

# Active Image Labeling and Its Application to Facial Action Labeling

Lei Zhang<sup>1</sup>, Yan Tong<sup>2</sup>, and Qiang Ji<sup>1</sup>

<sup>1</sup> Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute

<sup>2</sup> Visualization and Computer Vision Lab, GE Global Research Center

zhangl2@rpi.edu, tongyan@research.ge.com, qji@ecse.rpi.edu

**Abstract.** For many tasks in computer vision, it is very important to produce the groundtruth data. At present, this is mostly done manually. Manual data labeling is labor-intensive and prone to the human errors. The training data it produces often lacks in both quantity and quality. Fully automatic data labeling, on the other hand, is not feasible and reliable. In this paper, we propose an interactive image labeling technique for efficient and accurate data labeling.

The proposed technique includes two parts: an automatic labeling part and a human intervention part. Constructed on a Bayesian Network, the automatic image labeler produces an initial labeling of the image. A person then examines the initial labeling and makes some minor corrections. The selected human corrections and the image measurements are then integrated by the Bayesian Network framework to produce a refined labeling. To minimize the human involvement, an active user feedback strategy is developed, through which the optimal user feedback is determined, so that the labeling errors in the subsequent re-labeling process can be maximally reduced. The proposed framework combines the advantages of the human input with those of the machine so that the reliable, accurate, and efficient data labeling can be achieved. We demonstrate the validity of the proposed framework for interactive labeling of facial action units. The proposed methodology, however, is not limited to labeling of facial action units. It can be easily extended to other areas such as interactive image segmentation.













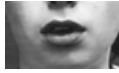
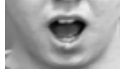
## 1 Introduction

For a variety of classification tasks in computer vision including image segmentation, feature detection, and object recognition, image labeling is needed in order to create the training data for the classifiers. Producing image groundtruth is typically carried out manually. However, manual image labeling is labor-intensive, and with limited throughput. In addition, image labeling is an error-prone process due to various reasons, such as the labeler's errors or the imperfect description of classes. Thus, a second (or more) pass(es) of labeling often by a different human labeler is usually required in order to fix the errors and inconsistencies in the earlier pass(es). To alleviate these problems, various interactive or semi-automatic learning tools have been used. For example, Levin et al.[1] used co-training in the context of visual car detection to improve the accuracy of a classifier using the unlabeled examples. The active learning is another machine learning technique to address these problems. Active learning techniques are sequential learning methods that are designed to reduce the manual training costs in achieving adequate

learning performance. It is being used increasingly to help reduce the amount of labeled training data by incorporating limited user feedback selectively and intelligently during training. In this paper, we will introduce a framework based on Bayesian Network for active labeling and demonstrate its use for facial action unit labeling.

Facial expression recognition represents an active research area in computer vision. Current research in facial expression recognition can be classified into two categories: facial expression recognition and facial action recognition. While the former focuses on recognizing the global facial patterns, the latter concentrates on recognizing the local facial motions. Since there are many different global facial expressions in real life, the current research in global facial expression recognition studies only a few prototype expressions (happiness, disgust, surprise, etc.). Local facial motion recognition focuses on recognizing the local facial actions as defined in Facial Action Coding System (FACS) [2]. FACS defines 44 facial Action Units (AUs) and 8 head pose action units at 5 asymmetric intensity levels, co-occurring in various combinations. Each AU represents a kind of facial muscular activity that produces facial appearance changes and is anatomically related to the contraction of a specific set of facial muscles. The FACS has been demonstrated to be a powerful means for detecting and measuring a large number of facial displays by observing a small set of visually discernable facial muscular movements [3]. Table 1 shows some commonly occurring AUs and their interpretations.

**Table 1.** A list of commonly occurring action units and their interpretations [4].

AU1 	AU2 	AU4 	AU5 	AU6 
Inner brow raiser	Outer brow raiser	Brow Lowerer	Upper lid raiser	Cheek raiser
AU7 	AU9 	AU12 	AU15 	AU17 
Lid tighten	Nose wrinkle	Lip corner puller	Lip corner depressor	Chin raiser
AU23 	AU24 	AU25 	AU27 	
Lip tighten	Lip presser	Lips part	Mouth stretch	

Compared with recognizing the global facial expressions, recognizing the local facial actions is expression-independent, can capture subtle facial changes, and is free of human emotional judgment/interpretation. In addition, from the recognized local facial actions, it is possible to perform various high level understanding such as facial expression recognition, human behavior analysis, cognitive state recognition, etc.

For facial action recognition, AU labeling is required to create the training data. Manual AU labeling of facial expression, in particular spontaneous expressions, is very time consuming, error prone, and expensive. It usually requires years of training be-

fore a person can become a certified AU coder. In addition, the inter-variation among AU coders is often large. AU labeling becomes even more difficult when multiple AUs co-occur. Given these difficulties, recent researches have focused on developing automatic AU coding systems [3]. Despite much progress in automatic AU recognition, the current automatic AU recognition systems are limited mainly to recognizing AUs from the posed expressions and with mostly the frontal faces. They remain ineffective for recognizing AUs from the spontaneous facial expressions due to in part the subtlety of facial actions in the spontaneous expressions and in part the face occlusion/distortion from head movements.

