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

Introduction

The surge of recent interest in multimodal information processing and multimodal interfaces has prompted the need for more sophisticated techniques for estimation and classification of data represented in different but *coupled* modalities. Namely, it has become necessary to devise techniques that take full advantage of more or less "correlated" information present in multiple modalities to enhance the estimation and classification performance within individual modalities. For instance, difficult recognition of visually perceived hand motions (gestures) could potentially benefit from (somehow) incorporating the accompanying speech data into the recognition process. Similarly, the accompanying speech could help in predicting the values of hand motion parameters necessary for efficient hand tracking. So far, numerous approaches employing loosely coupled unimodal techniques have been ported directly into the multimodal domain to alleviate the recognition problem. Various multimodal interfaces such as (Fukumoto, Suenaga, & Mase 1994; Cohen *et al.* 1997) rely on high level joint interpretation of different modalities. Initial feature estimation, prediction, and lower-level classification is performed independently within each of the modality domains. This approach, unfortunately, discards the inherent dependencies that may exist among different modes and ceases to exploit the benefits of multimodal coupling. Another drawback of classical tracking/classification approaches stems from the commonly found uncoupling of feature tracking and prediction from feature classification. However, classification very often involves higher level knowledge constraints, the presence of which can undoubtedly benefit the tracking/prediction process.

In this work, we propose a novel framework for multimodal object estimation/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 recently been brought into domains of gesture recognition (Schlenzig, Hunter, & Jain 1994), bimodal lip reading and speech interpretation (Adjoudani & Benoit 1995), and bimodal gesture/speech recognition and source separation (Brand 1997; Brand & Oliver 1997; Pavlović, Berry, & Huang 1997). In this framework, we extend the classical role of multimodal HMMs from statistical classifiers to feature predictors. Moreover, by fusing the multimodal formulation with higher level knowledge (grammars) we allow the influence of such knowledge to be reflected in feature prediction and tracking as well as in feature classification.

Multimodal Hidden Markov Models

A 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 dis-

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

tribution. In other words,

$$\begin{aligned} \mathbf{A} &= [a_{ij}]_{N\times N},\ a_{ij} = P(x_{t+1} = j \mid x_t = i), \\ \mathbf{b} &= [b_i]_{N\times 1},\ b_i = P(y_t = Y \mid x_t = i), \\ \pi &= [\pi_i]_{N\times 1},\ \pi_i = P(x_0 = i). \end{aligned} \quad (1)$$

where x_t denotes a hidden state from the set $\mathcal{X}$ of N possible discrete hidden states, and y_t denotes an observation from a set of observations $\mathcal{Y}$. $\mathbf{A}$ describes the time invariant distribution of hidden states at time $t+1$ conditioned on the first order Markov predecessors at time t. The observations are assumed dependent only on the current hidden states. 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 based on forward/backward probability propagation, Viterbi decoding, and Baum-Welsh reestimation (EM algorithm) exist for all three tasks (Rabiner & Juang 1993).

Multimodal hidden Markov models (MHMMs) can be defined as an obvious 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 processes, MHMMs have M such mutually coupled sets or M modes. Formally, a MHMM is a triplet $(\mathbf{A}, \mathbf{b}, \pi)$ where

$$\begin{aligned} \mathbf{A} &= [a_{\underline{k},\underline{l}}]_{\underline{k},\underline{l}\in\mathcal{X}},\ a_{\underline{k},\underline{l}} = P(\underline{x}_{t+1} = \underline{l} \mid \underline{x}_t = \underline{k}), \\ \mathbf{b} &= [b_{\underline{k}}]_{\underline{k}\in\mathcal{X}},\ b_{\underline{k}} = P(\underline{y}_t = \underline{Y} \mid \underline{x}_t = \underline{k}), \\ \pi &= [\pi_{\underline{k}}]_{\underline{k}\in\mathcal{X}},\ \pi_{\underline{k}} = P(\underline{x}_0 = \underline{k}). \end{aligned} \quad (2)$$

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. This dependence structure is depicted in Figure 1. Such dependence structure allows for different internal dynamics of each modality to exist (horizontal dependencies among hidden states in Figure 1) while still introducing inter-modal correlation (diagonal dependencies in Figure 1.)

