Combining User Modeling and Machine Learning to Predict Users' Multimodal Integration Patterns

Xiao Huang[1], Sharon Oviatt[1,2], and Rebecca Lunsford[1,2]

[1] Natural Interaction Systems
10260 Sw Greenburg Road Suite 400
Portland, OR 97223
{Xiao.Huang,Rebecca.Lunsford}@naturalinteraction.com
[2] Center for Human-Computer Communication
Computer Science Department
Oregon Health and Science University
Beaverton, OR 97006
oviatt@csee.ogi.edu

Abstract. Temporal as well as semantic constraints on fusion are at the heart of multimodal system processing. The goal of the present work is to develop *user-adaptive temporal thresholds* with improved performance characteristics over state-of-the-art fixed ones, which can be accomplished by leveraging both empirical user modeling and machine learning techniques to handle the large individual differences in users' multimodal integration patterns. Using simple Naïve Bayes learning methods and a leave-one-out training strategy, our model correctly predicted 88% of users' mixed speech and pen signal input as either unimodal or multimodal, and 91% of their multimodal input as either sequentially or simultaneously integrated. In addition to predicting a user's multimodal pattern in advance of receiving input, predictive accuracies also were evaluated after the first signal's end-point detection─the earliest time when a speech/pen multimodal system makes a decision regarding fusion. This *system-centered metric* yielded accuracies of 90% and 92%, respectively, for classification of unimodal/multimodal and sequential/simultaneous input patterns. In addition, empirical modeling revealed a .92 correlation between users' multimodal integration pattern and their likelihood of interacting multimodally, which may have accounted for the superior learning obtained with training over heterogeneous user data rather than data partitioned by user subtype. Finally, in large part due to guidance from user-modeling, the techniques reported here required as little as 15 samples to predict a "surprise" user's input patterns.

1 Introduction

Techniques for temporal information fusion are at the heart of designing a new generation of multimodal systems. Current state-of-the-art multimodal systems use *fixed temporal thresholds* based on previous modeling of users' natural modality integration patterns [1,2]. However, newer studies [3,4,5] show that there are significant individual differences among users in their multimodal integration patterns, and that adaptive temporal thresholds for multimodal systems could achieve substantial

S. Renals, S. Bengio, and J. Fiscus (Eds.): MLMI 2006, LNCS 4299, pp. 50 – 62, 2006.

improvements in processing speed, accuracy and overall performance [6,7]. Motivated by these newer results, our recent work addresses the development of user-adaptive temporal thresholds for future multimodal systems. The present paper explores combining empirical user modeling and machine learning techniques to quickly learn a user's multimodal integration patterns, and to adapt a multimodal system's temporal thresholds to that user.

1.1 Related Work on Individual Differences in Multimodal Integration Patterns

A series of studies conducted with users across the lifespan has indicated that individual child, adult, and elderly users all adopt either a predominantly *simultaneous* or *sequential* integration pattern during production of speech and pen multimodal constructions [3,4,5,7]. It can be summarized that: 1) previous lifespan data on speech and pen input from over 100 users shows that they are classifiable as either *simultaneous* or *sequential* multimodal integrators (70% simultaneous, 30% sequential); 2) a user's dominant simultaneous or sequential integration pattern can be identified almost immediately; and 3) their integration pattern remains highly consistent throughout an given interaction (88-97% consistent) and over time. 4) Based on previous data, it's clear that users' dominant multimodal integration pattern is strikingly consistent and resistant to change. In addition, behavioral and linguistic differences in the interaction styles of these two groups suggest underlying enduring differences in cognitive style [7].