To further push researches in spontaneous facial expression recognition, it is critical to have both sufficient and reliable AU-labeled facial expressions since supervised training of automatic facial action recognition systems require the groundtruth of AU labels. To achieve this goal, we propose an interactive AU labeling system that combines automatic AU measurements with limited yet selective human input to perform both accurate and fast AU labeling. We construct a unified probabilistic framework based on a Bayesian Network to incorporate both the automatic AU measurements and the human input. Furthermore, we also propose an active human intervention strategy based on the mutual information calculated from BN inference. This active human intervention strategy can provide the suggestion for the next human input so that the person can focus on certain errors with priority. It is potential to reduce the whole human involvement when these errors are first corrected. In addition, this framework can handle the uncertainty of the AU measurement by the automatic labeling technique and that of the human labeling. Compared with other interactive image labeling system, ours enjoys the following advantages: 1) incrementally allow the human input (from one or multiple human experts) to be incorporated at any stage; 2) systematically combine the human input with the automatic AU measurements, and 3) allow to determine the optimal human interaction so that human involvement can be kept to the minimum level.

## 2 Related Work

From the machine learning perspective, semi-automatic image labeling is a problem of semi-supervised active learning [5], which combines semi-supervised learning [6–12] with active learning [13–19]. The semi-supervised learning aims to design image classifier by making use of the unlabeled data, and is useful especially when the training data is limited; while the active learning mechanism aims to enlarge the useful information conveyed by the human feedback, and provides the annotators the most informative samples according to the current image classifier. One commonly used semi-supervised learning technique is the co-training technique [20, 21, 1], which aims to improve classifiers' learned information of the labeled data by maximizing their agreement on the unlabeled data.

The active learning proceeds sequentially, with the learning algorithm actively asking for the labels (categories) of some instances from a teacher (also referred to as membership queries). The objective is to ask the teacher to label the most informative instances in order to reduce the total labeling costs and accelerate the learning. For active learning, the key is the sample selection strategy, which, based on a prede-

finer criterion, selects the next samples for the user to label. A better criterion means the corresponding sample will provide a larger amount of performance enhancement after being manually labeled. For classification, the most commonly used criterion is the close-to-boundary criterion. Other criteria have also been proposed. The expected information gain is a natural measure to select the next sample whose label is going to be asked for. However, it is not sufficient for guaranteeing a large reduction in the expected prediction error. Freund et al.[22] enhance the Query by Committee method to get high expected information gain. Shen et al.[23] incorporate multi-criteria (informativeness, representativeness and diversity) for active learning. Besides using the distance from the classification boundary, Kapoor [24] propose the variance and uncertainty of a Gaussian Process for object categorization as the active learning criteria. Raghavaan et al.[25] improves active learning by combining the instance level feedback with the feature level feedback.

In speech recognition, the concept of active labeling is introduced to minimize the number of mislabelings by combining the output of machine labeling results with the human feedback [26]. Specifically, active labeling aims to minimize the number of utterances to be checked again by first automatically selecting the ones that are likely to be erroneous or inconsistent with the previously labeled examples. The user feedback is then requested to correct the selected erroneous labeling. The corrected labels are then feedback to the automatic labeler to reclassify the utterances. Despite its use in speech recognition, active labeling has not yet received due attention in computer vision.

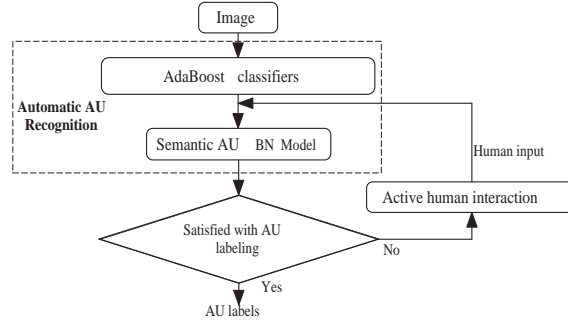
Although some human interventions are usually required for most of the AU recognition systems to achieve an accurate and reliable face detection and/or facial feature motion extraction for AU recognition [27], AU labeling is still performed either automatically or manually. In contrast, our interactive AU labeling system combines the advantage of automatic AU measurements (i.e., efficiency and objectivity) with the advantage of human labeling (i.e., reliability and accuracy) to perform both reliable and efficient AU labeling.