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! This can be achieved using the state grouping technique often employed in the domain of Bayesian networks (Frey 1998). 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. Well known classification, inference and learning techniques of unimodal HMMs can then be readily applied to MHMMs.

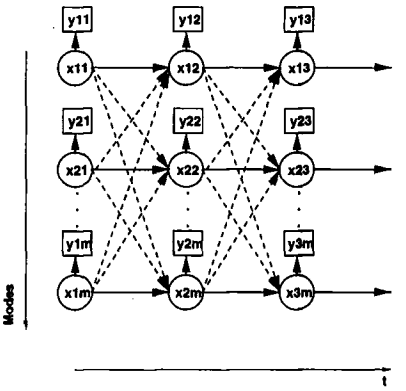

Figure 1: Independence graph for multimodal hidden Markov models. Solid and dashed arrows depict intra-modal and inter-modal dependencies, respectively. Each row contains states associated with one mode of the multimodal process.

Prediction

HMMs are often employed as classifiers of temporal sequences of features in conjunction with some classical feature predictors/trackers such as Kalman filters. Using this approach, however, decouples feature prediction from feature classification: features are estimated and predicted independently of how they are later classified. This can often result in degradation of the system performance. A more closely coupled prediction and classification may be beneficial to each other. For instance, knowing which class a hand motion belongs to can bear influence on which motion model parameters are used for the hand tracking. 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 $t+1$ can obtained as

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

where we use $\alpha_t^0(x_{t+1}) = \sum_{x_t} \alpha_t(x_t) P(x_{t+1} \mid x_t)$ and $\alpha_t(x_t) = P(x_t \mid \underline{y}_t)$ 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},

$$E[y_{t+1} y'_{t+1} \mid \underline{y}_t] = \frac{1}{P(\underline{y}_t)} \sum_{x_{t+1}} E[y_{t+1} y'_{t+1} \mid x_{t+1}]\ \alpha_t^0(x_{t+1}).$$

This can, of course, be generalized for an arbitrary $K \geq 1$ step prediction as well as the filtering, $K = 0$.

The above estimates of y_{t+1} and their variance obtained from the HMM hence 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. For instance, a video object feature (velocity of the hand in a sequence of images, for example) can be predicted based on previous values of that video feature as well as the accompanying audio features. 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. We discuss this notion in the following section.

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. An example is often found in speech recognition: phonemes as the smallest speech units are modeled using HMMs. Words are modeled as specific sequences of phonemes, and sentences are modeled using grammatical rules defined on words. Similar approaches can be employed to describe object motion, for instance, by defining a set of rules over the set of basic motion models. Classification of unknown motion or speech can then be tackled in this framework.

Consider a set of HMMs $\mathcal{H} = \{H_1, \ldots, H_W\}$ and a probabilistic grammar describing the temporal dependencies of the individual HMMs H_i in the set. One way to model such a grammar would be to view it as a Markov model $(\mathbf{A}_G, \pi_G)$ defined over the space $\mathcal{H}$, where

$$\mathbf{A}_G = [a_{Gij}]_{W \times W}, \; a_{Gij} = P(H_j \mid H_i), \qquad (3)$$

and $P(H_j \mid H_i)$ denotes the probability of model H_i followed by H_j. π_G denote initial model probabilities.

