Co-Adaptation of Audio-Visual Speech and Gesture Classifiers

C. Mario Christoudias, Kate Saenko, Louis-Philippe Morency and Trevor Darrell
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar Street, Cambridge, MA 02139, USA
cmch,saenko,lmorency,trevor@csail.mit.edu

ABSTRACT

The construction of robust multimodal interfaces often requires large amounts of labeled training data to account for cross-user differences and variation in the environment. In this work, we investigate whether unlabeled training data can be leveraged to build more reliable audio-visual classifiers through co-training, a multi-view learning algorithm. Multimodal tasks are good candidates for multi-view learning, since each modality provides a potentially redundant view to the learning algorithm. We apply co-training to two problems: audio-visual speech unit classification, and user agreement recognition using spoken utterances and head gestures. We demonstrate that multimodal co-training can be used to learn from only a few labeled examples in one or both of the audio-visual modalities. We also propose a co-adaptation algorithm, which adapts existing audio-visual classifiers to a particular user or noise condition by leveraging the redundancy in the unlabeled data.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms

Keywords

Semi-supervised learning, Adaptation, Audio-visual speech and gesture, Co-training, Human-computer interfaces

ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

1. INTRODUCTION

Human interaction relies on multiple redundant modalities to robustly convey information. Similarly, many human-computer interface (HCI) systems use multiple modes of input and output to increase robustness in the presence of noise (e.g. by performing audio-visual speech recognition) and to improve the naturalness of the interaction (e.g. by allowing gesture input in addition to speech). Such systems often employ classifiers based on supervised learning methods which require manually labeled data. However, obtaining large amounts of labeled data is costly, especially for systems that must handle multiple users and realistic (noisy) environments. In this paper, we address the issue of learning multi-modal classifiers in a semi-supervised manner. We present a method that improves the performance of existing classifiers on new users and noise conditions without requiring any additional labeled data.

There has been much interest recently in developing semi-supervised learning algorithms for problems with multiple views of the data. One such algorithm, co-training [2], improves weak classifiers learned on separate views of the labeled data by maximizing their agreement on the unlabeled data. Co-training has been shown to work for a variety of multi-view problems in natural language and other domains [3, 8, 9]. Typically, it works well in settings where only a small amount of labeled data is available along with a large amount of unlabeled data. A range of classifiers has been explored, including naive Bayes [2], mixture models [1], and support vector machines [12].

In the first part of the paper, we explore co-training for two audio-visual tasks: speech unit classification and user agreement detection. The first task is to identify a sequence of acoustic and lip image features as a particular word or phoneme. The second task is to determine whether a user has expressed agreement or disagreement during a conversation, given a sequence of head gesture and acoustic features. Although we only deal with isolated sequences, the algorithm can be extended to continuous recognition. As the core classifier in the co-training paradigm, we use the hidden Markov model (HMM), which is common for speech and gesture sequence classification.

Co-training was originally proposed for the scenario in which labeled data is scarce but unlabeled data is easy to collect. In multimodal HCI development, it may be feasible to collect enough labeled data from a set of users in a certain environment, but the resulting system may not generalize well to new users or environments. For example, a new user may gesture differently, or the room may become noisy when a fan is turned on. The semi-supervised learning problem then becomes one of adapting existing models to the particular condition. To solve this problem, we investigate a variant of co-training, which we call *co-adaptation*. Co-adaptation uses a generic supervised classifier to produce an initial labeled training set for the new condition, from which a data-specific classifier is built. The algorithm then improves the resulting data-specific classifier with co-training, using the remaining unlabeled samples.

We begin in the following section with a discussion of related work. Co-training is then described in the context of audio-visual classification in Section 3. Our co-adaptation algorithm is presented in Section 4. Experiments and results are described in Section 5. Finally, a summary and a discussion of future work are given in Section 6.

2. RELATED WORK

Co-training is a multi-view, semi-supervised learning algorithm originally developed by Blum and Mitchell [2]. It relies on multiple, independent views of the learning problem to learn a classifier from a small amount of labeled training data (see the following section for a more detailed discussion of the co-training algorithm). In the area of natural language processing, Collins and Singer [3] demonstrated how co-training can be used to learn a named entity recognizer from spelling and context views with little training data. Similarly, in the field of computer vision, Levin et al. [8] used co-training to learn a vision-based car detector from intensity and motion views.

Multimodal classification is well suited for multi-view learning because each modality provides a potentially redundant view to the learning algorithm. While the concept of multimodal co-training was mentioned as promising future work in the seminal Blum and Mitchel paper [2], it appears that there has been little subsequent work on cross-modal co-training. Li and Ogihara [9] use a multi-view learning algorithm applied to gene expression and phylogenetic data to perform gene function classification. Yan and Naphade [14] investigate co-training for semantic concept detection in video.