### 3 The Interactive AU Annotation System

#### 3.1 Overview of the Image Labeling Method

Considering various shortcomings with the manual AU labeling, we propose to build an interactive AU annotation system by which the human annotators can interact with an automatic AU recognition system in labeling AUs, especially for spontaneous facial expressions. Figure 1 outlines the flowchart of the proposed system. Given an image, the automatic AU recognition system first performs an initial labeling of AUs in the image. Based on the result of the automatic labeling, the human needs to determine which AU (or which sets of AUs) should be changed next in order to reduce the involvement in future. To achieve this, we employ an active human involvement strategy, by which we actively decide the optimal human intervention in order to maximize its effectiveness. Given the identified AU, a human can then manually set its state. This completes the human input stage. The automatic AU recognition system then systematically and incrementally combines the current human input with the existing knowledge of AUs (e.g. AU measurements and the previous human inputs) to correct the mislabeled AUs

through a belief propagation. This process repeats until the human is satisfied with the AU labeling. The system can incrementally combine the image data and the human input to produce progressively improved AU labeling. In addition, using this active human labeling scheme, a human can focus only on the most important AUs and, therefore, can significantly reduce the human involvement. In this way, both the labeling speed and accuracy are improved.



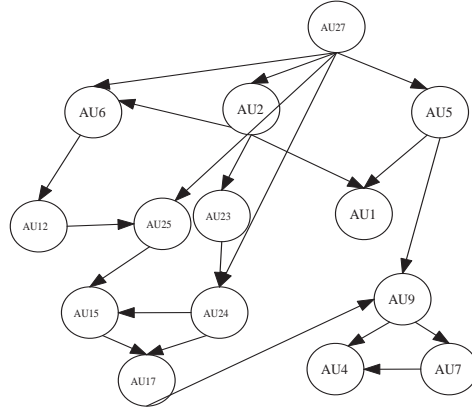
**Fig. 1.** The flowchart of the proposed interactive AU labeling system.

### 3.2 Automatic AU Recognition

To perform an effective interactive labeling, we have developed an automatic AU recognition system. The system consists of two major components as shown in Figure 1. The first component is an image-based AU classification system based on Adaboost. The second component is a probabilistic AU model that captures the spatial and semantic dependencies among AUs. The final AU recognition results are based on combining the AdaBoost classification results with the AU model. In the paragraphs to follow, we summarize our automatic AU recognition system.

**AU Classification Using AdaBoost** We first perform the face and eye detection from images. Given the knowledge of eye centers, the face region is normalized and convolved pixel by pixel by a set of multiscale and multiorientation Gabor filters. Then, the measurement for each AU is obtained through a computer vision technique based on Gabor wavelet features and AdaBoost classifiers similar to [28]. However, this frame-by-frame AdaBoost classification method is susceptible to image noise and inaccurate image alignment. In addition, it cannot perform effectively when multiple AUs co-occur or when facial deformations are subtle, which happens frequently for spontaneous facial expressions. Therefore, besides measuring each AU individually, it is more important to incorporate other related prior knowledge to help improve AU recognition under these difficult conditions.

**AU Modeling with A Bayesian Network** Based on the previous study by Tong et al.[29], there are semantic relationships (the co-occurrence relationships and the mutually exclusive relationships) among AUs. These relationships exist due to either certain facial expressions or the underlying facial anatomy and physiology. Specifically, some AU combinations are physiologically impossible to occur together, while other AUs co-occur mainly because of certain facial expressions. For example, AU6 (cheek raiser) tends to happen with AU12 (lip corner puller)when smiling. On the other hand, AU25 (lips part) can hardly happen with AU24 (lip presser) simultaneously. We will exploit these relationships for automatic AU labeling. In particular, following the work by Tong et al.[29], we use a Bayesian Network (BN), as shown in Figure 2, to capture the co-occurrence and the mutually exclusive relationships among AUs. A BN is a directed acyclic graph (DAG) that represents a joint probability distribution among a set of variables, where the nodes represent the variables and the links among those nodes represent the conditional dependency among the variables.

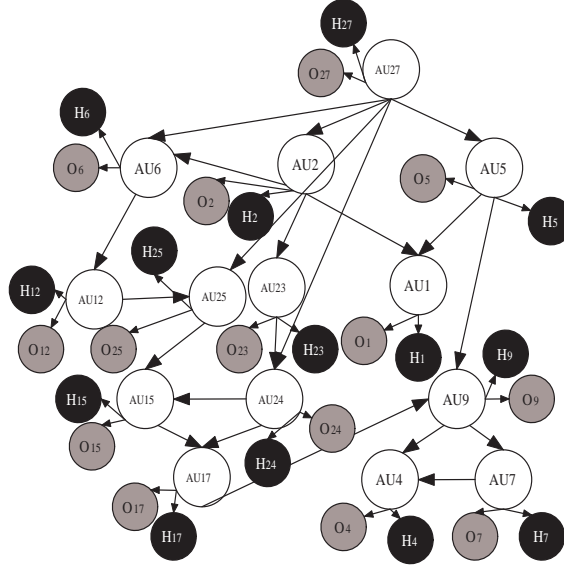


**Fig. 2.** A BN model for modeling the semantic relationships among AUs. Each white circle represents an AU, and the link between two white circles represents the semantic relationship between the two AUs.

