Interactive Labeling of Facial Action Units

Lei Zhang, Yan Tong, and Qiang Ji
Rensselaer Polytechnic Institute
zhangl2@rpi.edu, tongyan@ge.com, qji@ecse.rpi.edu

Abstract

For many computer vision problems, it is very important to produce the groundtruth data. Manual data labeling is labor-intensive and prone to the human errors, whereas fully automatic data labeling is not feasible and reliable. In this paper, we propose an interactive labeling technique for efficient and accurate data labeling. Constructed on a Bayesian Network (BN), the automatic image labeler produces an initial labeling of the image. A human then examines the initial labeling and makes minor corrections. The human corrections and the image measurements are then integrated by the BN framework to produce a refined labeling. We demonstrate the capability of this technique on labeling Facial Action Units.

1 Introduction

For a variety of classification tasks in computer vision including image segmentation, object 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 labelers' variations or the imperfect description of different classes. To alleviate these problems, various interactive or semi-automatic learning tools have been used. For example, Levin et al. [6] used co-training for 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 and achieve adequate learning performance. It has been used increasingly to help reduce the amount of labeled training data by incorporating limited user feedback selectively and intelligently during training.

Facial expression recognition represents an active area of research in computer vision. Among the different research directions, the 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 expressions by observing a small set of visually discernable facial muscular movements [7]. Table 1 shows some commonly occurring AUs and their interpretations.

Table 1. Examples of commonly occurring action units and their interpretations [5].

For facial action recognition, AU labeling is required to create the training data. Manual AU labeling of facial expression is very time consuming, error prone, and expensive. It usually requires a long time of training before 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. To further push the researches in 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 AU labels.

For this purpose, we propose an interactive AU labeling system that combines the automatic AU measurements with the selective human inputs to perform both accurate and efficient AU labeling. We construct a unified probabilistic framework based on a BN to incorporate both the automatic AU measurements and the human input. 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 inputs (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) propagate the influences of new human inputs through a principled belief propagation.

2 AU Labeling System

Our automatic AU labeling system consists of two major components. One is an image-based AU classification system based on Adaboost. The other is a probabilistic AU model that captures the spatial and semantic relationships among AUs.

We first perform the face and eye detection from images. Given the knowledge of eye centers, the measurement for each AU is obtained through a computer vision technique based on Gabor wavelet features and AdaBoost classifiers similar to [1]. However, this frame-by-frame AdaBoost classification 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. Therefore, besides measuring each AU individually, it is more important to incorporate other prior knowledge to help improve AU recognition under these difficult situations.

Based on the previous study by Tong et al. [8], there are useful semantic relationships (i.e. 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. Following the prior work by Tong et al. [8], we use a BN, as shown in Figure 1, to capture the co-occurrence and the mutually exclusive relationships among AUs. In addition, the BN also incorporates the human inputs.

Specifically, the nodes in the BN represent AUs and

the links between them capture their relationships (co-occurrence or mutual exclusiveness). The BN can be constructed and parameterized using the training data with a standard learning technique [4]. We employ this BN model as our base algorithm for their effectiveness. Compared with other machine learning methods such as Support Vector Machine and Neural Network, the 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 both the AU measurements and the human input, we associate each AU node AU_i with two measurement nodes represented by a grey circle and a black circle. The grey circle O_i represents the AU measurement obtained through the computer vision technique (i.e. AdaBoost classifier). The link between AU_i and O_i represents the measurement uncertainty, which can be 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. In this way, the BN model systematically combines the objective image measurements of AUs, the subjective human labeling, and their uncertainties.

In the BN model, each AU node has two states: "presence" and "absence". Its parameter is characterized by the Conditional Probabilistic Tables (CPTs) $p(AU_i|pa(AU_i))$, where $pa(AU_i)$ denotes the parent configuration of AU_i. Given the training data (AU labels), these parameters can be learned [3]. For the measurement nodes, the parameters $p(O_i|AU_i)$ or $p(H_i|AU_i)$ are assigned according to the recognition accuracy of the computer vision technique or the human confidence, respectively. Based on the BN model, the states of the AU nodes are inferred using the junction tree inference approach.

3 Interactive AU Labeling

Given the BN model, 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.

$$AU_1^*, \cdots, AU_N^* =$$
$$\operatorname*{argmax}_{AU_1, \cdots, AU_N} p(AU_1, \cdots, AU_N | O_1, \cdots, O_N, H_1, \cdots, H_K)$$
(1)

where $O_1, \cdots, O_N$ represent all the image measurements and $H_1, \cdots, H_K$ represent all the available hu-
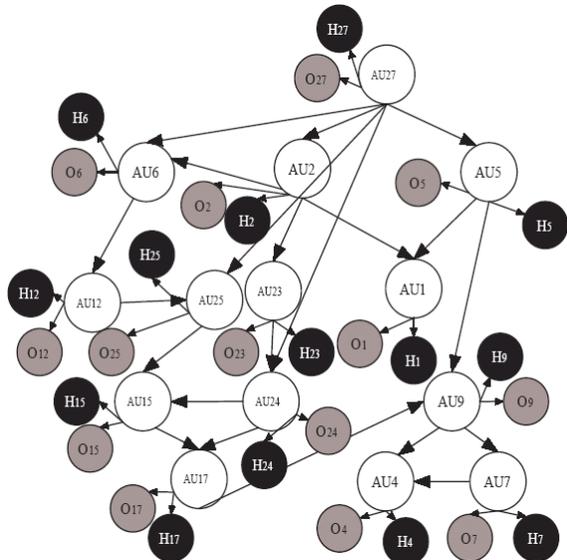
Figure 1. The BN model for interactive labeling of facial action units.

man inputs. Based on the conditional independence in BN, the joint probability can be factorized as follows:

$$p(AU_1, \cdots, AU_N | O_1, \cdots, O_N, H_1, \cdots, H_K) \propto \quad (2)$$

$$\prod_i^N p(AU_i | pa(AU_i)) \prod_i^N p(O_i | AU_i) \prod_k^K p(H_k | pa(H_k))$$

where $pa(AU_i)$ represents the parent configuration of AU_i and $pa(H_k)$ represents the kth 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 from the AdaBoost classifier.