In this paper, we investigate the use of multimodal co-training for learning speech and gesture classifiers. To our knowledge, this is the first paper to use co-training in the context of audio-visual speech and gesture. We demonstrate that co-training can be successfully used to learn from noisy audio-visual speech and gesture data.

The development of user-adaptive multimodal interfaces is a growing area of research. Adaptation to a user's multimodal discourse patterns is known to be important, as users exhibit different interaction styles based on factors such as age and environment [13]. While we focus on improving the accuracy of low-level appearance, motion, and acoustic models, we believe our appoach will also be useful in adapting timing and fusion parameters. A different approach to multimodal adaptation is to design a system where the user adapts to the system's recognition capabilities while the system attempts to simultaneously adapt to the user [11]. In the context of audio-visual HMMs, maximum likelihood linear rotation (MLLR) has been recently used for speaker adaptation [7]. Semi-supervised recognition of agreement and disagreement in meeting data using prosodic and word-based features was proposed in Hillard, Ostendorf, and Shriberg [6].

Algorithm 1 Co-training Algorithm

```
Given a small labeled set $L$, a large unlabeled set $U$, $k$ views, and parameters $N$ and $T$:
Set $t = 1$
repeat
  for $i = 1$ to $k$ do
    Train classifier $f_i$ on view $i$ of $L$
    Use $f_i$ to label $U$, move $N$ most confidently labeled samples to $L$
  end for
  Set $t = t + 1$
until $t = T$ or $|U| = 0$
```

3. AUDIO-VISUAL CO-TRAINING

The intuition behind the co-training algorithm is that classifiers operating on independent views of the data can help train each other by sharing their most confident labels. The generic co-training procedure is given as Algorithm 1. Initially, a small set L of labeled examples is used to train weak classifiers in each view. We call this the *seed* set. Then, the seed classifiers are used to assign labels to the unlabeled set U, and the N most confidently labeled samples from each classifier are moved to L. The expanded seed set is then used to re-train the classifiers. This continues for several iterations, until either the maximum number of iterations T is reached, or the set U becomes empty. The success of the algorithm depends on two assumptions: the conditional independence of the views, and the sufficiency of each view to learn the target function.

Although co-training has been applied to natural language [3] and other single-modality tasks (e.g. [8]), it is unclear whether the assumptions required for its success will hold in the case of multi-modal HCI problems. We will now discuss what makes these problems different and how it may affect the training algorithm.

Co-training exploits the redundancy in the disjoint sets of features used to identify categories. Such redundancy is, in fact, what makes multimodal tasks seem so well-suited to co-training: The spoken utterance "yes" and a head nod are redundant indications of user agreement; facial appearance and voice both convey user identity, etc. However, the assumption that each modality is sufficient for classification does not always hold. For example, the user can indicate agreement just by nodding and not providing any spoken feedback, or by nodding while saying something that does not explicitly state agreement. Another issue related to sufficiency is that the observations belonging to a particular category may not be aligned in time across modalities and may have variable-length segmentations. In this paper, we make sure that for each segmented time period, each view in the training data is sufficient to identify the correct class.

The other assumption made by the co-training paradigm is that of class-conditional independence of views. This seems like a reasonable assumption in the case of multiple modalities. In fact, the same assumption is made by many multimodal fusion models which express the class-conditional likelihood of a multimodal observation as the product of the observation likelihoods for each modality.

Finally, the original formulation of the co-training algo-

Algorithm 2 Co-Adaptation Algorithm

```
Given user-independent classifiers $f_i^{UI}$, $i = 1, ...k$, a user-dependent unlabeled set $U$ and parameters $N$, $M$ and $T$:
set $S = \emptyset$
for $i = 1$ to $k$ do
    Use $f_i^{UI}$ to label the $M$ highest-confidence samples in $U$ and move them to $S$
end for
Set $t = 1$
repeat
    for $i = 1$ to $k$ do
        Train user-dependent classifier $f_i$ on view $i$ of $S$
        Use $f_i$ to label $N$ highest-confidence samples in $U$ and move them to $S$
    end for
    Set $t = t + 1$
until $t = T$ or $|U| = 0$
```

rithm [2] relies on weak classifiers trained on a small quantity of labeled data to provide new labels at each iteration. To ensure that the quality of the labeled data does not deteriorate, the classifiers need to either have a low false positive rate, or reliable confidence estimates. While this may be possible for text classification tasks, it is harder to acheive for noisy multi-modal observations. In our formulation, which uses HMM classifiers, we compute confidence values as follows. Let x_i be an observation in modality i, and y be one of $1, ..., n$ labels. Then the posterior probability of y given x_i is

$$P(y|x_i) = \frac{P(x_i|y)P(y)}{\sum_{u=1}^{n} P(x_i|u)P(u)} \quad (1)$$

where the likelihood of x_i given the label is obtained from the HMM classifier f_i for each class. We use the posterior probability computed in (1) as the confidence value to assess the reliability of labels assigned to the unlabeled samples during co-training.