Based on previous research, Table 1 summarizes the multimodal input ratio of ten adults while interacting with a map-based multimodal system [5,7,8]. Participants interacted multimodally on 62% of the tasks and unimodally on 38%, and there were large individual differences in the ratio of multimodal interaction ranging from 22% to 92%. Given hand annotated data, previous research has indicated that a human rater could predict both a user's dominant multimodal integration pattern (i.e., simultaneous/sequential) and their likelihood of interacting multimodally (i.e., versus unimodally) with 100% accuracy after only 15 commands [8]. All of these findings indicate

Table 1. Average percentage of unimodal vs. multimodal interactions, and sequential vs. simultaneous integration patterns for different user's multimodal interactions ([8])

Subject	Multimodal	Unimodal	SIM	SEQ
1	69%	31%	87%	13%
2	92%	8%	100%	0%
3	62%	38%	90%	10%
4	62%	38%	97%	3%
5	84%	16%	99%	1%
6	89%	11%	98%	2%
7	22%	78%	5%	95%
8	69%	31%	72%	28%
9	41%	59%	97%	3%
10	28%	72%	0%	100%
Consistency	73.6%		93.5%	

that users' multimodal interaction and integration patterns are fertile content for incorporating machine-learning techniques. Future multimodal systems that can *detect and adapt to a user's dominant multimodal integration patterns* could yield substantial improvements in multimodal system robustness and overall performance.

1.2 Related Work Applying Machine Learning to Multimodal Data

In order to build adaptive temporal thresholds, a multimodal system has to be able to learn and adapt to each user's input patterns. However, the general study of adaptive information fusion for multimodal systems is still in its infancy [5]. Apart from standard stream-weighting techniques for optimizing multimodal signal recognition, more recent work has begun investigating and developing new machine learning techniques in areas like adaptive information fusion for audio-visual speech processing, user authentication, and activity classification [9, 10, 11, 12]. Typically, such research uses graphical models (Hidden Markov Models or Bayesian Belief Networks and their extensions) to build models of the relation between different modalities.

For example, Bengio [9, 12] proposed an asynchronous Hidden Markov Model for audio-visual speech recognition and user authentication. This work takes advantage of the inherently close "temporal coupling" of speech and lip movements as modalities, and it requires a large amount of high-quality video and acoustic training data. For example, after training conducted with 185 recordings from 37 subjects, performance with the AHMM exceeded that of an HMM (i.e., yielding 88.6% correct for 9 digits at 10 dB signal-to-noise ratio). In the case of data on users' speech and pen input, these modes are not as closely aligned temporally, and sometimes do not occur in combination at all. As such, this is a more challenging problem than processing closely-coupled multimodal data. One major under-acknowledged prerequisite for processing multimodal speech and pen input is *accurate clustering of users' speech and pen signals into multimodal versus unimodal constructions* before fusion and semantic interpretation take place, which is one goal of the present research.

In other work by Oliver, layered HMMs [10] have been used to infer 6 distinct human activities in an office environment from users' audio-visual activities and mouse input after 1 hour of training. With respect to data requirements, Oliver's methods were more efficient than most others outlined in the literature (i.e., 10 mins. of training for each of 6 activities). High accuracy also was reported for activity classification (over 99%), although generalization across variations in office conditions (e.g., changes in lighting) is known to be a limitation with this type of approach.

In work conducted by Lester, et. al [11], static classifier and HMM models were combined to predict users' physical activities with a wide variety of multimodal sensors (e.g., auditory activity, acceleration) while people were mobile. After training on 4 hours of data at a frequency of 4 Hz, mobile users' activities could be identified with 85% accuracy. In summary, most previous learning models have required relatively large amounts of training data. In contrast, one goal of our current work is to develop accurate learning models for predicting users' multimodal interaction patterns after *minimal training samples* (i.e., as few as 15 samples total, learned over 1-3 mins.), such that they can be deployed easily during *real-time multimodal processing*.

In earlier work on the development of adaptive multimodal processing techniques for handling users' integration patterns, Gupta developed adaptive temporal

thresholds for fusion based on BBNs, which was implemented within a speech/pen multimodal system. In this work, he reported a 40% performance improvement after training on 495 samples, compared with systems that use fixed temporal thresholds [6]. Apart from empirical user modeling in this area (cited in section 1.1), in past work we also have implemented simple Bayesian Belief Network models with discrete variables, which achieved prediction accuracies of 85% in classifying users' multimodal integration patterns after only 15 training samples [8].

