Submitted to AAAI Workshop on Representations for Multi-modal Human-Computer Interaction – short paper

Multimodal Tracking and Classification of Audio-Visual Features

Vladimir Pavlović and Thomas S. Huang *
University of Illinois at Urbana-Champaign
{vladimir,huang}@ifp.uiuc.edu

Abstract

The surge of interest in multimodal interfaces has prompted the need for novel estimation and classification techniques for data from different but coupled modalities. Unimodal techniques ported to multimodal domain have only exhibited limited success. We propose a new framework for feature tracking and classification based on multimodal knowledge-constrained hidden Markov models (HMMs). Typical role of HMMs as statistical classifiers is enhanced by their new role as multimodal feature trackers (predictors). Moreover, by fusing the multimodal formulation with higher level knowledge we allow the influence of such knowledge to be reflected in feature tracking and classification.

Introduction

The surge of interest in multimodal interfaces has prompted the need for more sophisticated techniques for estimation and classification of data represented in different but *coupled* modalities. Numerous approaches employing loosely coupled unimodal techniques have been ported into the multimodal domain with limited success. For instance, various multimodal interfaces such as (Fukumoto, Suenaga, & Mase 1994; Cohen *et al.* 1997) rely on high level joint interpretation of different modalities. This approach, unfortunately, discards some low-level dependencies that may exist among different modes. Another drawback of classical tracking/classification approaches stems from the lack of coupling between feature tracking and feature classification.

In this work, we propose a novel framework for multimodal object tracking and classification based on multimodal knowledge-constrained hidden Markov models. Hidden Markov models are a commonly used statistical tool in the field of speech recognition (Rabiner & Juang 1993). They have been recently brought into domains of gesture recognition (Schlenzig, Hunter, & Jain 1994), bimodal lip reading (Adjoudani & Benoit 1995), and bimodal gesture/speech recognition (Pavlović, Berry, & Huang 1997) and source separation (Brand 1997). In this framework, we extend the role of multimodal HMMs from statistical classifiers to feature trackers and classifiers. Moreover, by fusing the multimodal formulation with higher level knowledge we allow the influence of such knowledge to be reflected both in feature tracking and classification.

*This work was supported by National Science Foundation Grant IRI-9634618.

Multimodal Hidden Markov Models

Hidden Markov model (HMM) is a doubly stochastic process, a probabilistic network with *hidden* and *observable states*. Each HMM can be defined as a triplet $(\mathbf{A}, \mathbf{b}, \pi)$, where $\mathbf{A}$ represents the (hidden) state transition matrix, $\mathbf{b}$ describes the probabilities of the observation states, and π is the initial hidden state distribution. Three types of tasks are usually associated with a system modeled as a HMM: observation classification, hidden state inference, and learning of model parameters. Efficient algorithms exist for all three tasks: Viterbi decoding, forward/backward probability propagation, and EM iterative learning (Rabiner & Juang 1993).

Multimodal hidden Markov models (MHMMs) can be defined as a simple extension of the classical unimodal HMMs, similar to (Brand 1997). Instead of having a single set of hidden and observable states describing one type of process, MHMMs have M such mutually coupled sets (M modes). Formally, a MHMM is a triplet $(\mathbf{A}, \mathbf{b}, \pi)$ where $\mathbf{A} = [a_{\underline{k},\underline{l}}]_{\underline{k},\underline{l}\in\mathcal{X}}$, $a_{\underline{k},\underline{l}} = P(\underline{x}_{t+1} = \underline{l} \mid \underline{x}_t = \underline{k})$, $\mathbf{b} = [b_{\underline{k}}]_{\underline{k}\in\mathcal{X}}$, $b_{\underline{k}} = P(\underline{y}_t = \underline{Y} \mid \underline{x}_t = \underline{k})$, and $\pi = [\pi_{\underline{k}}]_{\underline{k}\in\mathcal{X}}$, $\pi_{\underline{k}} = P(\underline{x}_0 = \underline{k})$. Here, $\underline{k} = [k_1 k_2 \dots k_M]'$, $k_i = 1, \dots, N_i$, denotes a vector of indices in the space $\mathcal{X}$ of all M-dimensional indices. Analogous to HMMs, $\mathbf{A}$ now describes the joint probability distribution of M multimodal states conditioned on their M *multimodal* predecessors.

Given the above definition of a MHMM, the problems of inference and learning may seem difficult to tackle. However, every MHMM can be readily transformed into an equivalent HMM using the state grouping technique often employed in the domain of Bayesian networks. An M-modal state in $(N_1, N_2, \dots, N_M)$ dimensional X space can be represented as a unimodal state in a one dimensional set of $N_1 \times N_2 \times \cdots N_M$ different states. Classification, inference and learning techniques of unimodal HMMs can then be readily applied to MHMMs.

