Segmentally Boosted HMM

Computational Perception Lab

People

- Pei Yin (primary contact)
- Prof. Irfan Essa,
- Prof. Thad Starner, and
- Prof. James M. Rehg

Goal

Discriminative feature selection for hidden Markov Models in sequence classification.

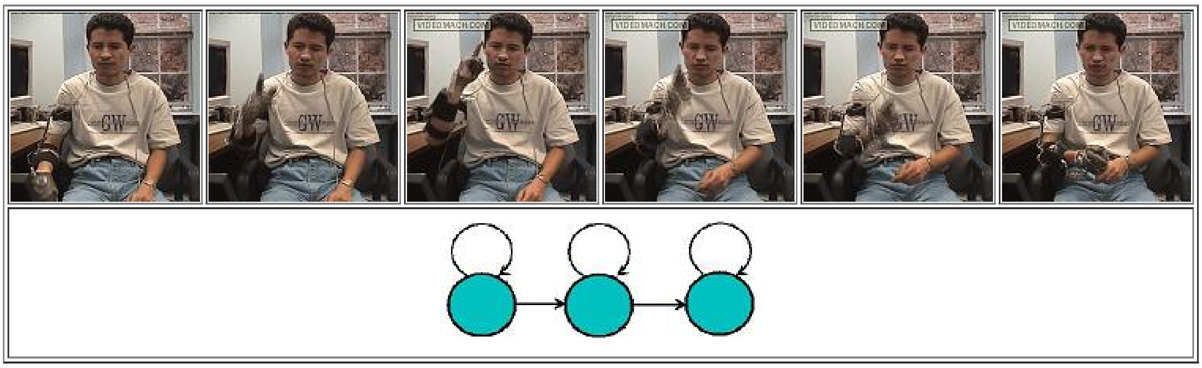
Introduction

Speech recognition, gesture recognition, DNA analysis and many other pattern recognition tasks for time-series data are sequence classification problems, which predict of a single label for an entire sequence. The most successful technique for sequence classification is the hidden Markov Model (HMM). The recognition accuracy and efficiency of HMMs can be improved with discriminative features. Traditional feature selection methods require the data being independently and identically distributed (i.i.d.), but time sequences usually contain strong temporal correlation between the adjacent observation frames. Furthermore, features in time sequences may be "sometimes informative", that is, discriminative only in some segments of a sequence. In this research, we propose Segmentally-Boosted HMMs (SBHMMs), which is able to address the problems of both temporal correlation and segmentally-informative features by assuming "piecewise i.i.d." Experiments show that the SBHMMs consistently improve traditional HMM recognition in American Sign Language recognition, human gait identification, lip reading and speech recognition. The reduction of error ranges from 17% to 70%.

Conditional Random Fields [Lafferty, et.al., 2001] or Tandem models [Hermansky, et.al., 2000] require a state-level labeling by human or forced alignment. In practice, such labeling may not be possible. For example, to recognize the sign brother, how can the human labeler precisely supervise the training for the first state when he does not even know the states meaning or how many states comprise the sign? Without such labeling, the discriminative feature selection will be limited to the level of the entire signs or phonemes. In contrast, SBHMMs are truely designed for sequence classification, in which the sub-sequence components are unknown, and our experiments show that the feature selection at the sub-sequence (state) level achieves superiour performance than that at the sequence level.