1.3 Why Combine Empirical User Modeling and Machine Learning Techniques?

It is well known that in many cases machine-learning techniques [13,14] (e.g., HMMs, Neural Networks) can be computationally intractable unless one has prior knowledge to bootstrap machine-learning models, which is one of our motivations for combining empirical user modeling and machine learning techniques. In addition, other advantages of using empirical modeling to guide machine learning applications include that it can indicate: 1) what content is most fertile for applying learning techniques; 2) what gains can be expected if learning techniques are applied; and 3) when different learning techniques should be applied to handle different subgroups of users adequately. Instead of selecting information sources through trial and error, user modeling also can 4) guide the selection of information sources; 5) indicate how to apply learning techniques so they are transparent and avoid destabilizing users' performance; and 6) reveal how many training samples are needed to train a learning model to achieve a given level of performance. If a model requires too many training samples, it may be inappropriate for real-time learning and/or may not be fertile territory for applying machine learning techniques to certain real-world problems.

1.4 Specific Goals of This Research

In this paper, we combine user modeling and machine learning techniques in an effort to predict users' multimodal integration patterns. Our goals include: (1) conducting further empirical work to discover what type of information may best predict users' multimodal input patterns, which then could be leveraged as prior knowledge to bootstrap machine learning, and (2) determining the best training strategy for improving the predictive accuracy and speed of learning users' multimodal input patterns. With respect to the second goal, we investigated a) the impact of training sample size to discover the optimal amount of training data, b) training over each user's data, data partitioned by user types, and training over all (heterogeneous) user data together, and c) the efficiency of the leave-one-out train-test technique with the present multimodal data, which also can evaluate how well a multimodal system can adapt to a new "surprise" user's input patterns.

Finally, we wanted to develop and test a predictive model that could be used during real-time multimodal system's decision-making regarding whether to fuse input signals before attempting lexical interpretation. Rather than predicting the type of user input completely in advance of receiving it, we evaluated the model's predictive power after end-point detection of a first signal, which is the earliest time at which a speech/pen multimodal system would attempt signal fusion and interpretation.

2 Empirical Study on Users' Multimodal Interaction Patterns

2.1 Study Overview

Data used in this research were collected from 10 volunteer users while interacting spontaneously with a multimodal map-based system. During practice, participants completed 5 tasks using speech only, 5 using pen only, and 5 using both speech and pen. After training, participants were told they could interact with the system any way they wished for the remainder of the session, using speech input, pen input, or multimodal input. To ensure there was no effect of recognition-based system errors on the users' interaction choices, a high-fidelity Wizard-of-Oz system was employed with errors generated at a fixed rate of 20% for all conditions. For further details, see [5].

Input from participants was coded as either unimodally or multimodally delivered. If unimodal, input was scored as either involving speech input or pen input. When multimodal, the integration pattern was coded as either simultaneous (i.e., speech and pen input at least partially overlapped in time), or a sequential one (i.e., one input mode delivered before the other, with a lag between modes). Each participant also was classified as having a dominant unimodal/multimodal and simultaneous/sequential pattern if 60% or more of their input could be classified as that type. An independent second scorer carefully double-checked all of the real-time unimodal and multimodal judgments and multimodal integration patterns to verify their accuracy.

