MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
3-33

Inter-modal Learning and Object Concept Acquisition

Katsuhiko Ishiguro † Nobuyuki Otsu †† † Yasuo Kuniyoshi †

†Graduate School of Information Science and Technology University of Tokyo Tokyo 113-0033, Japan

††National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba-shi 305-8568, Japan

{ishiguro, otsu, kuniyosh}@isi.imi.i.u-tokyo.ac.jp

Abstract

We propose an inter-modal learning system which acquires concepts about objects from auditory and visual information. The system extracts features from the input of spoken words and images, and analyzes the statistical correlation between both modalities. We use kernel-based multi-variate analysis methods and information theoretic criteria. In our experiments, the system acquired semantically meaningful concepts about shapes and colors, and also exhibited concept generalization and specialization spontaneously. Our results show that the system is capable of flexible and adaptive concept acquisition.

1 Introduction

Because of the rapid increase of multimedia data, the "multi-modal" framework has turned into a growing research field and a few works have emerged on multi-modal (audio-visual) information processing [4, 1, 5, 11, 3].

One of the main applications for audio-visual information processing is image retrieval from video contents [7, 12]. For that purpose, systems must understand and index contents. Then systems are required to *abstract* or *conceptualize* essential meanings (as symbols) from various auditory and visual signals. Preferably, this conceptualization process should be performed by the systems themselves in an unsupervised learning manner since it is impossible to define and describe all the concepts manually for real-world databases.

Clearly, this issue is similar to the "concept / language acquisition problem" faced by real-world intelligence, because human infants heavily rely on multi-modal information (e.g. visual stimuli of a toy car, auditory input of the toy's name and tactile feelings of the toy in the hands) in their concept acquisition periods. Infants have to solve the problem of how to combine multi-modal information and how to generate "concepts".

One solution is to employ the statistical correlation of multi-modal input, such as frequent co-occurrence of visual and auditory stimuli or the mutuality of multi-modal information sources.

In this paper, we propose an autonomous learning system, which acquires "concepts" based on the statistical correlation of audio-visual information. To deal with multi-modal data, we employ an inter-modal learning method using multi-variate data analysis and information theory.

2 Approach

Let us describe what we intend to make our system perform:

Imagine that you are shown the images of black objects (e.g. cars, balls, pens) and you hear the word "black" every time you see an image. Based on the frequent co-occurrence of images and spoken words, it is natural to assume that the spoken word "black" is correlated with the color of images (black). So you can hypothesize that the concept BLACK is represented with black color in visual modality and the spoken word "black" in auditory modality.

Fig. 1 is the sketch of acquired concept BLACK. And we show the overall procedure in our system in Fig. 2.

2.1 Definition and representation of concepts

The input data consists of a set of still image of an object and associated spoken word (e.g. name or color of the object), subsequently called an "incident". By using input incidents, the system then estimates the correspondence between auditory and visual features in terms of the statistical correlation, and finally outputs the acquired concepts. In this paper, we define a "concept" as what is indicated by the inter-modal correspondence.

In our system, each concept is represented with three elements: "essential spaces", "degree of confidence" and "main incidents".

"Essential spaces" are provided for each modality and represent intrinsic features of the concept. In the example of Fig. 1, a visual essential space reflects the blackness of images and an auditory essential space regards the word "black" as the typical auditory input.

"Degree of confidence" in each essential space is used to calculate the likelihood that an incident belongs to a particular concept. In the case of Fig. 1, blackness of the image and the spoken word's similarity to "black" are calculated.

"Main incidents" are the prototype incidents of the concept. We consider that main incidents represent the content of the concept. If the incident is evaluated with high confidence in both essential spaces, the system adds this incident to a list of main incidents. In Fig. 1, an incident which combines the spoken word "black" with an image of a black car is supposed to become a main incident. But an incident of