Specifically, the nodes represent AUs and the links between the nodes capture the relationships (co-occurrence or mutual exclusiveness) among AUs. The model can be constructed and parameterized using the training data with a standard learning technique [30]. We employ this BN model as our base learning algorithm for their effectiveness in many learning tasks. Compared with other learning methods such as Support Vector Machine and Neural Network, BN has the capabilities of incorporating different types of human inputs through its hierarchical structure, incorporating the human input at any stage of the labeling, systematically combining the human inputs with the existing data, and exerting the impacts of human inputs on other entities through belief propagation.

To integrate the AU measurements and the user feedback, we extend the BN model in Figure 2 for interactive labeling as shown in Figure 3. Specifically, we associate each

AU node  $AU_i$  with two measurement nodes, which are represented by a grey circle and a black circle, respectively. The grey circle  $O_i$  represents the AU measurement obtained through the computer vision technique (i.e., AdaBoost classifier in this work), and the link between  $AU_i$  and  $O_i$  represents the measurement uncertainty resulting from the computer vision technique. This uncertainty is obtained from analyzing the performance (accuracy) of the AdaBoost classifier for each AU. The black circle  $H_i$  represents the human input for this AU, and the link between  $AU_i$  and  $H_i$  represents the reliability and confidence of the human labeling. The confidence is determined by the person who is giving the human input. For a domain expert who is very sure about his/her judgment, the confidence is set as 1. For other persons, this confidence is set depending on the level of his/her expertise. In this way, the BN model systematically combines the objective image measurements of AUs, the subjective human labeling, and their uncertainties.



**Fig. 3.** An extended BN model for interactive labeling. The white circle represents each AU node that should be labeled. The grey circle represents the measurement for each AU node, which is obtained using the AdaBoost classifier. The black circle represents the human labeling for each AU node, which is obtained through interactive labeling.

Given the model structure shown in Figure 2, we need to define the states for each node and learn the model parameters associated with each node. Each AU node has two states for representing its "presence" and "absence" states. Its parameter is characterized by the Conditional Probabilistic Tables  $p(AU_i|pa(AU_i))$ , where  $pa(AU_i)$  denotes the parent configuration of  $AU_i$ . Given the training data (AU labels), these parameters can be learned [31]. For each measurement node, the parameters  $p(O_i|AU_i)$  or

$p(H_i|AU_i)$  are assigned manually based on the recognition accuracy of the computer vision technique or human confidence, respectively. Based on the BN model, the states of the AU nodes are inferred using the junction tree inference approach (see Section 3.3 for more details).

**Active Human Input Determination** Given the AU label assignment by the automatic AU recognition, we need decide 1) which AU is mislabeled, and 2) among the mislabeled AUs, which AU's state shall be corrected next. An important observation for AU labeling is that it is not straightforward for a human to quickly identify which AU is mislabeled. Identifying the mislabeled AUs requires both time and expertise. It is even more difficult to decide which AU shall be corrected next in order to minimize the subsequent user feedback. To achieve this goal, it is important to identify the AU not based on its obvious mistake, but based on its potential to maximally correct other mislabeled AUs. We employ the active user feedback strategy for this purpose. The idea of the active user feedback is to select the mislabeled AUs to solicit the user's feedback so as to achieve the maximal reduction in the overall estimation uncertainties and minimize the misclassification by the BN model after incorporating the user's feedback. The most informative AU is the one that maximally reduces the overall estimation uncertainty.

For this purpose, we need develop a measure to quantify the utility of each AU. This measure reflects the estimated contribution of a certain AU to the reduction of uncertainty in the subsequent labeling process when its label is manually corrected. The mutual information between the overall model uncertainty and the user feedback is one of such measures. It measures the potential of the user feedback in reducing the overall uncertainty of the model. Mutual information is found to be quite effective in the relevancy feedback for image retrieval and is easy and quick to compute. Specifically, let  $H_k$  be the human labeling of the  $k^{th}$  AU, and let  $AU_1, \dots, AU_N$  be all the AU nodes in the BN model, the mutual information between  $H_k$  and  $AU_1, \dots, AU_N$  can be computed as follows:

$$I(AU_1, \dots, AU_N; H_k) = \sum_{AU_1, \dots, AU_N} p(AU_1, \dots, AU_N, H_k) \log \frac{p(AU_1, \dots, AU_N, H_k)}{p(AU_1, \dots, AU_N)} \quad (1)$$

We use this mutual information measure to return a ranked list of AUs sorted in a decreasing order of importance. A person can then use this list to guide the selection of the next AU for correction.