An easy way to integrate this grammar in the HMM framework arises when one observes that the set $\mathcal{H}$ with grammar $(\mathbf{A}_G, \pi_G)$ can be viewed as one complex HMM. This complex HMM is defined over the set of $N = N_1 + N_2 + \ldots + N_W$ hidden states formed from the hidden states of all individual HMMs. The probability transition matrix of this set, $\mathbf{A}_{complex}$ can be easily obtained from individual model's transition matrices, entry and exit state distributions, and $\mathbf{A}_G$. The observation distributions are simply carried over from the individual HMMs.

By viewing the probabilistic grammar-constrained set of HMMs as a complex HMM itself enables us to directly apply the tools of classification, inference, and prediction of simple HMMs to this case. This, in turn, introduces higher level knowledge constraints to all those tools with all of it's benefits and possible drawbacks. Furthermore, straightforward extensions of this approach can be applied to multimodal HMMs yielding knowledge-constrained multimodal classification, inference, 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 the often used left-to-right HMMs), the complexity of complex HMMs and MHMMs becomes quite tractable.

Experimental Results

Our 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 for interaction with immersive virtual environments described in (Pavlović, Berry, & Huang 1997). The setup allows a user to interact with a virtual 3D environment using hand gestures, such as pointing and simple symbolic motions, and spoken commands. For example, the user would point with her/his hand at an object and say "select." Once the object is selected, the user could make it rotate counter-clockwise by saying "rotate left" and performing a hand gesture symbolizing rotation to the left. In total, a set of twelve gestural commands and fourteen spoken words was used to interact with the environment.

In the original setup, gestures and speech were initially independently recognized using unimodal HMMs and then jointly interpreted on the word level. Unencumbered hand tracking is accomplished using a set of one or two video cameras. Analysis and prediction of hand motion from the video stream was originally obtained using a second-order Kalman predictor, as described in (Pavlović, Berry, & Huang 1997).

The multimodal HMMs were constructed from unimodal models of the original setup, the known intra- and inter-modal grammars, we have constructed a joint MHMM of the modeled audio/video process. Intra-model state transition probabilities were approximated by relative transition frequencies between the unimodally segmented states. Our training set consisted of a sequence of 40 multimodal commands. The test set was a different sequence of 40 commands performed by the same user. This model was then used to perform multimodal gesture feature prediction and multimodal gesture/speech classification.