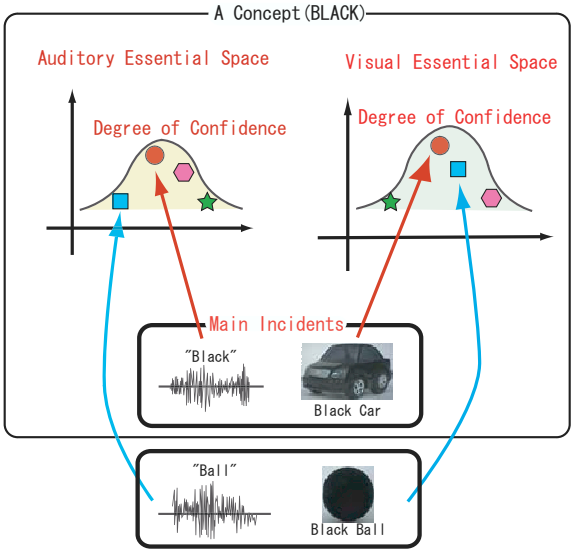

Figure 1: Sketch of a concept

an black ball image and "ball" will be rejected because of the improper auditory input.

2.2 Visual and Auditory features

As visual features, we extract Higher-order Local Auto-Correlation (HLAC) features [9] from YUV images. HLAC features have several preferable characteristics such as segmentation-free or translation-invariant. HLAC features are defined as:

$$\int_S f(\boldsymbol{r}) f(\boldsymbol{r}+\boldsymbol{r}_1) \dots f(\boldsymbol{r}+\boldsymbol{r}_M) d\boldsymbol{r} \qquad (1)$$

where $\boldsymbol{r}$ denotes a reference pixel in an image $\boldsymbol{S}$, $f(\boldsymbol{r})$ is a pixel value at $\boldsymbol{r}$, and $\boldsymbol{r}_i$ are displacements from $\boldsymbol{r}$. M is an order of HLAC. We can obtain number of HLAC features by altering the local displacement patterns $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_M\}$. Images are fed to the system with various resolutions down-sized into $\frac{1}{4} \times \frac{1}{4}$, $\frac{1}{16} \times \frac{1}{16}$. We extract HLAC features (35 dimensions) from the original and downsized images.

As auditory features, we utilize Mel Frequency Cepstrum Coefficients (MFCC). MFCC is widely used in audio signal processing as a standard auditory feature and reflecting human's perceptual measurements. We used segmented spoken words. From each frame we extract MFCC, ΔMFCC (differential coefficients of MFCC) and power. MFCC and ΔMFCC are extracted up to 12th order.

2.3 Inter-modal correlation analysis methods

The main issue addressed in this paper is the analysis of the correlation of features derived from multiple information sources. One possibility is to use Canonical Correlation Analysis (CCA). Or, we can measure "synchrony" of information sources with mutual information. Based on these premises, most of previous works are roughly classifiable in two categories: One inspired by CCA [1, 11], and one based on mutual information [5, 10, 3].

It is interesting to note that there have hardly been any attempts at combining CCA and mutual information. The fact that CCA gives an optimized projection for maximizing mutual information in appropriate conditions [2] also

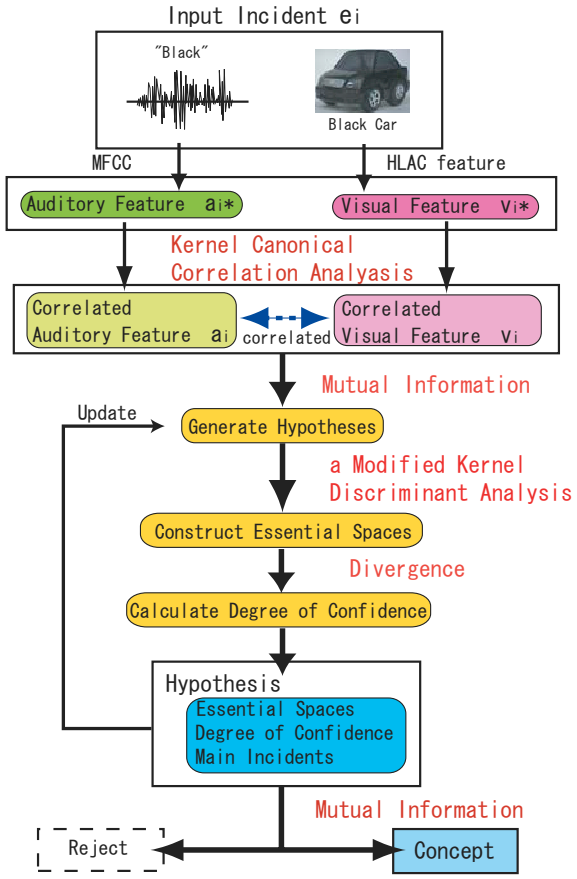

Figure 2: Overall procedure

encourages us to combine CCA and mutual information. In this paper, we present a framework that uses both mutual information and CCA, and adopt Kernel CCA (KCCA) [2] to avoid the limitation of linearity in CCA.

