quad (5)$$

where δ is the Kronecker delta function that equals to one if two arguments are the same or zero otherwise. $p(\alpha|w_m^i)$ indicates the probability of the word w_m^i generated by the component α. In the initialization, we use k-mean to cluster visual features and each cluster is assigned to one Gaussian. The mean and variance were computed based on the clustering results as well as the initial $R_{n\alpha}^i$. In addition, we calculate the co-occurrence of visual clusters and words and set the initial values of Z_{nm}^i. Then the EM-based algorithm performs the E-step and the M-step successively until convergence. In E-step, we compute the parameters of $p(v_n^i|\alpha)$ and $p(w_m^i|\alpha)$ by Equation (4) and (5). In M-step, we reestimate $R_{n\alpha}^i$ and Z_{nm}^i using Equation (2).

Experiment

Nine users participated in the experiments. They wore a head-mounted CCD camera to capture a first-person point of view. Visual data were collected at the resolution of 320 columns by 240 rows of pixels. Acoustic signals were recorded using a headset microphone at a rate of 22 kHz with 16-bit resolution. A Polhemus 3D tracker was utilized to acquire 6-DOF head positions at $40Hz$ to detect scene change. Users were seated at a desk on which there were nine office objects shown in Figure 3. They were given two tasks of introducing those objects to a newcomer (e.g. a human agent or a service robot recently purchased): (1) what are those objects and where are they located. (2) how to use those objects to accomplish the office task of writing down phone orders and maintaining purchasing records. They were asked to perform each task three times. Table 1 shows sample transcripts of these two tasks. Before the start of each trial, we randomly arranged the positions of objects and rotated them by 20^o to obtain images containing different sets of objects from different views. The system collected audio-visual pairs consisting of verbal descriptions and image sequences. The average length of the first task was 129 seconds and the second task was 186 seconds. The data collected from the first task formed Dataset 1 and Dataset 2 was made up of the data collected from the second task.

Task 1 – I see on the table a glue stick. – then in the front of me is a pink pencil sharpener. that you can use to shape the pencil. – in the back is a white stapler that is used to put pieces of paper together. .....
Task 2 – welcome to your first day of work. – this is your desk and your job is to answer the phone and take the orders. – when taking orders over the phone, you write it down on the pad of paper over there. .....

Table 1: Examples of natural descriptions

After learning, the system obtained grounded lexical items including the clusters of visual objects with their corresponding linguistic labels. To evaluate experimental results, we defined the following two measures for symbol grounding: (1) **symbol grounding accuracy** measures the percentage of the perceptual representations of objects which are correctly associated with linguistic labels. (2) **symbol spotting accuracy** measures the percentage of word-object pairs that are spotted by the computational system. This measure provides a quantitative indication about the percentage of grounded lexical items that can be successfully found. Table 2 shows the results of symbol grounding.

	Dataset 1	Dataset 2
symbol grounding	83.6%	86.2%
symbol spotting	81.9%	82.5%

Table 2: Results of symbol grounding

To evaluate the performance of categorizing visual objects, we compared our multimodal learning algorithm with several unimodal supervised and unsupervised methods that we ran on the same data set. The experiments were conducted using visual-only and audio-visual multimodal categorizations. For visual features, the clustering method based on Gaussian Mixture Model (GMM) was run which gave average accuracy of 66%. Moreover, Support Vector Machine (SVM) (Burges 1998) and Self-Organizing Map (SOM)(Vesanto & Alhoniemi 2000) were applied to visual data as supervised classification methods. Table 3 shows the comparison of all those methods. As we can see, the multimodal learning method yielded significantly better results than the performances achieved by the unsupervised learning methods. Moreover, the results showed that only using the multimodal information co-occurring in the environment is enough to give the categorization accuracy better than that of supervised SOM and within 7% of supervised SVM results. Considering the fact that our method deals with the correspondence and clustering problems simultaneously, these results are impressive. Note that the supervised methods provide benchmarks for evaluation and comparison, and we understand that more advanced classification methods could achieve better results for the same data set. However, the purpose of the experiments is to demonstrate the role of co-occurring data from multiple modalities in multimodal clustering. Specifically, we are interested in the possible improvement of performance in a fully unsupervised way by integrating more information at the sensory level. The limitation of this method is that a visual object and the corresponding spoken name are supposed to share the same latent node. Since some spoken words do not have visual correspondences, the algorithm tries to distribute them into several other nodes, which biases the estimates of parameters of those nodes. Thus, the algorithm has to fit some irrelevant data in this uniform structure.