### 3.3 Interactive AU Labeling through Probabilistic Inference

Given the BN model (Figure 3) for interactive AU labeling, the AU labeling is performed through a probabilistic inference by maximizing the joint probability of all AUs given their image measurements and all available human inputs, i.e.

$$AU_1^*, \dots, AU_N^* = \underset{AU_1, \dots, AU_N}{\operatorname{argmax}} p(AU_1, \dots, AU_N | O_1, \dots, O_N, H_1, \dots, H_K) \quad (2)$$

where  $AU_1, \dots, AU_N$  represent all the AUs that should be labeled,  $O_1, \dots, O_N$  represent all the image measurements that are obtained through the computer vision technique, and  $H_1, \dots, H_K$  represent all the available human inputs for the AUs. Based on



the conditional independence in BN, the joint probability can be factorized as follows:

$$p(AU_1, \dots, AU_N | O_1, \dots, O_N, H_1, \dots, H_K) \propto \quad (3)$$

$$\prod_{j=1}^N p(AU_j | pa(AU_j)) \prod_{j=1}^N p(O_j | AU_j) \prod_{k=1}^K p(H_k | pa(H_k))$$

where  $pa(AU_j)$  represents the parent configuration of  $AU_j$  in the BN model and  $pa(H_k)$  represents the  $k^{th}$  AU with the human input.

In this way, the AUs can be labeled semi-automatically and incrementally by combining the subjective knowledge from the human coder and the objective image measurements. For efficiency, we only allow one human correction on one AU in each iteration. This single correction on one AU may affect the AU labeling results for all AUs through belief propagation in the BN.

## 4 Experimental Results

To demonstrate our proposed interactive AU labeling system, we perform the AU labeling experiments on two databases. The first database is the Cohn-Kanade DFAT-504 database [4]. This database is collected under controlled illumination and background, and has been widely used for evaluating facial AU recognition system. The Cohn-Kanade database only contains the frontal and posed facial expressions. On the other hand, it is also very important to study the labeling for spontaneous facial expressions. We therefore also evaluate our system on the second database, i.e., a spontaneous facial expression database. The second database consists of images collected from: (1) Multiple Aspects of Discourse (MAD) research lab at the University of Memphis [32], (2) Belfast natural facial expression database [33], and (3) videos obtained from the website (<http://www.youtube.com/>). In these images, the subjects are displaying various natural facial expressions (such as frown and confusion) with often subtle facial appearance changes. Furthermore, the head poses are not limited to the frontal view in these images. Due to these reasons, the automatic AU recognition from spontaneous facial expression is more challenging.

### 4.1 Interactive AU Labeling on Posed Facial Expressions

We first perform the interactive AU labeling on the posed facial expression images from the Cohn-Kanade database. We divided the database into 8 sections, each of which contains about 1000 images from different subjects. Each time, we randomly choose 7 sections for training the AdaBoost classifier for each AU and the BN model as shown in Figure 3. Then, we randomly select 50 samples from the remaining section for testing. Hence, there are totally 400 samples used for testing the automatic AU labeling, the AU labeling with arbitrary human input, and the AU labeling with active human input.

For the posed facial expression database, besides comparing with the fully automatic AU recognition, we also compare two methods of interactive AU labeling with two different ways of human intervention. In the first method, the human coder arbitrarily chooses a mislabeled AU for correction. The human coders in our experiments are

**Table 2.** Comparison of AU labeling accuracy on Cohn-Kanade database using the automatic AU recognition with only the image measurements (denoted by "BN"), the interactive labeling with the image measurements and the arbitrary human input, and the interactive labeling with the image measurements and the active human input, respectively.

AU label		AU1		AU2		AU4		AU5		AU6		AU7		AU9	
		Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #
Total test samples		57	343	38	362	60	340	20	380	40	360	36	364	15	385
Error rate		FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP
BN		0.210526	0.0831	0.1205	0.0292	0.1826	0.0793	0.1852	0.0643	0.1081	0.0483	0.144737	0.058036	0.125	0.0188
Arbitrary input	iter1	0.017544	0.0219	0	0.0058	0.1167	0.0329	0.15	0.033	0.075	0.0152	0.055556	0.01497	0.066667	0.008
	iter2	0.052632	0.0063	0	0	0.05	0.0066	0.05	0.011	0.05	0.0061	0.027778	0.005988	0.066667	0.0027
	iter3	0	0	0	0	0.05	0.0033	0	0.0027	0	0	0	0.002994	0	0
	iter4	0	0	0	0	0	0	0	0	0	0	0	0.002994	0	0
	iter5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Active input	iter1	0.087719	0.0094	0.1053	0.0116	0.1167	0.0263	0.15	0.0192	0	0.0212	0.055556	0.011976	0.066667	0.0053
	iter2	0.017544	0.0031	0.0526	0	0.05	0	0.1	0.0082	0	0.003	0	0	0	0
	iter3	0.017544	0	0	0	0.0167	0	0.05	0	0.003	0	0	0	0	0
	iter4	0	0	0	0	0.0167	0	0	0	0	0	0	0	0	0
	iter5	0	0	0	0	0	0	0	0	0	0	0	0	0	0