4. CO-ADAPTATION ALGORITHM

Co-training was proposed for the scenario where labeled data is scarce but unlabeled data is easy to collect. In certain multimodal HCI applications, it may be feasible to collect a lot of labeled data to train a model on a particular set of users and environmental conditions (audio noise level, lighting, sensing equipment, etc.) However, such a model may not generalize well to new users and conditions.

To address this issue, we propose an adaptive version of the co-training algorithm that bootstraps a data-dependent model from a data-independent model trained on a large labeled dataset. Suppose we obtain unlabeled data from a new condition, such as a new user. We first use the user-independent model to specify a small seed set of labeled examples using its most confident predictions. A user-dependent model is then trained on this initial seed set and improved with cross-modal learning on the rest of the unlabeled data. The resulting co-adaptation algorithm is summarized as Algorithm 2.

The intuition behind the co-adaptation algorithm is that, while the overall performance of the generic model may be poor on new users or under new noise conditions, it can still be used to accurately label a small seed set of examples. The initial seed classifier can then be improved via co-training. Since the new classifier is trained using samples from the new working condition (i.e., new user and environment), it has the potential to out-perform the original generic classifier in the new setting, especially when user variation or difference in environment is large.

Note that, in Algorithm 2, a new user-dependent model is trained on the unlabeled data instead of adding the new labels to the user-independent labeled set. The advantage of this approach is that it is better suited to situations where there is a large imbalance between the amount of labeled and unlabeled data. Alternatively, we could use the new labels to adapt the parameters of the existing model using an HMM adaptation technique such as maximum likelihood linear rotation (MLLR)[7]. The advantage of training a separate user-dependent model is that it enables us to use data-dependent features. For example, we can train a new model with higher-resolution visual observations, or apply data-dependent principal component analysis (PCA). We leave this as a future work direction.

5. EXPERIMENTS

To evaluate our co-training framework, we apply it to two different multimodal tasks: speech unit classification and agreement recognition in human-computer dialogue. Both tasks exploit the audio and the visual modalities, and are typical examples of HCI applications.

In all experiments, we use correct classification rate (CCR) as the evaluation metric, defined as

$$\text{CCR} = \frac{\text{\# sequences correctly classified}}{\text{total \# of sequences}}.$$

We compare the co-adaptation algorithm to two other semi-supervised methods [4]. The first method uses the top N most confidently classified examples from one modality to train a classifier in the other modality. As we show in our experiments, this method is only beneficial when the relative performance of the classifiers on the unlabeled data is known a priori, so that stronger classifiers can be used to improve weaker ones. We show that co-adaptation can achieve the same or better improvements in performance without the need for such prior knowledge.

The second baseline we consider is single-modality bootstrapping, which does not use cross-modal learning, but rather learns a semi-supervised classifier separately in each modality. It is similar to co-adaptation (Algorithm 2), except that each classifier operates on its own copy of U and S, and classification labels are not shared across modalities. As we demonstrate in our experiments, cross-modal learning algorithms are better at improving weak classifiers than single-modality bootstrapping, especially when one modality is more reliable than the other.

In the following experiments, we use left-to-right HMMs with a mixture of Gaussians observation model.

5.1 Audio-Visual Agreement Recognition

In this section, we apply multimodal co-training to the task of recognizing user agreement during multimodal interaction with a conversational agent. In this setting, the user interacts with an agent using speech and head gestures. The agent uses recognized head nods (or head shakes) and agreement utterances in the user's speech to determine when the user is in agreement (or disagreement). In unconstrained

Classifier	Seed	Co-training	Oracle
Audio	88.4 ± 9.9	91.7 ± 9.2 (p=0.03)	95.1 ± 5.4 (p<0.01)
Visual	95.5 ± 4.4	96.8 ± 3.6 (p=0.07)	97.5 ± 2.8 (p<0.01)

Table 1: Co-training of multimodal agreement classifiers. Each column shows the mean CCR over 15 test subjects, ± the standard deviation. The p-value comparing the performance of the seed and co-trained classifiers, and the seed and oracle classifiers is also displayed.

speech, there are a variety of utterances that can signify agreement, making recognition of agreement difficult with user-independent classifiers, as agreement utterances may vary per user. In this paper, we focus on classifying "yes" and "no" utterances and nod and shake head gestures, and seek to improve these classifiers using unlabeled data.