2.2 Results: Correlated Multimodal Integration Patterns

Participants' ratio of simultaneous versus sequential multimodal integrations was strongly correlated to their ratio of multimodal versus unimodal interactions, p=0.92. In fact, 85% of the variance in a participant's likelihood of generating a multimodal versus unimodal construction could be accounted for just by know

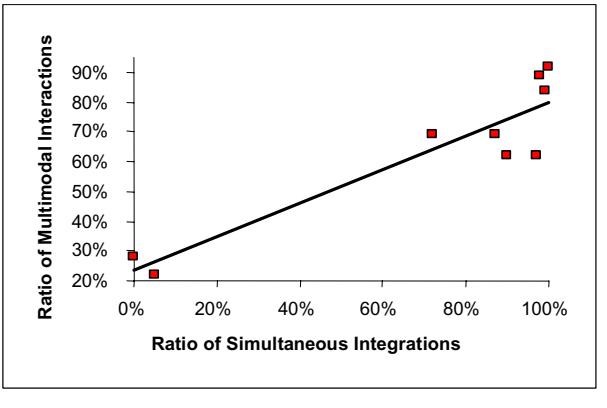

Fig. 1. Linear regression between participants' ratio of simultaneous to sequential multimodal integrations, and their overall likelihood of interacting multimodally

ing their ratio of simultaneous to sequential multimodal integrations during multimodal interactions, which was significant, F=38.3 (df=1,7) $p<.00005$, two-tailed. Figure 1 shows the best fitting linear regression, with the ratio of multimodal interactions increasing in relation to a participant's ratio of simultaneous integrations.

2.3 Discussion

This correlation between participants' ratio of simultaneous to sequential integrations and their likelihood of interacting multimodally is very substantial. Given that users' multimodal integration patterns were highly consistent, this finding provides powerful predictive leverage on correctly classifying a user's subsequent signal patterns. In fact, knowing a user's dominant integration pattern can account for 85% of all the variance in their likelihood of interacting multimodally during subsequent input. In future work, it remains an open question how best to leverage this strong correlation to bootstrap optimal predictive power, although two possibilities are (1) incorporation of more relevant information source into new models, and (2) pursuit of heterogeneous training during machine learning which, given adequate modeling, would be a prerequisite for detecting the regularities between these correlated signal patterns.

3 Machine Learning Approaches for Input Pattern Prediction

In this section, we first provide an introduction to Bayesian Belief Networks and its simplified version, Naïve Bayes. We then compare and present the results of three different training strategies. The general goal was to investigate how best to combine user modeling and machine learning techniques to build a new generation of adaptive multimodal interfaces. More specifically, our results provide guidance for developing user-adaptive temporal thresholds for fusion in future multimodal systems.

3.1 Introduction to Bayesian Belief Network and Naïve Bayes

A Bayesian Belief Network (BBN) [15] is a graphical model that encodes probabilistic relations among discrete related variables. A BBN model can infer causal relations and handle situations where some data are limited or missing. Furthermore, it also is an ideal representation for combining prior knowledge and new training samples.

Naïve Bayes models, a simplified version of BBN, are simple to implement and efficient to train and use, typically producing reasonable predictions compared with more complex learning-based models. However, by assuming that variables are independent and equally important, they also can cause skewed results, especially if many of the variables are interrelated. Because of the ease of implementing Naïve Bayes, we chose this model as a starting point for the present exploratory work even though the multimodal information sources we are modeling are known to be interrelated.

We used the Matlab toolkit [16] to implement a Naïve Bayes model (Figure 2) for this study. The model represents the joint probability distribution of seven variables (four input, three output): 1) Type of current signal: an input variable that represents the type of the modality represented in the current signal (speech, pen, or neither/silence); 2) Duration of current signal: an input variable that has two values, 1 if the duration is longer than the average duration, and 0 if less; 3) Last multimodal

integration pattern: an input variable value of the last multimodal integration pattern (simultaneous, sequential or neither if a unimodal interaction); 4) Last command type: an input variable of the last interaction's unimodal/multimodal value; 5) Type of next signal: an output variable that represents what the new next signal is (i.e., if the interaction is unimodal, the next signal would be silence); 6) Command type: an output variable that represents whether the interaction is unimodal or multimodal; 7) Multimodal integration pattern: an output variable that represents the predicted temporal relationship between the current signal and next signal.