An example of multimodal gesture and speech pa-

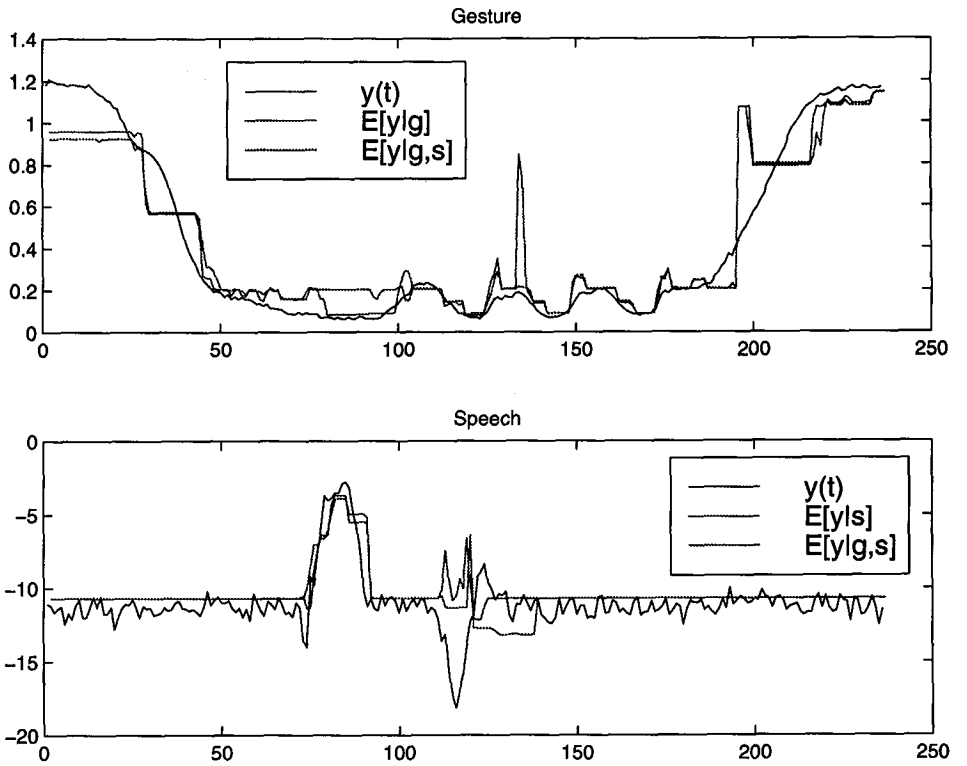

Figure 2: One step prediction of hand angle and a cepstral coefficient using multimodal knowledge-constrained HMM. Blue lines, red lines, and green lines depict measured features, unimodal estimates, and bimodal estimates, respectively.

rameter prediction on a sequence of test data is depicted in Figure 2. As the example indicates the multimodally predicted gesture feature is closer to the real (measured) data then the one predicted unimodally. However, one should note the effect of hidden space discretization in HMMs. Namely, the predicted values are "quantized" about the levels of observation means associated with the hidden states. If the number of hidden states is sufficiently high, the discretization will not significantly effect the prediction. On the other hand, higher number of hidden states results in increased computational complexity and demand on larger training data sets. To circumvent this problem, we have formulated a mixed-state HMM which encompasses both discrete and continuous hidden states (Pavlovic & Huang 1998).

Gesture and speech recognition was also tested on a short sequence of data. The results were again encouraging: using gestures alone (unimodal recognition) the recognition rate was close to 80%. Once the multimodal model was employed, the recognition rate improved to 94%. An example of unimodal and multimodal classification on a sequence of data is depicted in Figure 3. Unfortunately, these encouraging recognition results are counterweighted by the complexity of inference in the high dimensional multimodal state space. Sparsity of the system significantly effects this complexity. Yet, training of the model parameters is largely not affected by the sparsity. To address this issues we are currently devising approximate learning techniques based on variational inference(Jordan *et al.* 1998).

Conclusions

The recent gain in popularity of multimedia and multimodal interfaces has prompted the need for more sophisticated techniques for estimation and classification of multimodal data. Classical approaches employing loosely coupled unimodal techniques applied to the multimodal domain have shown limited success, possibly due to the loss of inherent dependencies that may exist among different modes at lower levels of integration. Moreover, the lack of tight coupling between feature prediction and feature classification in the classical approaches may further reduces the performance of such techniques in the multimodal domain. In this work, we propose a novel probabilistic network framework for multimodal object track-

ing/classification which fuses the feature tracking and classification enhanced by constraints of a higher level knowledge. Results of our test indicate the feasibility of this approach. Despite these encouraging results two problems still remain: computational complexity induced by a high-dimensional state space, and "discretization" of the estimation and prediction spaces. These problems will be addressed in the future using approximate inference and learning techniques and mixed-state HMMs.

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., and Oliver, N. 1997. Coupled hidden Markov models for complex action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 994–999.

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.

Frey, B. 1998. *Graphical Models for Machine Learning and Digital Communication.* MIT Press.

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

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1998. An introduction to variational methods for graphical models. In Jordan, M. I., ed., *Learning in graphical models.* Kluwer Academic Publishers.

Pavlovic, V., and Huang, T. S. 1998. Mixed state hidden markov models. in preparation.

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.