5.1.1 Dataset

For our experiments on agreement recognition, we collected a database of 15 subjects interacting with a virtual avatar. In each interaction, the subject was presented with a set of 103 yes/no questions and was asked to respond with simultaneous speech and head gesture, and to use only "yes" and "no" responses along with head nods and shakes. Each interaction was recorded with a monocular video camera and lasted 10-12 minutes. A log file with the start and end times of each spoken utterance from the avatar was kept. During each interaction, a remote keyboard was used by the experimenter to trigger the dialogue manager after each subject's response. The end times of the subject's answers were also logged. The video sequences were then post-processed using the avatar's log file to extract the responses of each subject. The sequences were manually labeled to identify positive and negative responses and answers where subjects used extraneous speech or did not speak and gesture at the same time were discarded. To keep the responses to roughly the same length, any responses longer than 6 seconds were also discarded. The resulting data set consisted of 1468 agreement and disagreement audio-visual sequences.

To extract features for the visual classifiers we used a modified version of the 6-degree of freedom head tracker in [10] modified to perform monocular tracking. This tracker was used to compute 3D head rotation velocities for each subject. For each answer segment we applied a 2-second, 64-sample, windowed fast-Fourier transform (FFT) to the x, y and z head rotation velocities computed at 0.1 second intervals within the segment. The x, y and z frequency responses at each time window were then concatenated into a single 99-dimensional observation vector. For the audio agreement classifiers, we used 13 dimensional Mel-frequency cepstral coefficients (MFCCs) computed at 100Hz from the audio of each answer segment.

5.1.2 Results

In this section we present our experiments on co-training speech and gesture agreement classifiers. We first present results on co-training and then demonstrate our co-adaptation technique.

To begin we evaluate co-training for the construction of a user-dependent agreement classifier from a few labeled examples. For this experiment, we use Gaussian speech and gesture classifiers (1-state HMMs with 1 mixture component). We evaluated the co-training algorithm using leave-n out cross-validation on each subject, where we split the data of each subject into 90 percent train and 10 percent test for each round of cross-validation. At each round the training data is split into an unlabeled training set and a labeled seed set of 3 positive and 3 negative examples. To remove bias due to a particular choice of seed set or unlabeled train and test set, co-training was evaluated over three cross-validation trials for each subject where the seed set as well as unlabeled train and test sets were chosen at random.

Table 1 displays the result of the agreement co-training experiment with $N = 4$ and iterating until all the unlabeled training data is labeled with co-training (see Algorithm 1). The table displays the average classification accuracy, averaged across all 15 subjects and three trials. In the table the performance of the co-trained audio and visual Gaussian classifiers are also compared to oracle performance, obtained by training the audio and visual agreement classifiers using a fully supervised paradigm, i.e., with ground truth labels on all the training data, and evaluating these classifiers on the test set. The table also gives the p-values of the difference in classifier performance before and after co-training computed using statistical t-tests. Through co-training we were able to increase overall performance of the audio classifier by 3.3 percent with a p-value of p=0.03, meaning that this increase is statistically significant. Similarly, we were able to gain a marginally significant increase in the performance of the visual classifier by 1.3 percent with a p-value of p=0.07.

Next we evaluate our co-adaptation algorithm. For this experiment, we used 5-state HMMs with 2 mixture components, and ran our co-adaptation algorithm with $M = 4$, $N = 4$ and 3 iterations. We performed leave-one out experiments where we trained user-independent audio and visual classifiers on 14 out of the 15 subjects in our dataset an ran co-adaption on the left out subject. For each subject we ran co-adaptation on random splits of the data, 90 percent train and 10 percent test, and averaged the results over 10 trials. Figure 1 displays the classification accuracy of the user-independent and user-dependent audio and visual classifiers obtained with co-adaptation. The user-dependent HMM classifiers obtained with co-adaptation either matched or improved performance over the user-independent classifiers. As was the case in our previous experiment the main improvement of co-adaptation is seen in the audio modality as the user-independent visual classifiers are already performing quite well on each subject.