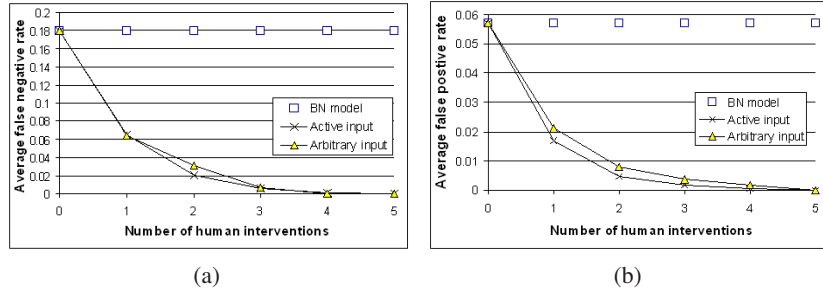
AU label		AU12		AU15		AU17		AU23		AU24		AU25		AU27	
		Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #
Total test samples		64	336	26	374	53	347	23	377	24	376	147	253	30	370
Error rate		FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP
BN		0.1	0.0367	0.2609	0.0741	0.1619	0.0738	0.4091	0.0591	0.2857	0.0696	0.172881	0.089912	0.042857	0.0126
Arbitrary input	iter1	0.046875	0.0193	0.0385	0.0363	0	0.0312	0.1304	0.0162	0.125	0.0384	0.040816	0.022026	0.033333	0
	iter2	0	0.0032	0	0.0056	0	0.0156	0.0435	0.0108	0.0833	0.0219	0.013605	0.017621	0	0
	iter3	0	0.0096	0	0	0	0.0093	0.0435	0.0081	0	0.011	0	0.004405	0	0
	iter4	0	0.0064	0	0	0	0.0093	0	0	0.0055	0	0	0	0	0
	iter5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Active input	iter1	0.015625	0.0193	0.0385	0.0251	0.0189	0.0405	0.1739	0.0135	0	0.0329	0.07483	0	0	0
	iter2	0	0.0032	0	0.0168	0	0.0187	0.0435	0.0054	0	0.0082	0.020408	0	0	0
	iter3	0	0	0	0.0084	0	0.0125	0	0	0	0	0	0	0	0
	iter4	0	0	0	0	0	0.0062	0	0	0	0	0	0	0	0
	iter5	0	0	0	0	0	0	0	0	0	0	0	0	0	0

domain experts with certain familiarity with AU recognition. They are allowed to use their experiences to select the mislabeled AU for correction. The second method uses the proposed active human input strategy to identify the mislabeled AU for correction. In each iteration, only one human intervention is allowed to correct one AU label for both methods. Table 2 and Figure 4 summarize the results for the two user feedback methods.

Table 2 reports the AU labeling accuracy for each AU individually by using different methods. In Table 2, "FN" and "FP" represent the false-negative rate (i.e. the error rate of positive samples) and false-positive rate (i.e. the error rate of negative samples) for each AU, respectively. From Table 2, we can compare how often the active input can drive the individual AU recognition error to zero before the arbitrary input can. The active input performs better in 14 cases, while the arbitrary input performs better in 7 cases. They perform the same in 7 cases. These results demonstrate the superiority of the active input to the arbitrary input. Compared with the arbitrary human input, less human intervention is involved using the proposed active human input strategy. Usually, only two human corrections will be sufficient to obtain the accurate AU labeling for all 14 AUs. It makes the AU labeling much more efficient than the manual labeling and much more accurate than the fully automatic labeling.

Figure 4 reports the average AU labeling performance for all AUs. In Figure 4, we can find that both false-negative rate and false-positive rate decrease significantly with

the help of a few human corrections. For example, the average false-negative rate decreases from 17.9% (automatic AU labeling) to 6.4% and the average false-positive rate also decreases from 5.7% (automatic AU labeling) to 1.7% with only one human correction using the proposed interactive AU labeling method with active human input. The two curves are not separated significantly mainly due to two reasons. First, the automatic AU recognition already achieve very good accuracy. There are not many errors to correct. As a result, the difference of the corrected AU labels in these two methods can only be a small fraction of the total AU samples. Second, the human coders also use their experiences to select the AU to be corrected. They somewhat try to select certain AUs with high priority for correction, too. This implicit use of domain knowledge also reduces the performance difference between two methods.



**Fig. 4.** Average AU labeling performance on Cohn-Kanade database using the automatic AU recognition with only image measurements, the interactive labeling with image measurements and arbitrary human input, and the interactive labeling with image measurements and active user input, respectively. x axis represents the number of human corrections (iterations), while y axis represents the (a) average false-negative rate and (b) average false-positive rate, respectively.

For the AUs that are hard to label using the automatic AU labeling, the improvement is impressive. For example, with one human correction, the false-negative rate of AU15 (lip corner depressor) decreases from 26.09% (automatic AU labeling) to 3.85%, and its false-positive rate also decreases from 7.41% (automatic AU labeling) to 2.51%. The false-negative rate of AU24 (lip presser) decreases from 28.57% (automatic AU labeling) to 0%, and its false-positive rate also decreases from 6.96% (automatic AU labeling) to 3.29% using the proposed interactive AU labeling method with active human input.