Let us denote the set of auditory features as $\{\boldsymbol{a}_i^*\}_{i=1}^N$ and the set of visual features as $\{\boldsymbol{v}_i^*\}_{i=1}^N$. N is the number of input incidents $\boldsymbol{e}_i = (\boldsymbol{a}_i^*, \boldsymbol{v}_i^*)$. It is noted that $\boldsymbol{a}_i^*$ is not an vector, but an array of MFCC vectors with a variable length (equals to the number of frames). We convert them into correlated auditory and visual features, subsequently denoted as $\{\boldsymbol{a}_i\}_{i=1}^N$ and $\{\boldsymbol{v}_i\}_{i=1}^N$.

We define kernel gram matrices as follows:

$$\boldsymbol{K}_A = \left\{k_a(\boldsymbol{a}_i^*, \boldsymbol{a}_j^*)\right\} \quad \boldsymbol{K}_V = \left\{k_v(\boldsymbol{v}_i^*, \boldsymbol{v}_j^*)\right\} \qquad (2)$$

where k_a and k_v are the kernel functions (in this paper, we use Gaussian kernels). Since auditory features $\boldsymbol{a}_i^*$ and $\boldsymbol{a}_j^*$ are not vectors, we measure the distance of auditory features as the difference of HMM's model likelihood. We solve the following eigenvalue problems:

$$\left(\boldsymbol{K}_A^2 + \gamma_A \boldsymbol{I}\right)^{-1} \boldsymbol{K}_A \boldsymbol{K}_V \left(\boldsymbol{K}_V^2 + \gamma_V \boldsymbol{I}\right)^{-1} \boldsymbol{K}_V \boldsymbol{K}_A \boldsymbol{W}_A = \boldsymbol{W}_A \boldsymbol{\Lambda}^2 \quad (3)$$

$$\left(\boldsymbol{K}_V^2 + \gamma_V \boldsymbol{I}\right)^{-1} \boldsymbol{K}_V \boldsymbol{K}_A \left(\boldsymbol{K}_A^2 + \gamma_A \boldsymbol{I}\right)^{-1} \boldsymbol{K}_A \boldsymbol{K}_V \boldsymbol{W}_V = \boldsymbol{W}_V \boldsymbol{\Lambda}^2 \quad (4)$$

where $\boldsymbol{W}_A$ and $\boldsymbol{W}_V$ are the coefficient matrices for kernel gram matrices and γ_A and γ_V are regularization terms.

The next step is to calculate the mutual information. Mutual information represents the amount of information about one modality conveyed from the other modality. In auditory and visual canonical space, we define spherical regions F_A

and F_V, each is centered at a correlated feature $\boldsymbol{a}_i$ and $\boldsymbol{v}_i$ and has a radius of r_a and r_v, respectively. We compute a mutual information between F_A and F_V as follows:

$$I(F_A, F_V) = \sum_{s,t}^{0,1} P(A=s, V=t) \log_2 \frac{P(A=s, V=t)}{P(A=s)P(V=t)} \tag{5}$$

Binary variables $s, t \in \{0, 1\}$ indicate whether an incident *is* contained in the region (1) or *not* (0).

The system seeks the optimum radii that give maximum mutual information. If the mutual information is greater than a threshold, our system generates a "hypothesis" of a concept and bundles contained incidents. Temporarily, we choose mutual incidents among F_A and F_V as the hypothesis's "main incidents" .

2.4 Calculating elements of a concept

Then, the system calculate the three elements, namely "essential spaces", "degree of confidence" and "main incidents".