Table 2 displays the average classification accuracy of the user-independent and user-dependent classifiers obtained with co-adaptation, averaged over the 15 subjects. The user-dependent audio classifiers obtained with co-adaption do significantly better than the user-independent models, with an average improvement of 4.4 percent and a p-value of 0.023. In Table 2 we also compare our co-adaptation algorithm to single-modality bootstrapping with $M = 10$, $N = 10$ and 3 iterations, and found that unlike our approach the difference in performance between the user-independent and

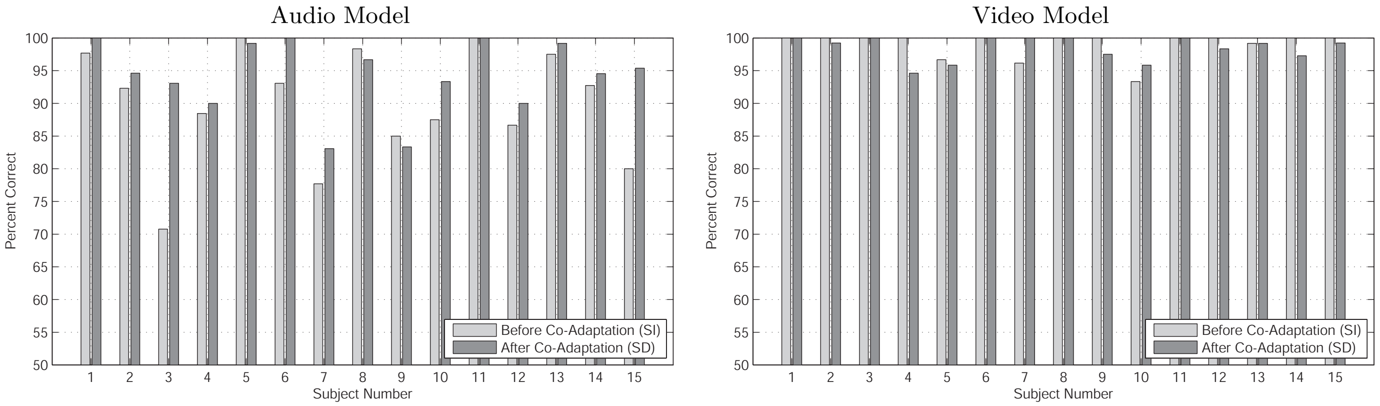


Figure 1: Detailed results for co-adaptation of multimodal agreement classifiers (summarized in Table 2). The CCR rate of the user-dependent and user-independent classifiers are shown for each of the 15 test subjects. The light bars show the CCR of the user-independent classifiers and the dark bars show the CCR of the user-dependent classifiers found with co-adaptation.

Classifier	User Independent	Co-Adaptation	Single-Modality Bootstrap
Audio	89.8 ± 8.8	94.2 ± 5.6 (p=0.023)	91.3 ± 8.7 (p=0.414)
Visual	99.0 ± 2.0	98.5 ± 1.8 (p=0.332)	98.5 ± 2.3 (p=0.411)

Table 2: Co-adaptation of multimodal agreement classifiers. Each column shows the mean CCR over the 15 test subjects, ± the standard deviation. The p-value comparing the performance of each method to that of the user-independent model is also shown.

user-dependent audio HMM classifiers obtained with single-modality bootstrapping was not significant (p-value equal to 0.414). This is because co-adaptation, unlike single modality bootstrapping, was able to leverage the good performance of the visual classifiers to significantly improve the performance of the audio agreement classifier.

5.2 Audio-Visual Speech Classification

Audio-visual speech unit classification uses acoustic features extracted from the speech waveform and image features extracted from the speaker's lip region. It has been widely reported that visual input helps automatic speech recognition in the presence of acoustic noise (e.g. [5]). However, while recording audio-visual speech data is becoming easier, annotating it is still time-consuming. Therefore, we would like to see whether co-training can help exploit unlabeled data for this task.

To satisfy the sufficiency assumption, it should be possible to distinguish between the speech units using only lipreading. This is possible if, for example, the units are digits recognized as whole words: "one", "two", etc. In this paper, we evaluate our algorithm on the task of phoneme unit classification. To ensure sufficiency, we clustered several phonemes together so that the resulting "visemes" are visually distinguishable:

1: b, p, m, f, v

2: w, uw, oy, ao, ow, r

3: sh, zh, ch, jh, s, z

4: ae, aw, ay, ey, aa

5.2.1 Dataset

For evaluation, we used a subset of the multi-speaker audio-visual database of continuous English speech called AVTIMIT [5]. The database contains synchronized audio and video of 235 speakers reading phonetically balanced TIMIT sentences in a quiet office environment. There are 15 sentences per speaker, so the number of sequences in the dataset is between 20 and 60 per viseme, per speaker. To simulate noisy acoustic conditions, speech babble noise was added to the clean audio to achieve a 0 db signal-to-noise ratio. The result is similar to a noisy public place, such as a busy coffee shop. The database contained phonetic transcriptions produced by forced alignment, which we converted to viseme labels via the mapping shown in the previous section. Since the original database was labeled, we simulated unlabeled data sets omiting the labels.