Conclusion

This paper presents a multimodal learning system that is able to not only ground spoken names of objects in visual perception but also learn to categorize visual objects using teaching

	Dataset 1	Dataset 2
SOM (supervised)	68.9%	69.6%
SVM (supervised)	85.5%	87.2%
GMM (unsupervised)	67.3%	65.2%
multimodal	79.9%	80.9%

Table 3: Results of categorization of visual objects

signals encoded in co-occurring speech. Our main motivation is to endow an autonomous system with the learning ability by being situated in user's everyday environments and processing multisensory data in an unsupervised mode.

The principal technical contribution of this paper is to explore the possibility of utilizing the spatio-temporal and cross-modal constraints of unlabeled multimodal data in the symbol grounding and clustering problems. Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes, which could be very difficult in dealing with sensory data. We show that the unimodal clustering problem can be fundamentally simplified by incorporating cross-modal cues. Also, we show that the EM algorithm extends readily to multimodal classification and that, importantly, the parametric forms of the individual modalities can be arbitrarily different in a general framework. Although the experiments reported in this paper focus on processing audio-visual data to ground language in visual perception, both the general principle and the specific algorithm are applicable to various applications in autonomous robot, multimodal interface, wearable computing, and multimedia index.

References

Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2).

Cohen, P. R.; Oates, T.; Beal, C. R.; and Adams, N. M. 2002. Contentful mental states for robot baby. In *Eighteenth national conference on Artificial intelligence*, 126–131.

Comanicu, D., and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Analysis Machine Intelligence* 24:603–619.

Coradeschi, S., and Saffiotti, A. 2003a. An introduction to the anchoring problem. *Robotics and Autonomous Systems* 43(2-3):85–96.

Coradeschi, S., and Saffiotti, A., eds. 2003b. *the special isssue on perceptual anchoring*, volume 42.

de Sa, V. R., and Ballard, D. 1998. Category learning through multimodality sensing. *Neural Computation* 10:1097–1117.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society* 39(1):1–38.

Duygulu, P.; Barnad, K.; de Freitas, J.; and Forsyth, D. 2002. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Proceedings of European Conference on Computer Vision*. Copenhagen: Springer.

Fellbaum, C., ed. 1998. *WordNet An Electronic Lexical Database*. MIT Press.

Harnad, S. 1990. The symbol grounding problem. *physica D* 42:335–346.

Hofmann, T., and Puzicha, J. 1998. Statistical models for co-occurrence data. Technical Report AI Memo 1625, Artificial Intelligence Laboratory, MIT.

Mel, B. W. 1997. Seemore: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation* 9:777–804.

Mitton, R. 1992. A computer-readable dictionary file based on the oxford advanced learner's dictionary of current english.

Quine, W. 1960. *Word and object*. Cambridge, MA: MIT Press.

Regier, T. 1996. *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MA: MIT Press.

Roy, D., and Pentland, A. 2002. Learning words from sights and sounds: A computational model. *Cognitive Science* 26(1):113–146.

Satoh, S.; Nakamura, Y.; ; and Kanade, T. 1997. Name-it: Naming and detecting faces in video by the integration of image and natural language processing. In *Proc. IJCAI'97*, 1488–1493.

Schiele, B., and Crowley, J. L. 2000. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision* 36(1):31–50.

Siskind, J. M. 1995. Grounding language in perception. *artificial Intelligence Review* 8:371–391.

Sleator, D., and Temperley, D. 1993. Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*.

Steels, L., and Vogt, P. 1997. Grounding adaptive language game in robotic agents. In Husbands, C., and Harvey, I., eds., *Proc. of the 4th European Conference on Artificial Life*. London: MIT Press.

Swain, M. J., and Ballard, D. 1991. Color indexing. *International Journal of Computer Vision* 7:11–32.

Thrun, S., and Mitchell, T. 1995. Lifelong robot learning. *Robotics and Autonomous Systems* 15:25–46.

Vesanto, J., and Alhoniemi, E. 2000. Clustering of the self-organizing map. *IEEE Transactions in Neural Networks* 11(3):586–600.

Wachsmuth, S., and Sagerer, G. 2002. Bayesian Networks for Speech and Image Integration. In *Proc. of 18th National Conf. on Artificial Intelligence (AAAI-2002)*, 300–306.

Weng, J.; McClelland, J.; Pentland, A.; Sporns, O.; Stockman, I.; Sur, M.; and Thelen, E. 2001. Artificial intelligence: Autonomous mental development by robots and animals. *Science* 291(5504):599–600.