## 4.2 Interactive AU Labeling on Spontaneous Facial Expressions

Currently the public available facial expression databases provide a large number of training data for the posed facial expressions, whereas the resource of spontaneous databases is limited. Due to the increasing interest in the spontaneous facial expression recognition, it is more important and desirable to label the AUs for the spontaneous

facial expression. In the second set of experiments, we evaluate the proposed interactive labeling method on a spontaneous facial expression database described previously. Specifically, we use 1300 images for training the automatic AU recognition network, and use different 300 images for testing AU labeling.

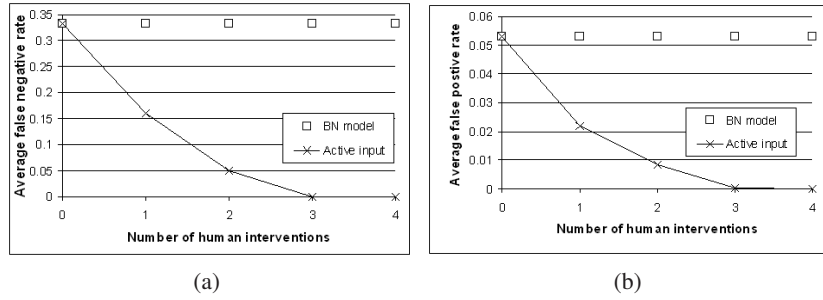
Table 3 and Figure 5 compare the AU labeling accuracies on the spontaneous facial expressions using the automatic labeling using BN inference with only the image measurements, and the interactive labeling using BN inference with both the image measurements and the proposed active human input. Table 3 reports the AU labeling accuracies for 12 AUs that commonly occur in the spontaneous facial expression database. Figure 5 reports the average AU labeling performance for all AUs.

**Table 3.** Comparison of AU labeling accuracies on a spontaneous facial expression database using BN inference with only the image measurements, and the interactive labeling with both the image measurements and the active human input.

AU label	AU1		AU2		AU4		AU5		AU6		AU12	
	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #
Total test samples	74	226	74	226	24	276	51	249	88	212	97	203
Error rate	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP
BN	0.1493	0.0798	0.2239	0.0423	0.5714	0.0474	0.1111	0.0172	0.2361	0.0406	0.297619	0.067708
Active input	Iter1	0	0.028	0.0484	0.014	0.75	0.0039	0	0.0126	0.0882	0	0.060976
	Iter2	0	0	0	0	0.5	0	0	0	0.0147	0	0.012195
	Iter3	0	0	0	0	0	0	0	0	0	0	0
	Iter4	0	0	0	0	0	0	0	0	0	0	0

AU label	AU15		AU17		AU23		AU24		AU25		AU27	
	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #	Pos #	Neg #
Total test samples	19	281	28	272	35	265	15	285	271	29	36	264
Error rate	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP
BN	0.2353	0.0436	0.5263	0.0389	0.6429	0.0158	0.5556	0.0144	0.1635	0.1818	0.285714	0.045977
Active input	Iter1	0	0.054	0	0.0425	0.0909	0	0.625	0.0072	0.0455	0.0909	0.205882
	Iter2	0.0667	0.0252	0	0.0232	0	0	0	0.0038	0.0455	0	0
	Iter3	0	0	0	0.0039	0	0	0	0	0	0	0
	Iter4	0	0	0	0	0	0	0	0	0	0	0



**Fig. 5.** Average AU labeling performance on a spontaneous facial expression database using BN inference given only the image measurements, and the interactive labeling by active human input, respectively. x axis represent the number of human corrections (iterations), while y axis represents the (a) average false-negative rate and (b) average false-positive rate, respectively.

Due to the low intensity, non-additive effect, and individual difference of spontaneous facial action, the purely automatic AU labeling cannot achieve an accurate and reliable AU labeling. However, from Figure 5, we can see that an accurate AU labeling can be achieved with a few human interactions. For example, we can achieve about 5% false-negative rate and 1% false-positive rate with only two corrections. The improvement is especially significant for the AUs that are difficult for the automatic AU labeling. Compared with the automatic AU labeling, the false-negative rate of AU23 (lip tightener) decreases from 64.29% (automatic AU labeling) to 9.09%, and its false-positive rate also decreases from 1.58% (automatic AU labeling) to 0% with one human correction using the proposed interactive AU labeling method with active human input. These results demonstrate the effectiveness and efficiency of the proposed interactive labeling approach.