For each label, the data sample consisted of a sequence of acoustic observations and a corresponding sequence of visual observations. The 42-dimensional acoustic feature vector, sampled at 100 Hz, contained 14 mel-frequency cepstral coefficients (MFCCs), their derivatives and double derivatives. Visual features were extracted from a 32-by-32 region centered on the lips, and consisted of an 8-by-8 subgrid of the discrete cosine transform (DCT) followed by a PCA transform to further reduce the dimensionality to 30 coefficients.

5.2.2 Results

In all of the following experiments, the number of HMM states was set such that the average sequence contained three frames per state, resulting in 3-4 states for the audio HMM and 1 state for the visual HMM. The number of Gaussian mixture components was set such that there was a minimum number of training samples per dimension for each component, up to a maximum of 20 components.

Classifier	Supervised	Co-training	Single-modality Bootstrap	Oracle
Audio	59.1 ± 5.6	67.0 ± 9.1 (p<<.01)	60.9 ± 7.7 (p=.10)	94.0 ± 1.1
Video	56.8 ± 10.5	66.2 ± 10.2 (p<<.01)	54.8 ± 12.2 (p=.10)	73.3 ± 4.5

Table 3: Co-training results on the speech dataset. Each column shows the mean CCR over 39 test speakers, ± the standard deviation. "Supervised" refers to the seed classifier performance. In parentheses, we show p-values for co-training and the single-modality bootstrap baseline relative to the supervised classifier.

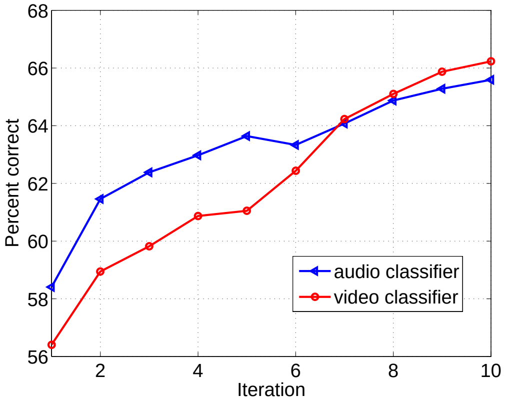


Figure 2: Learning rate of the co-training algorithm on the speech dataset. The plot shows the CCR after each iteration for the audio and video classifiers. The first iteration corresponds to the CCR of the seed classifier.

Our first goal is to show that we can improve classifiers that are poorly trained because of the lack of labeled training data by co-training them on unlabeled data. We thus look for the case then the amount of labeled data is too small, i.e., when adding more training data reduces the test error rate. For the speech dataset, this happens when the labeled set L contains 4 sample sequences per class. First, we train the supervised HMM classifiers on a randomly chosen L for each user, and test them on the remaining sequences. The results, averaged over all users, are shown in the first column of Table 3. Next, we co-train these initial classifiers, using N=4, M=2 and 9 iterations (after which the unlabeled set became depleted.) The results, in the second column of Table 3, show that co-training is able to signifícanly improve the performance in each modality, unlike single-modality bootstrapping (shown in the third column). For reference, the last column shows oracle performance, or what we would get if all of the labels added by co-training were correct. Note that, while the co-trained video classifier is approaching oracle performance, the audio is still far below that level. However, this dataset did not contain a lot of data per speaker. Perhaps, if more unlabeled data were available, the performance would continue to increase, following the trend shown in Figure 2.

Our second goal is to use our adaptive CT algorithm to improve existing user-independent (UI) models when new, unlabeled data becomes available. We train the initial UI audio and visual HMM classifiers on a large labeled dataset consisting of 50 users and approximately 20K samples. Then, for each of the users in the unlabeled dataset, we perform co-adaptation as described in Section 4, using N=25% of all samples, M=2, and 7 iterations. The UI and the final user-dependent (UD) co-trained classifiers are then tested on all of the data for each user.

First, we evaluate the case where the audio noise level in the labeled data matches the noise level in the unlabeled data. In this case, we are mostly adapting to the user. The results are shown in Table 4. The first observation is that the UI video classifier does not do much better than the UD supervised classifier (first column of Table 3.) Our co-adaptation algorithm improves the visual performance significantly, while the audio performance stays the same. One explanation for this is that audio is helping the video as the stronger of the two modalities. Therefore, we compare this to bootstrapping from the stronger audio modality (see "Audio-Bootstrap" in Table 4), and see that it has similar results, doing a little better on video, but a little worse on audio. However, bootstrapping from the video modality does much worse, actually degrading the audio classifier's performance.

