Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE: D5.3
A Framework for Cognitive Fusion

Date of deliverable:31.12.2007
Actual submission date: 06.02.2007

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **IDIAP Research Institute**

Revision [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D5.3 A FRAMEWORK FOR COGNITIVE FUSION

:

IDIAP Research Institute (IDIAP)
Oregon Health & Sciences University (OHSU)

Abstract:
The article first describes a general strategy for dealing with unexpected audio-visual events. The key aspect of this particular strategy is the comparison between of two processing streams, one being dominated by a top-down predictive process based on a prior experience, the other one based on a bottom-up information from the scene. We also discuss some issues in information fusion in such a scheme, as well as the issues of the necessary decision making resulting from the observed inconsistencies between the two processing streams.

Table of Content

1. Motivation

Surprising and unexpected sensory data confronting an organism could be important since they could represent a new danger or a new opportunity. Consistent with the classical information-theoretic concepts, information content of individual events is inversely proportional to their prior probability, i.e. the low probability unexpected events in the event stream carry the most information [5]. Miss-interpretation of unexpected inputs are therefore likely to be more serious that misclassifications of expected inputs. In addition to the low probability, the organism must have the capability to estimate the potential utility of the event in order to redirect its attention and processing resources to the surprising inputs. In the following sections we lay out a framework that can hypothetically account for the behaviour of natural cognitive system. This framework consist of three components: (1) a process of detecting that a given input deviates from the expectations based on prior knowledge, (2) multilevel fusion that enables an organism to discover coherence among different input streams and across levels of processing to make inferences about the identity of the observed events and (3) a decision theoretic framework that uses expected utility to determine the allocation of attentional resources.

2. Cognitive Processing of Unexpected Low Prior Probability Events

We note that for the purpose of understanding cognitive systems, and to avoid a variety of philosophical issues, rare or unexpected events are defined in terms prior probability of occurrence in a given context; i.e., $P\{\mathbf{L}\,|\,\mathbf{C}\}$, where L is the label and C is the context. With this interpretation of the prior probability, a variety of perceptual and cognitive experiments have demonstrated the effect of low prior probability. Although human recognition performance on unexpected stimuli is lower than that on the expected ones, the performance is quite remarkable. The results of this research lead to the development of a variety of techniques to incorporate models of the expected stimuli using simultaneous bottom up and top-down strategies (in a variety of fields from robotics to software engineering). The focus of the majority of these efforts has been on combining the top-down with bottom up processes in order to maximize the performance on stimuli and events expected in a given context. In contrast, the DIRAC project is focused on the interpretation of inputs and events that are not expected in a given context. It is important to note that the approach is centred on those situations where the low prior probability is not due to "noisy" measurements.

The proposed hypothetical framework for the cognitive processes involved in processing of sensory inputs and interpreting events or objects is illustrated in Fig. 1. A sensory input, **X,** is first transformed by the sensor-specific signal pre-processing stages and then used in two ways. First, the bottom up tract, shown as the lower path in Fig. 1, attempts to estimate the unconditional probability that a feature set **Q**, i.e., $p = P\{\mathbf{Q}\,|\,\mathbf{X}\}$. Simultaneously, the sensory input triggers a predictive process in the upper path that evokes top-down knowledge from the past experience appropriate for the current context and generates predictions for the features implied by the input and a model of the context, $p_c = P\{Q\,|\,\mathbf{X},\mathbf{M}_c\}$ where the probability of the features set **Q** given a model of the contextual effects, $\mathbf{M}_C$, e.g., A comparison between the two sets of features is then used to assess the congruence between the two representations. There are numerous metrics to characterize the discordance between

the two estimates of the feature set. In our initial approach we used an asymmetric metric, such as the Kullback-Leibler Divergence, given by

$$D(p, p_c) = \sum_{\forall Q} P\{Q \mid \mathbf{X}\} \log \frac{P\{Q \mid \mathbf{X}, \mathbf{M}_c\}}{P\{Q \mid \mathbf{X}\}}. \quad (1)$$

A reliable amount of discordance is then interpreted to indicate an unexpected object or event and combined with appropriate utilities can influence the decision regarding a subsequent action such as to escape, to fight, to gather more information, etc.).