We selected these variables during initial modeling for three reasons. First, they are available and fully annotated in the dataset. Second, they are either discrete or can be rendered as discrete variables, which is compatible with constraints entailed in building discrete BBN models. Third, they represent basic signal and command-level information sources, which are good candidates for initially attempting to predict users' command type and integration pattern.

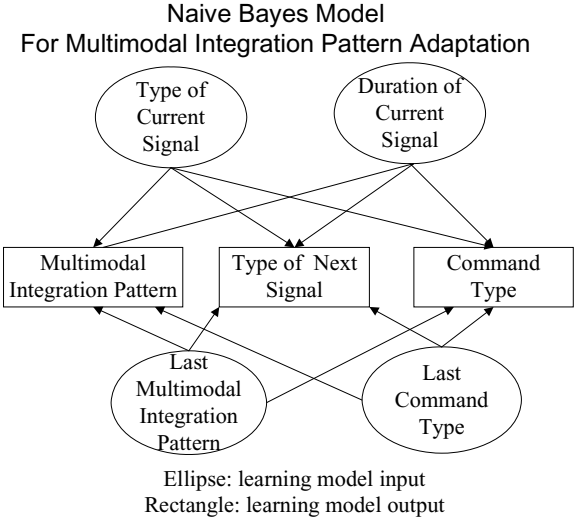

Fig. 2. Machine Learning Model (Naïve Bayes)

3.2 Tests of Machine Learning Approaches

As mentioned earlier, one goal of this work was to investigate means for combining user modeling with machine learning techniques to improve the speed, accuracy and generalizability of predicting users' command type and integration pattern. Towards that end, we investigated different training strategies to determine the best learning models and strategies for predicting users' signal input.

For the first test, we used different size training sets, increasing the number of training samples from 5 to 10, 15, 30, and 45, with the goal of determining how few samples are enough to train a user-adaptive learning model that can adequately predict users' multimodal input patterns. Determining the optimal training sample size

is important in order to know if a given domain is a candidate for online learning and also to avoid overtraining. If a learning model needs too many training samples, then real-time online learning could become intractable.

The second test was to partition train/test sets according to the two main types of user interaction pattern (i.e. unimodal or multimodal), and to determine whether these input patterns which represent different user groups may benefit from applying different learning techniques. As a result, a Naïve Bayes model was built for each user subset. The goal was to examine whether partitioning based on prior user-modeling knowledge could bootstrap the accuracy of prediction yielded by machine-learning.

The third test involved applying the leave-one-out technique, which is a typical strategy for organizing training and test subsets of data during evaluation of machine learning methods. We divided the entire dataset into two subsets: a set including the data from one subject (A) and another set containing data from the rest of the subjects (B). Set B was defined as the training set, while set A was the test set. Training and test data were recomputed 10 times for each of the 10 subjects in this manner, and then averaged. Unlike the partitioning during test 2, this average predictive accuracy represented training across the full heterogeneous group of all diverse users. To the extent the model shown in Figure 2 incorporates information sources involved in the correlated signal patterns reported in section 2 (i.e., between users' multimodal integration pattern and their likelihood of interacting multimodally), then predictive accuracy would be expected to improve with training over the more heterogeneous data involved in this third test, in comparison with user-partitioned training in test 2.