Since it is usually not known what level of noise the system will encounter during its deployment, the labeled data collected for training the user-independent models is often clean. However, the case when the test environment is noisier than the training data is precisely when visual input helps the most. Therefore, a compelling application of our algorithm would be to adapt not only to a new user, but to noise in the audio. We repeat the previous experiment, but with UI audio models trained on clean data. The results are shown in Table 5. In this case, it is the audio modality that is "weaker", judging from the UI performance in the first column. This time, co-adaptation improves both modalities: the visual from 59.8% to 69.0%, and the audio from 52.8% to 69.9%. On the other hand, bootstrapping from either the video or the audio modalities does worse, with the latter significantly degrading UI visual performance. Finally, the last column shows that single-modality bootstrapping does worse than co-adaptation. The detailed CCR results obtained before and after co-training for each user are shown as bars in Figure 3. In most cases, our algorithm either improves the UI model performance (by as much as 134% in the case of user 8's visual model), or does not make it worse.

6. CONCLUSIONS

In this paper, we investigated the multi-view semi-supervised co-training algorithm as a means of utilizing unlabeled data in multimodal HCI learning problems. Intuitively, the method uses single-modality classifiers to help train each other by iteratively adding high-confidence labels to the common training set. We extended the confidence-based co-training method to HMM classifiers, and showed that it not only learns user-specific speech and gesture classifiers using just a few labeled examples, but it is more accurate than

Classifier	User Independent	Co-Adaptation	Audio-Bootstrap	Video-Bootstrap
Audio	72.6 ± 4.5	72.0 ± 4.4 (p=.36)	70.2 ± 4.2 (p<<.01)	63.3 ± 11.8
Video	59.8 ± 11.3	68.1 ± 9.7 (p<<.01)	70.1 ± 6.2 (p<<.01)	62.4 ± 13.2

Table 4: User-adaptive co-training results on the speech data, matched labeled and unlabeled audio noise conditions. Each column shows the mean CCR over 39 test speakers, ± the standard deviation. p-values are relative to the UI classifier.

single-modality baselines. We also proposed an adaptive co-training algorithm, co-adaptation, and showed that it can be used to improve upon existing models trained on a large amount of labeled data when a small amount of unlabeled data from new users or noise conditions becomes available. When either the audio or the visual classifier is more accurate, our method performs as well as bootstrapping from the stronger modality, however, it does not require such knowledge. When both modalities are weak, such as when the user-independent audio speech classifiers are trained on clean audio, but the new condition is noisy, our method improves significantly over single-modality baselines. Interesting avenues of future work include the investigation of sufficiency, the use of co-adaptation to perform high-level adaptation of audio-visual classifiers (e.g., adapting their langauge model), the use of user-dependent observations and the use of HMM adaptation techniques (MLLR, MAP) in our algorithm.

7. REFERENCES

[1] S. Bickel and T. Scheffer. Estimation of mixture models using co-em. In *Proceedings of ICML Workshop on Learning with Multiple Views*, 2005.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 92–100, 1998.

[3] M. Collins and Y. Singer. Unsupervised models for named entity classification, 1999.

[4] B. Efron and R. Tibshirani. *An Introduction to the Boot-strap.* Chapman and Hall, 1993.

[5] T. J. Hazen, K. Saenko, C. H. La, and J. Glass. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proc. ICMI*, 2005.

[6] D. Hillard, M. Ostendorf, and E. Shriberg. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *HLT*, 2003.

[7] J. Huang, E. Marcheret, and K. Visweswariah. Rapid feature space speaker adaptation for multi-stream hmm-based audio-visual speech recognition. In *ICME*, 2005.

[8] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using cotraining, 2003.

[9] T. Li and M. Ogihara. Semi-supervised learning from different information sources. *Knowledge Information Systems Journal*, 7(3):289–309, 2005.

[10] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 803–810, 2003.

[11] S. Pan, S. Shen, M. X. Zhou, and K. Houck. Two-way adaptation for robust input interpretation in practical multimodal conversation systems. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, New York, NY, USA, 2005. ACM Press.