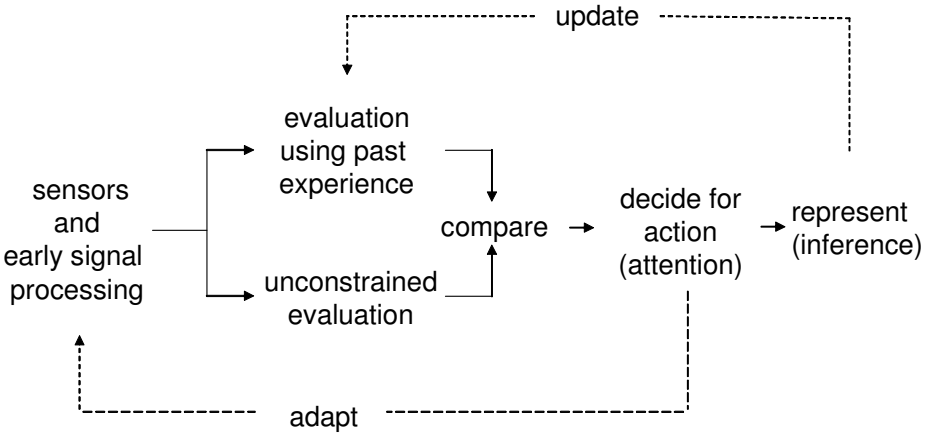

Fig. 1 Hypothesized process in discovery of unexpected items.

2.1. Unexpected Inputs in Machine Pattern Recognition

Automatic recognition by a machine is typically based on machine learning techniques that form internal representations of category boundaries or feature distributions derived from training sets. A training set consists of observable data from a sample of objects and corresponding labels that identify the objects. The internal representation is generally based on the most discriminating features thereby reducing the number of parameters needed to describe the data.. Recognition of unknown input is based on a comparison between the incoming features and the internal representation of the training set in combination with prior probabilities of expected objects. A well-trained machine, therefore, performs well on the objects similar to those in the training set but may perform poorly on previously unseen objects.

The automatic recognition of speech (ASR) is an instructive example of this classical approach to pattern recognition. A successful ASR attempts to emulate one of the most important cognitive functions used in communication among human beings, the recognition of a message in speech signal. The linguistic message in speech data is coded in a sequence of speech sounds (phonemes), substrings of phonemes represent words, sequences of words form phrases. A typical ASR attempts to infer the linguistic message in the phrase while ignoring a variety of other aspects of the utterance such as accent, emotional content, etc. Despite the ASR development efforts over the last fifty years supported by the industry as well as public funds, using hundreds hours of speech, estimating thousands of parameters, a reliable machine that would emulate human speech recognition abilities under realistic communication conditions is still an elusive dream. Significant improvements have been achieved using sophisticated techniques for utilizing the prior knowledge in the form of text-derived language model and in pronunciation lexicons. However, the heavy reliance on theses models has a very undesirable effect in that unexpected lexical items (words) in the phrase are typically replaced by acoustically acceptable in-vocabulary items. This is the major source of error in ASR [1,3,4]. Since the unexpected words are likely to be more informative [5], these errors are represent significant decrement in performance. Improving the machine

ability to handle these unexpected words would considerably increase the utility of ASR technology in practical applications. A successful approach to the detection of unexpected words is likely to generalize to many other machine recognition domains.

A technique for identification of unexpected words in automatic recognition of speech that builds on the above described principle has been recently developed and applied to identification of words that are not in the vocabulary of the machine [7] and is a topic of DIRAC deliverable D.2.5

2.2. Some Issues

A number of issues come up when pursuing this strategy. For instance, the problem of the unconstrained evaluation of the sensory event in machine recognition is certainly not a trivial one and will be addressed in our subsequent research work. Similarly, the problem of update of the prior knowledge so that the detected novel object or event can be added to the stored database and represented in terms of its features for future use in the top-down path. We must also note that unexpected stimuli may arise due to noisy and uncertain sensory inputs. Although it is clearly important to identify and distinguish situations with noise and uncertainty, the focus most of the work within DIRAC is in situations where the sensory signals have low uncertainty.