"Essential spaces" represents intrinsic features of the concept. We assume that main incidents (prototype) of a hypothesis (concept) concentrate at one point, and the other incidents are scattered far around them in the essential spaces. Kernel Discriminant Analysis (KDA) [8, 6] is employed to obtain such spaces, with a modified discriminant criterion shown in Eq. (6). This criterion means to *maximize* the total variance while *minimizing* the target class's variance:

$$maximize \quad tr\left(\Sigma_{Target}^{-1}\Sigma_{All}\right) \tag{6}$$

"Degree of confidence" is provided for each essential space in order to measure the likelihood that an incident's membership to a particular concept. We define this measure in each essential space by Eq. (7).

$$P(\boldsymbol{e}) = \exp\left(-\frac{\|\boldsymbol{e}-\boldsymbol{m}\|^2}{2\sigma^2}\right) \tag{7}$$

where $\boldsymbol{m}$ is an average of main incidents' essential features. Let us denote a corresponding essential feature of an incident with $\boldsymbol{e}$.

In our system, we adopt the σ^2 that minimizes the Divergence, statistical distance between the confidence functions of two modalities of one hypothesis (concept). Divergence is based on Kullback-Leibler Divergence in Information theory, and is formulated as follows:

$$D(p_A, p_V) = \sum_i (p_A(\boldsymbol{e}_i) - p_V(\boldsymbol{e}_i)) \log \frac{p_A(\boldsymbol{e}_i)}{p_V(\boldsymbol{e}_i)} \tag{8}$$

The probabilistic density functions p_A and p_V are substituted with the degree of confidence of the audio essential space and of the visual essential space, respectively.

"Main incidents" are the prototype incidents of the concept. If the incident is evaluated with high confidence in both essential spaces, the system adds this incident to a list of main incidents. The system tests each essential feature with the degree of confidence to select main incidents.

The system updates every hypotheses when a new incident is input. "Update" means the re-construction of essential features, and re-evaluation of the degree of confidence, and re-membering of main incidents.

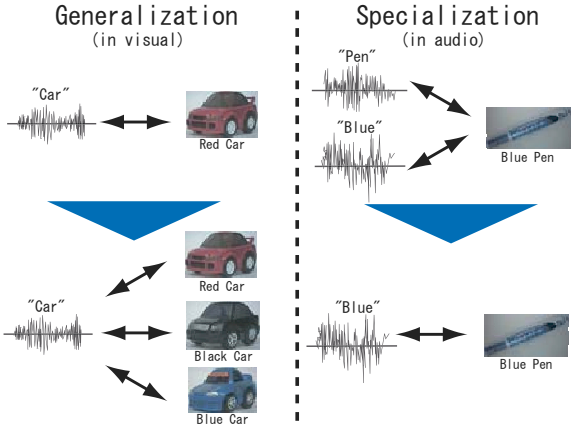

Figure 3: Examples of concept generalization and specialization

Figure 4: Examples of images

Finally, hypotheses are tested with mutual information to be acquired as concepts when the input of incidents ends.

2.5 Flexible changes of acquired concepts

Our system utilizes the statistical correlation of input multimodal data. This correlation varies according to the history of input, so we can expect the system to change the elements of concepts flexibly adapting to the input.

From the viewpoint of real-world machine intelligence, this means the system is able to *generalize* or *specialize* concepts (Fig. 3). Such adaptive and flexible treatment of concepts is actually important in infants' learning stage. We consider generalization and specialization as expansion and reduction of associations between the auditory and visual modalities in main incidents.

3 Experiment and Result

We prepared the images of models of cars, balls, and pens. Each has three color variations of red, blue and black. Thus we obtained 9 kinds of objects. Directions and scales are slightly changed through the capturing of images. Examples are shown in Fig. 4.

The spoken words are names or colors of these 9 objects. All words were spoken in Japanese. English translated words are shown in Table 1. In this setup, all images could be linked to multiple spoken words and all spoken words were associated with multiple images.

Input incidents were generated randomly from these auditory and visual data, but paired data are selected not to semantically contradict each other. We generated 100 inci-