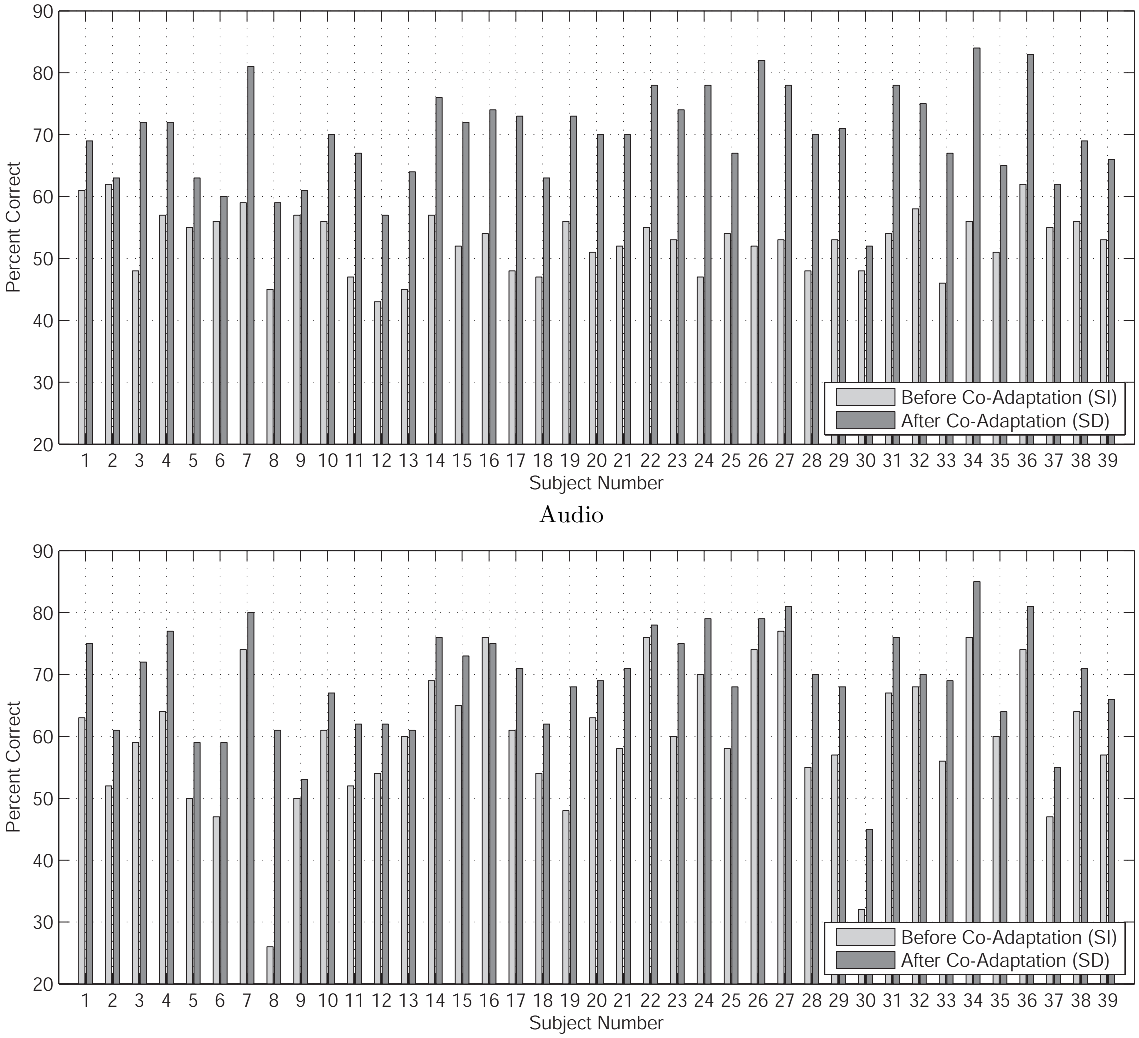
[12] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views*, 2005.

[13] B. Xiao, R. Lunsford, R. Coulston, M. Wesson, and S. L. Oviatt. Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. In *Proceedings of International Conference on Multimodal Interfaces*, 2003.

[14] R. Yan and M. Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *Computer Vision and Pattern Recognition*, pages 657–663, June 2005.

Classifier	User Independent	Co-Adaptation	Audio-Bootstrap	Video-Bootstrap	Single-modality Bootstrap
Audio	52.8 ± 4.8	69.9 ± 7.4 (p<<.01)	55.4 ± 4.5	63.3 ± 11.8	58.6 ± 4.4 (p<<.01)
Video	59.8 ± 11.3	69.0 ± 8.6 (p<<.01)	51.5 ± 7.9	62.4 ± 13.2	60.7 ± 12.1 (p=.03)

Table 5: Co-adaptation results on the speech data, mis-matched audio noise conditions. Each column shows the mean CCR over 39 test speakers, ± the standard deviation. p-values are relative to the UI classifier.



Audio

Video

Figure 3: Detailed co-adaptaion results for mismatched audio noise (summarized column 2 of in Table 5) for each of the 39 test speakers. The light bars show the UI models' CCR, the dark bars show CCR after co-adaptation.