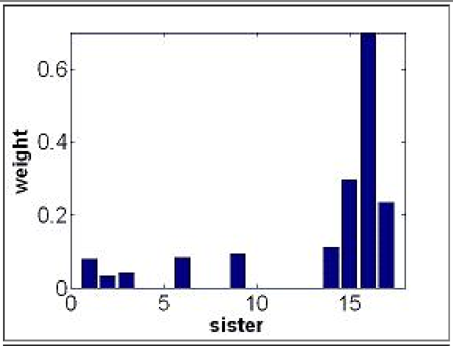
Approach

1. Train HMMs for input sequences (the example below is the sign "sister" in ASL).

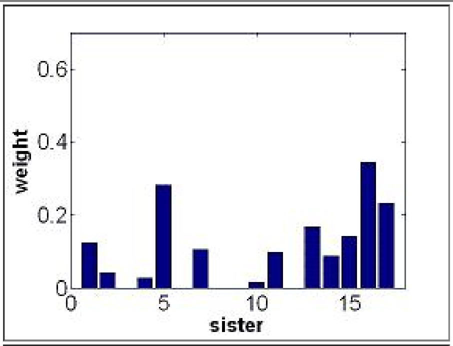


2. Label each frame automatically with its most likely state computed from the Viterbi decoding.

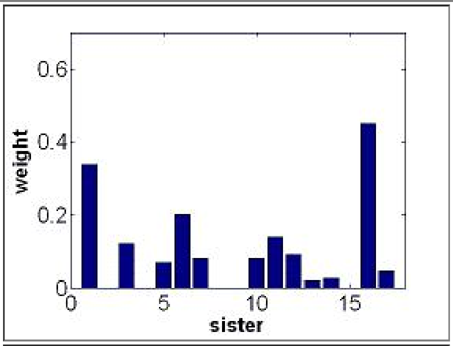
3. Train AdaBoost for this labeling.



feature weights for state 1



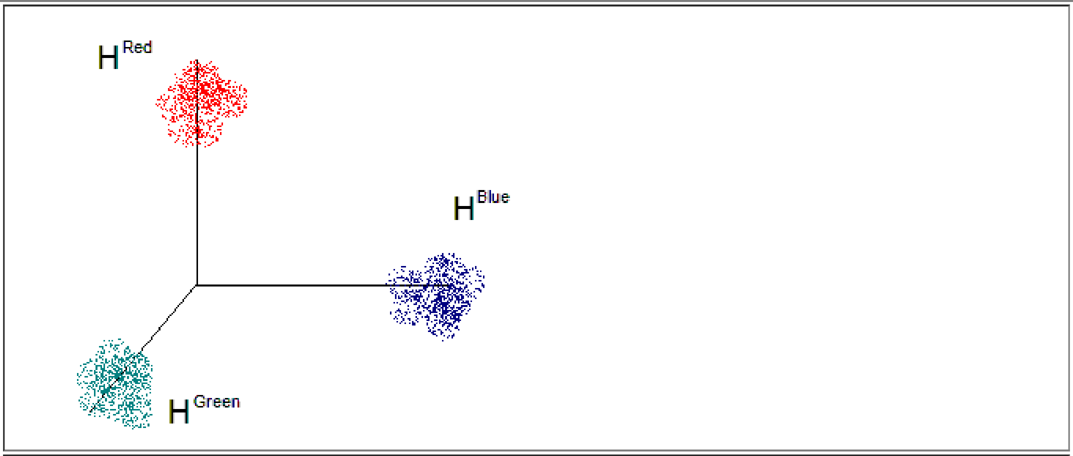
feature weights for state 2



feature weights for state 3

AdaBoost constructs classification ensembles for each state of every HMM. The ensembles linearly combine the decision stumps over 17 input features (accelerometer readings). The weights for the combination, which tell the importance of the features, are greedily chosen to achieve high accuracy in separating one state from all the other states (in the same HMM and in different HMMs). This figure show that our feature selection technique assigns more weight to the accelerometer readings number 15-17, the shoulder elevation, in state 1. This is consistant with the linguistics of ASL.

4. The new feature space V is the output space of the AdaBoost ensembles.



In real world applications, the dimension of V is the total number of classifiers, which is also the total number of the states in all HMMs. This figure shows a much simpler example, in which we only have three classes: Red, Green, and Blue. Any sample x in the original space is converted/projected to coordinate (H^(Red)(x), H^(Green)(x), H^(Blue)(x)). In the ideal (separable) situation, the samples with the same state label will be at the same corner in V, as shown in the figure. In the noisy (inseparable) situation, the aggregation property of the margins, which is described in [Yin, et.al., 2008], still drives the samples towards the "ideal" direction.