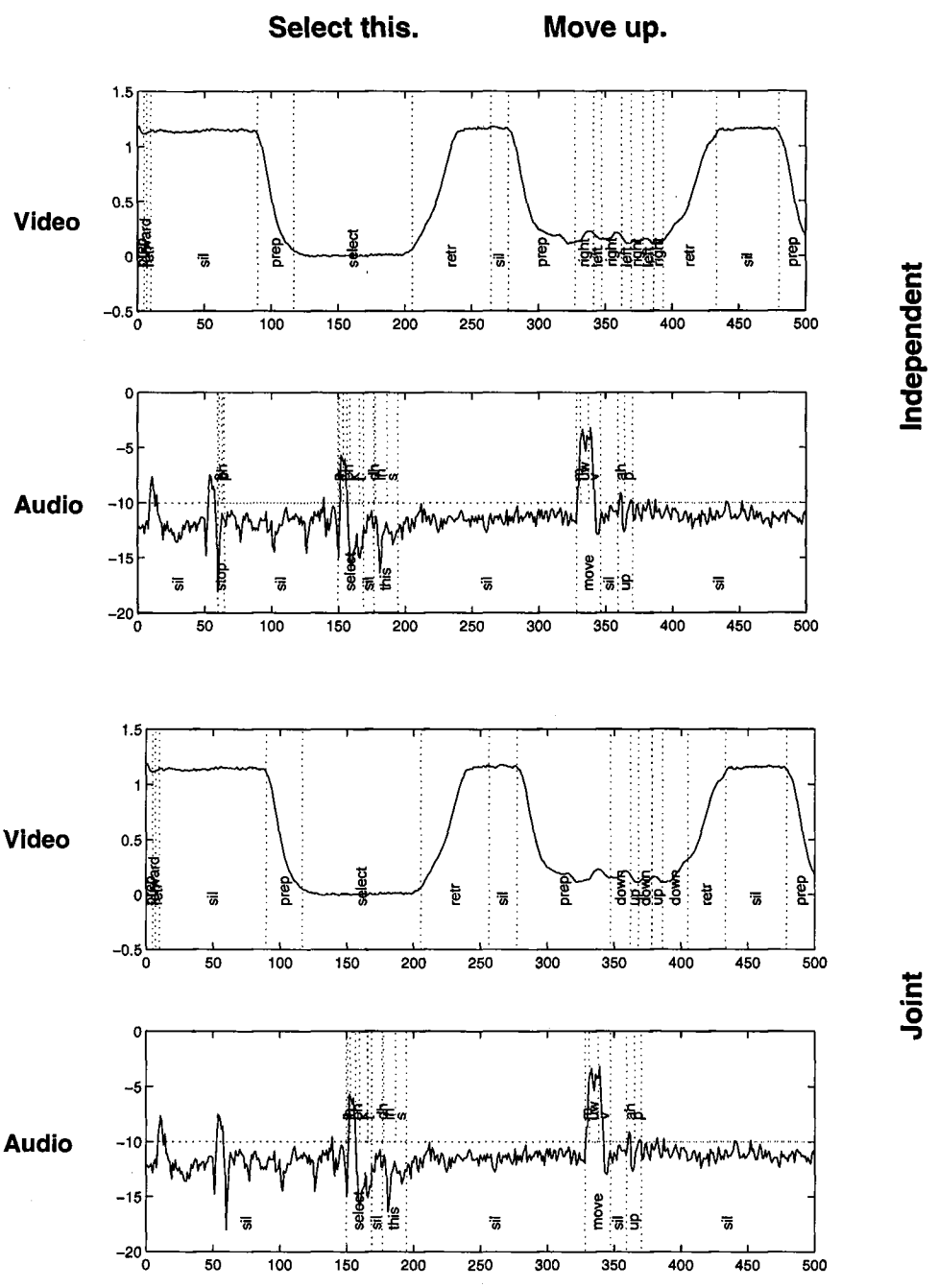

Figure 3: Recognition of spoken words and gestural actions. The figure shows results of temporal segmentation of hand gestures and speech using independent (top two graphs) and joint (bottom two graphs) inference. Depicted features for video and audio streams are the hand angle and a cepstral coefficient, 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.)