Tracking

HMMs are often employed as classifiers of temporal sequences of features in conjunction with some classical feature trackers such as Kalman filters. This approach, however, decouples feature prediction from feature classification. A more closely coupled prediction and classification may be beneficial to each other. HMMs represent a useful framework for such unification.

Consider a unimodal (or for that matter a multimodal) HMM as defined in the previous section. Given a set of observations $\underline{y}_t = [y_1 \cdots y_t]'$, it can be shown

that the *expected value* of an observation at time $t+1$, $E[y_{t+1} \mid \underline{y}_t]$, can obtained as

$$\hat{y}_{t+1} \frac{1}{P(\underline{y}_t)} \sum_{x_{t+1}} E[y_{t+1} \mid x_{t+1}] \; \alpha_t^0(x_t),$$

where we use $\alpha_t^0(x_{t+1}) = \sum_{x_t} \alpha_t(x_t) P(x_{t+1} \mid x_t)$ and $\alpha_t(x_t) = P(x_t, \underline{y}_t)$ denotes the forward probability, a product of the efficient forward probability propagation procedure (Rabiner & Juang 1993). Similar expression can be derived for the variance of y_{t+1}.

The above estimates of y_{t+1} and its variance eliminate the need for an additional Kalman-type predictor. Moreover, this prediction approach can be utilized in the framework of multimodal HMMs, thus effectively producing a *multimodal* estimate of the future observations in each of the coupled modes. This can greatly increase robustness of the prediction process. In addition, a higher level knowledge, such as grammars defined over sets of MHMMs, can be brought into play using this prediction approach.

Higher-Level Knowledge Constraints

Complex natural processes such as speech and object motion can rarely be accurately and efficiently described using a single model. It is more plausible to view such processes as being produced by a set of models governed by some higher level knowledge.

Consider a set of HMMs $\mathcal{H} = \{H_1, \ldots, H_W\}$ and a probabilistic grammar describing temporal dependencies of the individual HMMs H_i in the set. One way to model this grammar would be to view it as a Markov model $(\mathbf{A}_G, \pi_G)$ defined over the space $\mathcal{H}$, where $\mathbf{A}_G = [a_{G\,ij}]_{W \times W}$, $a_{G\,ij} = P(H_j \mid H_i)$, and $P(H_j \mid H_i)$ denotes the probability of model H_i followed by H_j. π_G denotes initial model probabilities.

An easy way to integrate this grammar into the HMM framework arises when one observes that set $\mathcal{H}$ with grammar $(\mathbf{A}_G, \pi_G)$ can be viewed as one complex HMM. Therefore, all classification and prediction tools of general HMMs can be readily applied to knowledge-constrained HMMs. This, in turn, introduces higher level knowledge constraints to prediction and classification. Furthermore, straightforward extensions of this approach can be applied to multimodal HMMs yielding knowledge-constrained multimodal classification and tracking.

Of course, complex HMMs or MHMMs designed in this fashion are defined over very high dimensional state spaces. However, by constraining the individual model topologies to sparse structures (such as left-to-right HMMs) and employing sparse grammars, the complexity of complex HMMs and MHMMs becomes tractable.

Experimental Results

Our preliminary experiments were aimed at testing the feasibility of the proposed framework. As the testbed application we chose a joint audio-visual interpretation of speech and unencumbered hand gestures acquired through a video camera for interaction with an immersive virtual environment (Pavlović, Berry, & Huang 1997). In the original setup, gestures and speech were independently recognized using unimodal HMMs and then jointly interpreted on the word level. Gesture tracking and parameter prediction from the video stream originally employed a second-order Kalman predictor.

Using the obtained unimodal models from the original setup, the known intra- and inter-modal grammars, we have constructed a joint MHMM of the audio/video process. This model was then used to perform multimodal gesture feature tracking and multimodal gesture/speech classification. An example of gesture parameter prediction on a sequence of test data is depicted in Figure 1. Gesture and speech recognition was also tested on a short sequence of data. The results were again encouraging and are depicted in Figure 2.

Conclusions and Future Work

Recent gain in popularity of multimodal interfaces has prompted the need for new techniques for estimation and classification of multimodal data. Classical approaches employing loosely coupled unimodal techniques have shown limited success, possibly due to the loss of low-level dependencies among modes. Moreover, the lack of tight coupling between feature tracking and classification in the classical approaches further degrades their performance. In this work, we propose a novel probabilistic network framework for multimodal prediction and classification enhanced by high-level knowledge. Preliminary results indicate the feasibility of this approach. Our current experiments are aimed at testing the robustness of tracking and classification for multimodal data corrupted by different noise levels as well as the influence of the modalities' coupling on the system performance. Future plans include extensions of this approach to mixed-state (discrete/continuous) HMM trackers/classifiers.

References

Adjoudani, A., and Benoit, C. 1995. Audio-visual speech recognition compared across two architectures. In *Proc. of the Eurospeech'95 Conference*, volume 2, 1563–1566.