The new feature space is more discriminative due to the feature selection procedure, and it fits the Gaussian observation model better in our experiments. For details please see [Yin, et.al., 2008].

In cases where there are too many states, unsupervised dimension reduction methods, such as PCA can be employed. We did not use such method in computing V for the evaluation reported in [Yin, et.al., 2008].

5. Train new HMMs in V. Those HMMs have higher accuracy due to the discriminative features.

Data

- **Georgia Tech Speech Reading Data**

Description: Continuous audio-visual speech recognition data, audio captured by one microphone at 16kHz and visual markers captured by Motion Capture devices at 120Hz.

Total Length	30m45s	MoCap Rate	120Hz
Training Data	24m42s	Testing Data	06m03s
Total Sentences	275	Total Phones	8468
Total Phonemes	39	Total Samples	> 200,000

Download: Compressed file (84MB): gtsr.rar

- **Georgia Tech Gait Recognition Data**

Description: Gait identification data captured by Motion Capture (MoCap) devices at 120Hz.

It contains 22 tracker readings for 15 subjects. We used three leg-related readings of the first five subjects following the convention in [Kim and Pavlovic, 2006]

Download: link (via FTP)

- **MIT American Sign Language Recognition Data**

Description: Continuous ASL data captured by video cameras mounted on the hat.

It contains 500 five-sign sentences composed of 40 different signs by one subject.

Download: starner97.zip from Contextual Computing Group (CCG) at Georgia Tech

- **Georgia Tech American Sign Language Recognition Data**

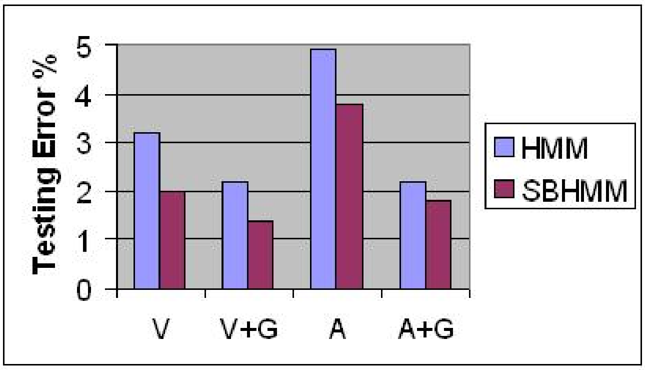
Description: Continuous ASL data captured by accelerometers on the gloves

It contains 665 four-sign sentences composed of 141 different signs by one subject.

Download: acceleglove.zip from Contextual Computing Group (CCG) at Georgia Tech.