Table 1: Used words, all recorded in Japanese

COLORS	Red	Blue	Black
NAMES	Car	Ball	Pen

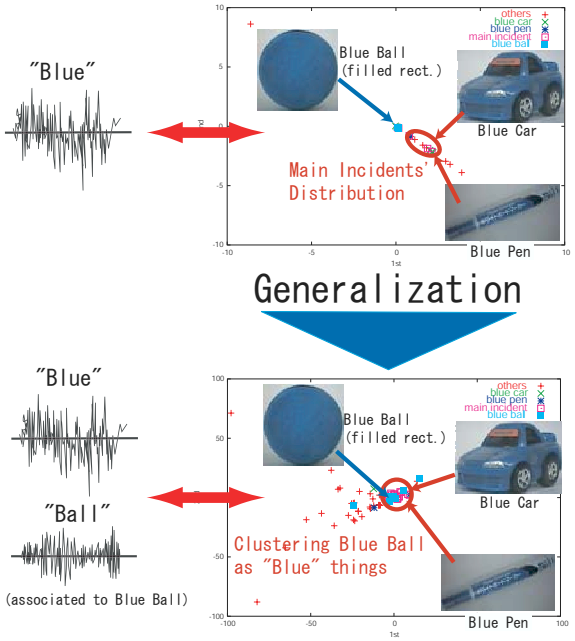

Figure 5: Example of concept generalization in visual space with the word "blue"

Table 2: Number of acquired concepts per trial

Times of Acquiring Concepts	generalization		specialization	
	auditory	visual	auditory	visual
1.8 ± 1.2	0.3	0.4	0.0	0.3

dents and input them to the system. The acquired concepts are verified by examining main incidents of each concept. We repeated this process for 10 times. We summarize the results of 10 experiments in Table 2. The threshold of mutual information for acquiring as a concept is set to 0.2.

Fig. 5 shows an example of concept generalization of color. In each figure, main incidents distribute in a circled area. The upper figure is the visual essential space at a birth of this hypothesis. The images of blue car and blue pen were associated with the word "blue", and the images of blue ball were not. The lower figure is the visual essential space when this was acquired as a concept finally. Plots of blue ball images (filled rectangles) distribute in main incidents' circled area. This means that the system generalized this concept to represent the all blue objects.

Our system not only acquired concepts, but also achieved concept generalization and specialization. It should be noted that these flexible changes in concept formation emerged spontaneously as the consequence of every calculations, and there are no intentional modules in our system.

4 Concluding Remarks

We proposed a system which acquires concepts of objects and colors by inter-modal learning using the statistical correlation of auditory and visual information.

As experimental results, the system acquired concepts about objects and colors properly, and also showed "concept generalization and specialization" which are seen in infants' cognitive developmental stages. This exhibits its capabilities of adaptive and flexible treatments of concepts.

But we still have problems in how should we utilize generalization and specialization for a rich concept structure. Future work will be aimed at considering the acquisition of meta-level concepts, related to the hierarchy or the synonymy of concepts.

We also have to consider the problem of several parameters and thresholds, manually decided in the proposed work. Some theoretic solutions for deciding those values are expected. And it is necessary to apply the system to a large real-world dataset since the size of data used in the experiments is limited.

Acknowledgments

We appreciate Max Lungarella and reviewers for their helpful comments.

References

[1] S. Akaho, S. Hayamizu, O. Hasagawa, T. Yoshimura, and H. Asoh. Concept acquisition from multiple information sources by the EM algorithm (in japanese). *The IEICE Trans. on information and systems*, Vol. J80-A, No. 9, pp. 1546–1553, 1997.

[2] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, Vol. 3, pp. 1–48, 2002.

[3] J. Fisher, III and T. Darrel. Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, Vol. 6, No. 3, pp. 406–413, 2004.

[4] O. Hasegawa, T. Kurita, S. Hayamizu, T. Tanaka, K. Yamamoto, and N. Otsu. Active agent oriented multimodal interface system. In *Proc. of IJCAI-95*, pp. 82–87, 1995.

[5] J. Hershey and J. Movellan. Audio-vision : Using audio-visual synchrony to locate sounds. In S. A. Solla, T. K. Leen, and K-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pp. 813–819. MIT Press, 2000.

[6] T. Kurita and T. Taguchi. A modification of kernel-based fisher discriminant analysis for face detection. In *Proc. of the 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 300–305, 2002.

[7] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, Vol. 25, No. 9, pp. 1075–1083, 2003.

[8] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pp. 41–48. IEEE, 1999.

[9] N. Otsu and T. Kurita. A new scheme for practical, flexible and intelligent vision systems. In *Proc. of IAPR Workshop on Computer Vision*, pp. 431–435, 1988.

[10] D. K. Roy and A. P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, Vol. 26, No. 1, pp. 113–146, 2002.

[11] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Advances in Neural Information Processing System 13*, pp. 814–820. MIT Press, 2000.

[12] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, Vol. 22, No. 12, pp. 1349–1380, 2000.