3.3 System-Centered Evaluation Metric of Machine Learning

In addition to predicting a given user's multimodal signal pattern *in advance of receiving a construction,* which was the learning metric compared during the first three tests, predictive accuracies also were evaluated during a fourth test *after the first signal's end-point detection*— which is the earliest time when a speech/pen multimodal system needs to make a decision regarding fusion. This *system-centered metric* was developed because we also need learning models that are capable of real-time prediction and decision-making about whether to complete lexical interpretation of an incoming signal after detecting its' end-point— or to wait and fuse the signal with a later arriving one before interpreting their joint meaning. Therefore, in this test we assumed that the model knows the end-point of the user's first signal. If the input pattern is multimodal and simultaneous, then readiness for lexical processing is clear and there is no need for prediction. Otherwise, the system needs to decide whether the present signal is unimodal, or multimodal but a part of sequentially-integrated construction. Instead of using a fixed temporal threshold which requires waiting 2-4 seconds before resolving this ambiguity, a system with a user-adaptive temporal threshold can weight the likelihood that an upcoming construction is unimodal or multimodal based on previous history. With this system-centered processing viewpoint in mind, a fourth machine learning test was conducted based on a new model.

4 Machine Learning Results

4.1 Increasing the Number of Training Samples

In this experiment, we built a Naïve Bayes model for each subject. The number of training samples varied from 5 to 10, 15, 30 and 45. The number of testing sample was fixed in all cases at 38. As shown in Figure 3, the average prediction accuracies for unimodal/multimodal and simultaneous/sequential for 5 and 10 samples were relatively low (5 samples: 64% and 58%; 10 samples: 74% and 68%). In contrast, using 15 samples, the performance improved substantially (79% and 81%). Further increasing the number of training samples provided minimal improvement beyond this (30 samples: 78% and 79%; 45 samples: 85% and 82%). These results are consistent with previous empirical results [7], in which using the first 15 samples for each user was sufficient to provide optimal classification of users' dominant input patterns.

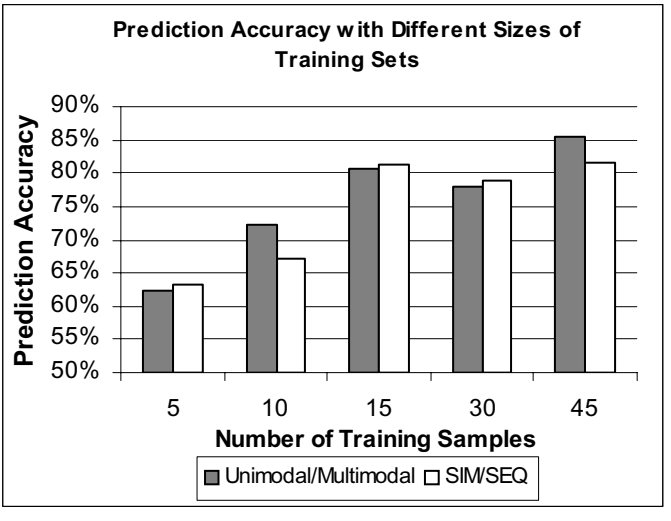

Fig. 3. Prediction accuracies of unimodal/multimodal and simultaneous/sequential patterns, with training sample sizes varying from 5, 10, 15, 30 to 45, respectively

4.2 Partitioning Training into Unimodal/Multimodal User Subsets

In this experiment, we partitioned the dataset into two user subsets. Set 1 includes all habitually unimodal subjects. Set 2 includes the multimodal subjects. We built a Naïve Bayes model for each subset, and compared the results with a model built for each individual subject. For Set 1, there are 45 training samples (i.e., 15 training samples from each of 3 unimodal subjects) and 214 test samples (i.e., 68 from each of 3 subjects). For Set 2, there are 105 training samples and 476 test samples (i.e., based on 7 subjects total). We also conducted a "Baseline" experiment by building a learning model for each individual subject (15 training samples and 68 test samples), yielding 10 total. The average prediction accuracies for unimodal/multimodal and simultaneous/sequential were 79.4% and 81.3%, respectively, for the baseline model. With the

data partitioned by user subtype, the average prediction accuracies were 83.5% and 77.9%, respectively, which was similar to accuracy of the baseline model.