## 5 Conclusion

In this research, we introduce an interactive image labeling system for effective combination of the automatic image labeling with the limited human labeling. By using a BN as the engine for the automatic image labeling and adopting the active user feedback strategy, the system is effective for image labeling in several aspects: 1) it allows human inputs to be incrementally incorporated at any stage; 2) it allows to combine different human inputs systematically; and 3) it employs an active human input scheme to minimize the amount of human input, thereby improving both labeling efficiency and accuracy. We have applied the framework to the interactive facial action units labeling. The experiments demonstrate the effectiveness of the system in improving both the AU labeling accuracy and the efficiency. Besides, the proposed framework is generic enough to be applied to different tasks in computer vision, including image segmentation, image retrieval, and object recognition, etc..

## References

1. Levin, A., Viola, P., Freund, Y.: Unsupervised improvement of visual detectors using co-training. *Int'l. Conf. on Computer Vision* (2003) 13–16
2. Ekman, P., Friesen, W.V., Hager, J.C.: *Facial Action Coding System: the Manual*. Research Nexus, Div., Network Information Research Corp., Salt Lake City, UT (2002)
3. Pantic, M., Bartlett, M.: Machine analysis of facial expressions. In Delac, K., Grgic, M., eds.: *Face Recognition*. I-Tech Education and Publishing, Vienna, Austria (2007) 377–416
4. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. *Proc. 4th IEEE Int'l Conf. Automatic Face and Gesture Recognition* (2000) 46–53
5. Zhou, Z.H., Chen, K.J., Dai, H.B.: Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Trans. on Information Systems* **24**(2) (2006) 219–244
6. In Chapelle, O., Weston, J., Schölkopf, B., eds.: *Semi-supervised learning*. MIT Press 2006
7. Cozman, F., C.I.C.M.: Semi-supervised learning of mixture models. *ICML* (2003)
8. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. *ICML* (2006)
9. Weston, J., Leslie, C., Zhou, D., Elisseeff, A., Noble, W.S.: Semisupervised protein classification using cluster kernels. In: *Advances in neural information processing systems* 16. MIT Press (2004)

10. Xu, L., Schuurmans, D.: Unsupervised and semi-supervised multi-class support vector machines. *Proc. of the 20th National Conf. on Artificial Intelligence* (2005)
11. Zhou, Z.H., Li, M.: Semi-supervised regression with co-training. *IJCAI* (2005)
12. Zhu, X.: Semi-supervised learning with graphs. Doctoral dissertation, Carnegie Mellon University (2005)
13. Ackley, D.H., Littman, M.L.: Generalization and scaling in reinforcement learning. In Touretzky, D.S., ed.: *Advances in Neural Information Processing Systems 2*. Morgan Kaufmann, San Mateo, CA (1990) 550–557
14. Boyan, J.A., Moore, A.W.: Generalization in reinforcement learning: Safely approximating the value function. In Tesauro, G., Touretzky, D.S., Leen, T.K., eds.: *Advances in Neural Information Processing Systems 7*. The MIT Press, Cambridge, MA (1995)
15. Gullapalli, V.: Reinforcement learning and its application to control. PhD thesis, University of Massachusetts (1992)
16. Lin, L.J.: Reinforcement learning for robots using neural networks. PhD thesis, Carnegie Mellon University (1993)
17. Maes, P., Brooks, R.A.: Learning to coordinate behaviors. *Proc. 8th National Conf. on Artificial Intelligence* (1990) 796–802
18. Zhang, W., Dietterich, T.G.: A reinforcement learning approach to job-shop scheduling. *IJCAI* (1995)
19. Chang, E.Y., Lai, W.C.: Active learning and its scalability for image retrieval. *ICME* (2004) 73–76
20. Christoudias, C., Saenko, K., Morency, L., Darrell, T.: Co-adaptation of audio-visual speech and gesture classifiers. *Int'l. Conf. on Multimodal Interfaces* (2006) 84–91
21. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. *Proc. of the Workshop on Computational Learning Theory* (1998) 92–100
22. Freund, Y., Seung, H., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168 (1997)
23. Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C.L.: Multi-criteria-based active learning for named entity recognition. *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004)
24. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. *ICCV* (2007)
25. Raghavan, H., Madani, O., Jones, R.: Active learning with feedback on both features and instances. *Journal of Machine Learning Research* 7 (2006) 1655–1686
26. Tur, G., Rahim, M., Hakkani-Tuk, D.: Active labeling for spoken language understanding. *Eurospeech* (2003) 2782–2789
27. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. *IEEE Trans. PAMI* 22(12) (2000) 1424–1445
28. Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., Movellan, J.R.: Automatic recognition of facial actions in spontaneous expressions. *J. Multimedia* 1(6) (September 2006) 22–35
29. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans. PAMI* 29(10) (October 2007) 1683–1699
30. Heckerman, D., Geiger, D., Chickering, D.M.: Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20(3) (1995) 197–243
31. Heckerman, D.: A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research (1995) 1–40
32. Multiple Aspects of Discourse research lab. <http://madresearchlab.org/>.
33. Douglas-Cowie, E., Cowie, R., Schroeder, M.: The description of naturally occurring emotional speech. *Fifteenth Int'l Congress of Phonetic Sciences* (2003)