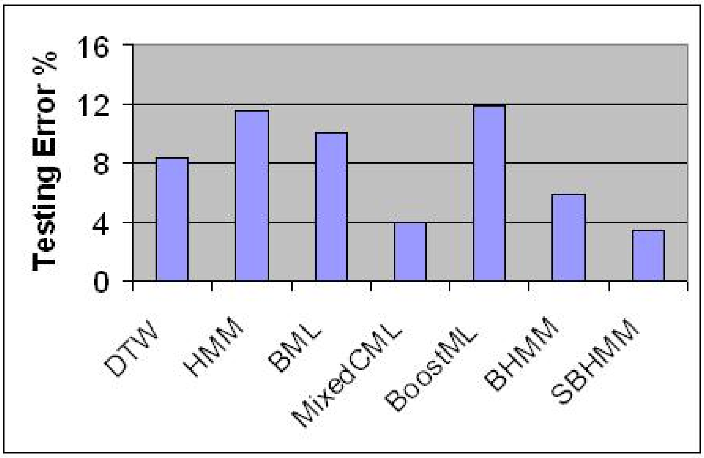
Experimental Results

- **American Sign Language Recognition on MIT data and Georgia Tech data**



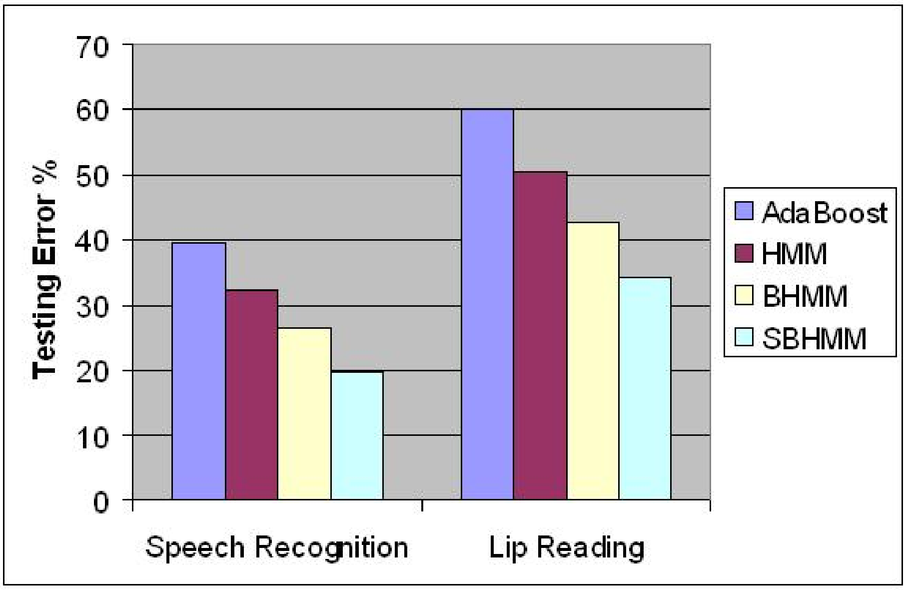
V: vision-based MIT data, A: accelerometer-based Georgia Tech data, and G: using grammar for postprocessing.

- **Georgia Tech Gait Recognition data**



The performance for DTW, HMM, BML, MixedCML, BoostedML are directly from [Kim and Pavlovic, 2006]. BHMM is boosted HMM in [Yin, et.al., 2004]. Please refer to these two papers for the details of the data and the experiments conducted.

- **Georgia Tech Audio-Visual Speech Recognition Data**



In this chart, we report the testing error for phoneme recognition. We only use audio information in speech recognition and visual information in lip reading. Note that audio-visual fusion may further reduce the recognition error as in [Yin, et.al., 2003]. For the details of the data and the experiments conducted, please refer to [Yin, et.al., 2004] and [Yin, et.al., 2008].

Publications

Pei Yin, Irfan Essa, Thad Starner, James M. Rehg, "Discriminative Feature Selection for Hidden Markov Models Using Segmental Boosting", in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Mar. 2008. (pdf) (bibtex).

Pei Yin, Irfan Essa, James M. Rehg, "Segmental Boosting Algorithm for Time-series Analysis," in Snowbird Machine Learning Workshop, Mar. 2007.

Pei Yin, Irfan Essa, James M. Rehg, "Asymmetrically Boosted HMM for Speech Reading," in Proc. of *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, pp II755-761, June 2004 (pdf) (bibtex)

Pei Yin, Irfan Essa, James M. Rehg, "Boosted Audio-Visual HMM for Speech Reading," in Proc. of *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, pp 68-73, Oct. 2003/held in conjunction with *ICCV-2003*. A version of this paper also appears in Proc. of *Asilomar Conference on Signals, Systems, and Computers*, pp 2013-2018, Nov. 2003 as an invited paper. (pdf) (bibtex)

Acknowledgements

We would like to thank Professor Biing-Hwang (Fred) Juang for his guidance with some of the acoustic processing parts of this research, Professor Aaron Bobick, Jianxin Wu, Qiang Fu, Matthew Mullin, Zhenhao Zhou, Yifan Shi, Jie Sun and Fabien Robert for the discussions, and our subjects for volunteering their time. This project is partly funded by National Science Foundation ITR grant #IIS-0205507 and IIS#0511900.

Code Download

Integrated with HTK

(coming soon)

Home	Projects	Publications	People	Courses	Sponsors	Co-Web CPL Swiki	GVU

Last Updated Apr. 7, 2008.