4.3 Leave-One-Out Test Method

In this experiment, we used the last 68 samples from a given user as the test set and the first 15 samples from the other users (135 total samples) as the training set, and then repeated the procedure 10 times, once for each subject. Using this training strategy, 88% of users' natural mixed input could be correctly classified as either unimodal or multimodal, and 91% of users' multimodal input could be correctly classified as either sequentially or simultaneously integrated, as shown in Figure 4. These high predictive accuracies exceeded the results achieved after partitioning training by user subtypes. This performance level may have derived in part from training across heterogeneous user data, which permitted learning of the strong correlation between multimodal information sources outlined in section 2.

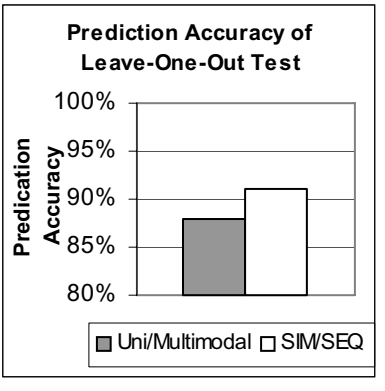

Fig. 4. Prediction accuracy for classifying unimodal/multimodal and simultaneous/sequential patterns based on the leave-one-out training strategy

4.4 Learning Methods Applied to System-Centered Fusion Process

The "Type of Current Signal" variable used in the previous model had three possible values: speech, pen and silence. In this experiment, we assumed the model is applied after the end-point of the first signal, and that the "Type of Current Signal" variable has one more value—"Both signals" (i.e., co-occurring). For this evaluation, we built a learning model for each subject. The first 15 samples were used for training, and the remaining 68 samples for testing. Using this model, 90% of users' natural mixed input could be correctly classified as either unimodal or multimodal, and 92% of their multimodal input was correctly classified as either sequentially or simultaneously integrated, as shown in Figure 5. These high accuracies indicate that real-time systems could be very effectively guided by user-adaptive predictions during the actual process of fusion and lexical interpretation.

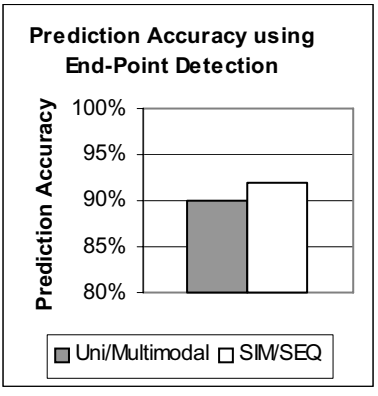

Fig. 5. Prediction accuracy for classifying unimodal/multimodal and simultaneous/sequential patterns based on end-point detection

5 Discussion and Future Work

In this paper, we combined user modeling with machine learning approaches to leverage better prediction of users' multimodal integration patterns. Motivated by previous empirical results, we investigated three different training strategies and two separate metrics of machine learning performance. Using just 15 training samples for each subject, the Naïve Bayes learning model achieved 79% prediction accuracy for unimodal/multimodal classification of users' input, and 81% accuracy for users' simultaneous versus sequential multimodal constructions. Increasing the number of training samples beyond this did not further enhance prediction accuracy. This result is consistent with past empirical studies, in which users' dominant patterns could be classified by humans with 100% accuracy after 15 samples, based on hand annotations. Secondly, we divided the training data into two user subgroups based on each subject's dominant multimodal interaction pattern (unimodal vs. multimodal). However, predictive accuracies based on partitioned data did not exceed rates achieved by training on more heterogeneous combined data during the leave-one-out test. In the third leave-one-out test, the machine learning model correctly classified 88% of users' natural mixed input as either unimodal or multimodal, and 91% of users' multimodal input as either sequentially or simultaneously integrated. These high predictive accuracies may have been due in part to the fact that this model was trained on hetero geneous user patterns, which would have enabled learning of the high correlation between information sources that was summarized on section 2. Finally, system-centered modeling that involved prediction after end-point detection of the first signal revealed accuracies for classifying unimodal/multimodal input of 90%, with classification of simultaneous/sequential multimodal integrations at 92%. These high accuracy rates based on a simple Naïve Bayes approach create a promising basis for developing a new generation of multimodal systems with adaptive temporal thresholds.