4 Experimental Results

To demonstrate our proposed interactive AU labeling system, we perform the AU labeling experiments on the Cohn-Kanade DFAT-504 database [5]. This database is collected under controlled illumination and background, and has been widely used for evaluating facial AU recognition system. 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. We randomly select 50 samples from the remaining section for testing. Hence, there are totally 400 samples used for testing.

In the interactive labeling, the human coders arbitrarily choose a mislabeled AU for correction. The human coders are domain experts with certain familiarity with
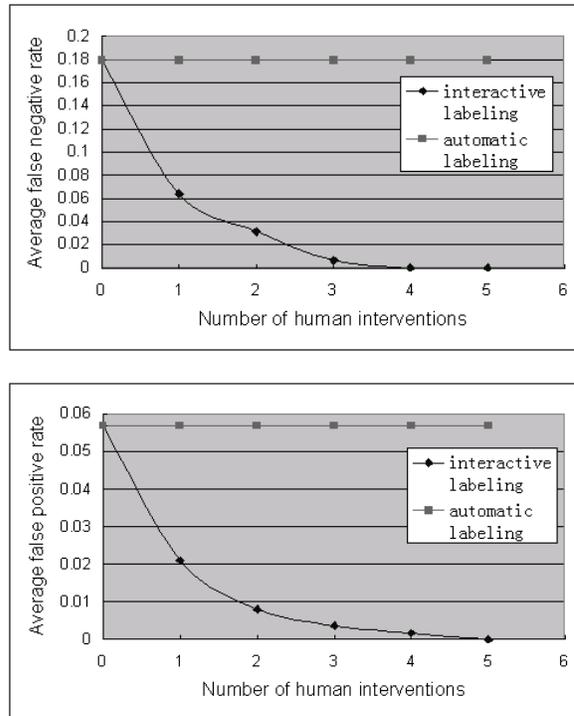
Figure 2. Average AU labeling performance on Cohn-Kanade database using the automatic AU labeling with only image measurements and the interactive labeling with image measurements and arbitrary human inputs.

AU recognition. They are allowed to select the mislabeled AU for correction based on their experiences. In each iteration, only one human intervention is allowed to correct one AU label.

Table 2 reports the accuracy for each AU labeling by using different methods. "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 see the human intervention significantly improves the labeling accuracy. Both the false positive rate and the false negative rate reduce greatly even with only a few interventions. 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.

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

Table 2. Comparison of the AU labeling accuracies on Cohn-Kanade database using the automatic AU labeling with only the image measurements and the interactive labeling with the image measurements and the arbitrary human inputs.

AU label		AU1		AU2		AU4		AU5		AU6		AU7		AU9	
Total test samples		Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#
		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
Automatic labeling		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
Interactive labeling	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

AU label		AU12		AU15		AU17		AU23		AU24		AU25		AU27	
Total test samples		Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#	Pos#	Neg#
		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
Automatic labeling		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
Interactive labeling	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	0.0055	0	0	0	0
	iter5	0	0	0	0	0	0	0	0	0	0	0	0	0	0

rate also decreases from 7.41% (automatic AU labeling) to 3.63%. The false-negative rate of AU25 (lips part) decreases from 17.29% (automatic AU labeling) to 4.08%, and its false-positive rate decreases from 8.99% (automatic AU labeling) to 2.2% using the proposed interactive AU labeling method.

Figure 2 reports the average AU labeling performance for all AUs. We can see 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 2.1% with only one human correction using the interactive AU labeling method.

5 Conclusion

In this paper, we propose an interactive image labeling system for effective combination of the automatic image labeling with the limited human labeling. Using a BN as the engine for the automatic image labeling, 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 systematically combine the image measurements, human inputs, and their uncertainties. We have applied the system to the interactive facial action unit labeling. The experiments demonstrate the effectiveness of the system according to both the AU labeling accuracy and the efficiency. Besides, the proposed framework is generic enough to be applied to different tasks in computer vision, includ-

ing image segmentation, image retrieval, object recognition, etc. In future, we will perform more extensive evaluation on different datasets and study the issue of automatic selection of the AU for correction.

References

[1] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *J. Multimedia*, 1(6):22–35, September 2006.

[2] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System: the Manual*. Research Nexus, Div., Network Information Research Corp., UT, 2002.

[3] D. Heckerman. A tutorial on learning with bayesian networks. *Technical Report MSR-TR-95-06, Microsoft Research*, pages 1–40, 1995.

[4] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3), 1995.

[5] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *Proc. 4th IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[6] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. *Int'l. Conf. on Computer Vision*, pages 13–16, 2003.

[7] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, 2007.

[8] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans. PAMI*, 29(10):1683–1699, October 2007.