Two particular issues, i.e. the issue of enhanced feature representation using the multi-level information fusion, and the issue of deciding for the action that relates to attention, are discussed in some more detail below.

3. Multilevel Information Fusion in Discovery of Unexpected Items

Information fusion happens on all levels of cognitive processing. The hierarchical, multi-resolution fusion provides way to integrate various sources at the most appropriate level. The ability to integrate information from multiple sources, i.e., information fusion, is, therefore, an important concept in detecting and identifying rare events. A pattern recognizer must be able to combine appropriately multiple features, multiple contextual variables, prior probabilities, etc. The important and novel aspect of our approach is that the combinations occur simultaneously at multiple levels, the information from multiple sources if integrated at the lowest possible level, and the combination is determined by an adaptive process driven, in part from the action (attention) control module as shown in Figure 1.

It appears that as the firing rates decrease with the increasing hierarchy of the processing level, so does increase the timing tolerance for the requirement of "synchrony" among the fused information streams. Thus, on the peripheral levels such as a fusion of two acoustic signal streams, each from one ear, the two signals with identical frequencies are combined at the signal level after adjusting the relative phase to maximize the resulting signal to noise ratio. If the signals have slight difference in frequency, i.e. the timing difference may be of the order of ms, the percept is a binaural beat.

However, if the signals are found to be uncorrelated, e.g. have non-overlapping frequency content, the combination is not performed on the peripheral level but rather is done on a higher level, where the firing rates are slower, and subsequently also the tolerance for the synchrony is larger. Then, the signals can still be considered "synchronous" even when the timing differences exceed the ms level.

Eventually, on the cortical level, the asynchronies up to hundred of ms can be tolerated in the information stream fusion within particular areas of cortex, as evidenced e.g. by human ability of comprehension of speech in highly reverberant environments.

So, when it comes to fusion of information from different modalities, possibly involving highest levels of cognitive processing, large asynchronies in audio-visual stimuli resulting e.g. from different speeds of propagation of acoustic and optical waves, can be tolerated for the fusion.

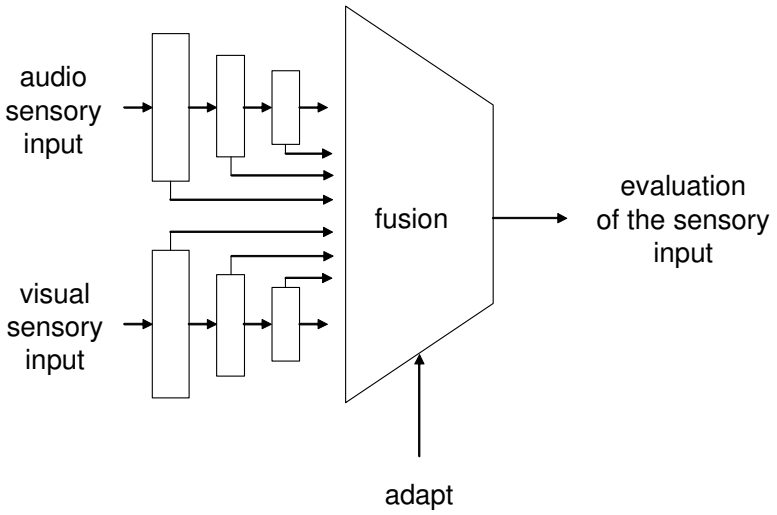

Fig. 2: One possible configuration of the sensory and early signal processing module

A simple example of the process involves a fusion of two acoustic signal streams, one from each of two spatially separated sensors such as one's two ears. If the two signals have the same frequencies, they are combined at the signal level after adjusting the relative phase to maximize the resulting signal to noise ratio. If the signals have slight difference in frequency, the percept is a binaural beat. If the two signals are uncorrelated, for example, have non-overlapping frequency content, the combination is performed at a higher level, such as on the level of the envelope of the signal. When the signal streams come from different modalities, the fusion may happen on even higher event level.