The long-term goal of this research is automatic learning and real-time system adaptation to users' multimodal integration patterns, as well as the development of new strategies for combining empirical user modeling with machine learning techniques to

bootstrap the accelerated, generalized, & improved reliability of information fusion in new types of multimodal systems─ including ones involving different modalities and applications. Based on this work, it is clear that empirical user modeling can guide machine learning techniques by uncovering fertile applications and valuable information sources. It also can provide insights into why machine learning succeeds when it does, which will be valuable for generalizing machine learning techniques successfully. Future work will need to explore the performance of more sophisticated learning models, such as asynchronous HMM models [12] and Markov Logical Networks [17] at handling this type of multimodal integration data. In addition, future work should develop learning models based on more precise continuous temporal information, so that users' average signal overlap or lag can be predicted more precisely.

Acknowledgments

Thanks to Benfang Xiao and Josh Flanders for assistance with data collection. This research was supported by DARPA Contract No. NBCHD030010 and NSF Grant No. IIS-0117868. Any opinions, findings or conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Project Agency, or the Department of the Interior.

References

1. S. Oviatt. Ten myths of multimodal interaction. Comm. of the ACM, Vol. 42(11). (1999) 74–81
2. S. Oviatt. Integration and synchronization of input modes during multimodal human computer interaction. In: Proc. of CHI. (1997) 415–422
3. S. Oviatt, R. Coulston, S. Tomko, B. Xiao, R. Lunsford, M. Wesson, and L. Carmichael, Toward a theory of organized multimodal integration patterns during human-computer interaction. In: Proc. of ICMI. (2003) 44-51
4. B. Xiao, R. Lunsford, R. Coulston, M. Wesson, and S. Oviatt. Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. In: Proc. of ICMI. (2003) 265–272
5. S. Oviatt, R. Coulston, and R. Lunsford. When do we interact multimodally? Cognitive load and multimodal communication patterns. In: Proc. of ICMI. (2004) 129-136
6. A. Gupta and T. Anastasakos, Dynamic time windows for multimodal input fusion, In: Proc. of Interspeech. (2004) 2293-2296
7. S. Oviatt, R. Lunsford and R. Coulston: Individual differences in multimodal integration patterns: What are they and why do they exist? In: Proc. of CHI. (2005) 241-249
8. X. Huang and S. Oviatt, Towards adaptive information fusion in multimodal systems, In: Proc. of MLMI. (2005) 15-27
9. S. Bengio. An asynchronous hidden Markov model for audio-visual speech recognition. In: Proc. of Advances in Neural Information Processing Systems. (2003) 1213–1220
10. N. Oliver, A. Garg and E. Horvitz, Layered representations for learning and inferring office activity from multiple sensory channels, Int. Journal on Computer Vision and Image Understanding, 96(2) (2004) 163-180
11. J. Lester, T. Choudhury and G. Borriello, A Practical approach to recognizing physical activities, To appear in the Proc. of Pervasive. (2006)

12. S. Bengio. Multimodal authentication using asynchronous HMMs. In: Proc. of AVBPA. (2003) 770–777
13. L. Rabiner, A tutorial on hidden Markov model and selected applications in speech recognition. In: Proc. of the IEEE, Vol.77, No.2 (1989) 257-286
14. R. Duda, P. Hart and D. Stork, Pattern classification, Morgan Kaufmann (2002)
15. D. Heckerman. A tutorial on learning with Bayesian networks. Learning in Graphical Modals. MIT Press (1999)
16. K. Murphy. The Bayes net toolbox for Matlab. Computing Science and Statistics, Vol. 33. (2001)
17. M. Richardson and P. Domingos, Markov Logic Networks, Machine Learning, Vol. 62. (2006) 107-136