Brand, M. 1997. Source separation with coupled hidden Markov models. Technical Report TR 427, Vision and Modeling Group, MIT Media Lab.

Cohen, P. R.; Johnston, M.; McGee, D.; Oviatt, S.; and Pittman, J. 1997. QuickSet: Multimodal interaction for simulation set-up and control. In *Proc. of the 5th Applied Natural Language Processing Meeting*. Washington, DC: Association of Computational Linguistics.

Fukumoto, M.; Suenaga, Y.; and Mase, K. 1994. "Finger-Pointer": Pointing interface by image processing. *Computers and Graphics* 18(5):633–642.

Pavlović, V. I.; Berry, G. A.; and Huang, T. S. 1997. Fusion of audio and visual information for use in human-computer interaction. In *Proc. Workshop on Perceptual User Interfaces*.

Rabiner, L. R., and Juang, B. 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey, USA: Prentice Hall.

Schlenzig, J.; Hunter, E.; and Jain, R. 1994. Recursive identification of gesture inputs using hidden Markov models. In *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, 187–194.

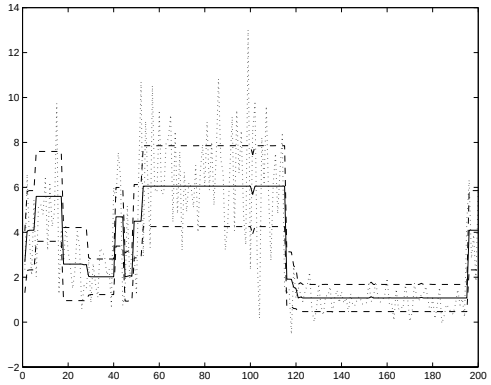

Figure 1: One step prediction of hand velocity using multimodal knowledge-constrained HMM. Dashed lines indicate standard deviation bounds on the predicted value.

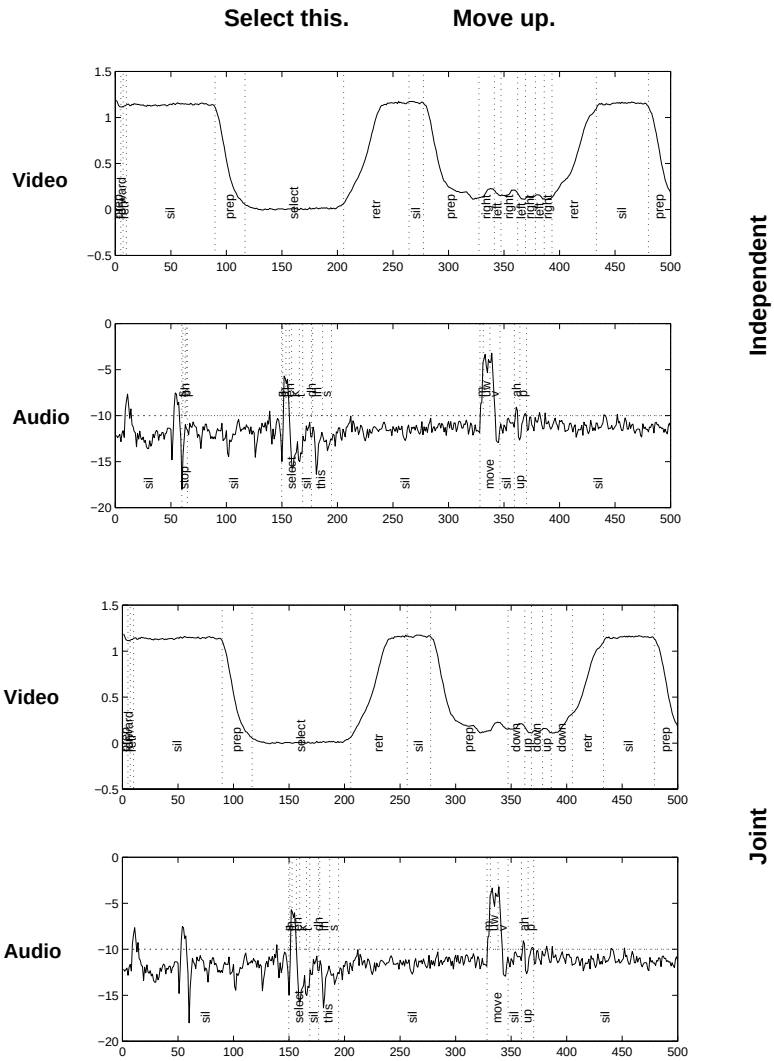

Figure 2: Recognition of spoken words and gestural actions. The figure shows results of temporal segmentation of hand gestures and speech using independent (top two graphs) and joint (bottom two graphs) inference. Depicted features for video and audio streams are the hand angle and a cepstral coefficient, respectively. Top line depicts correct sequence transcription. Note that joint interpretation eliminates a spurious "stop" in speech and correctly classifies "move up" in gestures. (Initial miss-labeling in the video stream is due to click noise.)