The underlying principles that govern the combination need to be investigated in considerable more details, but the current working hypothesis consists of the following observations:

1. Fusion of any two streams appears to occur at the lowest level at which the system can assess significant coherence (correlation) between the two streams. The coherence may occur following simple invertible transformations such as affine transformation (shift, scaling,).
2. The fundamental dimensions of features are determined by evaluating the coherence within a given context or environment

4. Attention Control in Processing of Unexpected Events

The research community has recognized the importance of the ability of the human perceptual and cognitive systems to detect outliers – objects whose features differ from others. Numerous experiments have carried out to examine which features are most effective in performing this discrimination in both spatial and sequential search tasks. For spatial tasks, an outlier stimulus has been defined by outstanding values in brightness, colour, etc, relative to the surrounding feature values. Examples include searching for a red object among green, an

oblique edge among vertical ones, etc. In the temporal form of the outlier detection task, a set of stimuli would be presented in a rapid serial visual presentation paradigm (RSVP), and the subjects' task would be to detect the "oddball." The definition of the "novel" stimulus, often implicit, has been in terms of relatively low frequency of occurrence of the feature values within the scope of the experiment.

Although the understanding of the perceptual processes underlying the detection of these perceptual outliers is very important, it does not capture the ability of the human cognitive system to recognize novel stimuli defined by a conjunction of features that have special "meaning" or value. For example, a new word in a given language would comprise the same phonemes as all the other words, but they would be combined in a novel way. The gait of a person carrying a concealed, heavy load may be similar to the walk of an obese person, but would be unusual for a fit individual. Taking a medication by an elder is certainly a typical behaviour for a particular elder, unless he just took the same medication several minutes earlier. An intelligence analyst viewing an airport scene may not pay much attention to a heavy truck parked next to a Boeing 757, unless that airplane is actually a regularly scheduled passenger flight.

The problem with detecting these unexpected events and objects is that novel conjunctions of usual features are quite frequent, although any individual conjunction may be unique. The need to focus one's attention on these unexpected stimuli must therefore be defined by the context and the "value" of such stimuli within a given situation or a task. Thus, even if we see a dog that appears to meow, we may ignore it unless we are in a psychological diagnostic test; in real life, we may interpret the sound as being generated by a hidden cat.

The process in Figure 1 associated with the decision to pay attention, must therefore include the notion of utility of an event. Even for humans, the ability to detect and interpret some of these unexpected stimuli requires some training and a prior specification of utility. Even if we hear a word that we never heard before, we may ignore it, unless it is in a context that imposes a high utility on that utterance for accomplishing a task, e.g., a policewoman is giving instructions in an emergency situation.

The detection and interpretation of the unexpected stimuli therefore requires the comparison of the feature values with those expected from the context by the predictive system, but the decision to pay attention to these requires the assignment of value (utility) to the discordance between the perceptual and expected values.

5. References

[1] Bazzi, I. (2002) "Modelling Out-of-Vocabulary Words for Robust Speech Recognition", MIT PhD. Thesis, Department of Electrical Engineering and Computer Science, 2002

[2] Bourlard, H. and Morgan, N., "Connectionist Speech Recognition - A Hybrid Approach", Kluwer Academic Publishers, 1994

[3] Chase, L. , "Error-Responsive Feed Back Mechanisms for Speech Recognizers", PhD Thesis, April 11, 1997.

[4] Hazen, T., and Bazzi, I., "A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring", ICASSP'01, Salte Lake City, Utah.

[5] Hermansky, H. and N. Morgan, "Automatic Speech Recognition", in Encyclopedia of Cognitive Science, L. Nadel, Ed., Nature Publishing Group, Macmilian Publishers, 2002

[6] MacKay, D.J.C. “Information Theory, Inference, and Learning Algorithms”, Cambridge University Press, 2004

[7] Ketabdar, H. and H. Hermansky: “Identifying and dealing with unexpected words using in-context and out-of-context posterior phoneme probabilities”, IDIAP Research Report 06-68, 2006