PERSON-SPECIFIC EXPRESSION RECOGNITION WITH TRANSFER LEARNING

Jixu Chen, Xiaoming Liu, Peter Tu, Amy Aragones

GE Global Research, Niskayuna, NY 12309, USA

ABSTRACT

A key assumption of traditional machine learning is that both the training and test data share the same distribution. However, this assumption does not hold in many real-world scenarios. For example, in facial expression recognition, the appearance of an expression may vary significantly for different people. Previous work has shown that learning from adequate person-specific data can improve facial expression recognition results. However, because of the difficulties of data collection and labeling, person-specific data is usually very sparse in real-world applications. Learning from the sparse data may suffer from serious over-fitting. In this paper, we propose to learn a person-specific facial expression model through transfer learning. By transferring the informative knowledge from other people, it allows us to learn an accurate person-specific model for a new subject with only a small amount of his/her specific data.

1. INTRODUCTION

In recent years, machine learning approaches have been successfully applied to the field of automatic facial expression recognition. However, many machine learning algorithms work well only under the assumption that the training and test data are drawn from the same distribution. In facial expression recognition, this assumption may hold for some prototypical and posed expressions, such as the "smiling" faces from the Cohn-Kanade DFAT database [6] (Figure 1(a)). Because the posed smile is quite consistent across subjects, current smile detection systems can easily achieve an accuracy of 97% [13] on the DFAT database (leave-one-subject-out cross validation). However, the identical-distribution assumption does not hold for complex and spontaneous expressions, like the "pain" expressions in the PAINFUL database [8] (Figure 1(b)). This database contains the spontaneous pain expressions of patients with shoulder injuries during their shoulder movement. We can observe large variation of the pain expressions across different subjects, such as eyes open or closed, mouth open or closed, etc. Because the training and test data may not share the same distribution, the performance of pain detection is worse than that of posed smile detection.

When the appearance of the facial expression changes across the subjects, learning a person-specific model is likely to achieve better performance than a generic model. However, in many real-world applications, it is expensive to collect and label a large amount of data for a specific new person. Thus, how to learn the person-specific model with limited person-specific data becomes a critical problem.

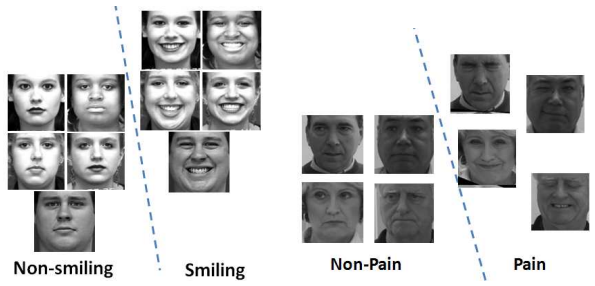

Fig. 1. (a) The posed smile expressions from DFAT database [6]. (b) The spontaneous pain expressions from PAINFUL database [8]. Pain expression has large variation across subjects.

In this paper, we exploit a new promising way to learn a person-specific model for expression recognition: transfer learning. Transfer learning represents a family of algorithms that transfer the informative knowledge from the source data to the new target domain. In our application, we take the pain expressions of other subjects as the source data and learn a person-specific model for a new target subject. We consider two transfer learning scenarios: inductive transfer learning (Sec. 3.1) and transductive transfer learning (Sec. 3.2). For the former, only a small amount of labeled target data are required to learn the robust target model without over-fitting. For the latter, the target data does not need to be labeled hence the data labeling work is entirely avoided. Finally, we compare various transfer learning algorithms and traditional learning algorithms in the PAINFUL database, and show significant improvement of the inductive transfer learning (Sec. 4).

2. RELATED WORK

Facial expression recognition has made considerable progress in recent years. A comprehensive review can be found in [16]. However, most of the current expression recognition research has focused on the posed expression under tightly controlled laboratory condition. There have been very little work on detecting natural spontaneous facial expression [11, 2] which may vary significantly across subjects.

An application of spontaneous facial expression recognition that would be of great benefit is pain/no-pain detection [7]. For instance, in intensive care units (ICU) [5], the improvement in patient outcomes has been achieved by pain monitoring. Lucey et al. [7] collected the spontaneous pain database from patients with shoulder injuries. Their pain detection system achieved 0.751 area under the ROC curve (AUC) using only appearance features, and achieved the best performance of 0.839 AUC by combining the shape and appearance features. In this paper, we propose to further improve this result through a person-specific pain detector.

Previous work [3, 12] has shown that a person-specific model out-performs a person-independent model in expression recognition when adequate person-specific data is available. However, if only a small number of training images for a new subject are available, learning a person-specific model increases the risk of over-fitting. In this paper, we propose to address this problem through transfer learning.

Transfer learning aims to extract knowledge from one or more source domains and improve the learning in the target domain. It has been applied to a wide variety of applications, such as object recognition [14], sign language recognition and text classification. For more details we refer the reader to the survey [9]. In this paper, we apply transfer learning algorithms to the task of person-specific facial expression recognition.

3. PERSON-SPECIFIC FACIAL EXPRESSION RECOGNITION

We first introduce the notation used for the transfer learning problem we intend to approach. Let's denote the training data of a new subject as target data $D_T = \{(\mathbf{x}_{T,i}, y_{T,i})\}_{i=1..N_T}$ and the training data of other subjects as source data $D_S = \{(\mathbf{x}_{S,i}, y_{S,i})\}_{i=1..N_S}$, where $\mathbf{x} \in \mathcal{X}$ is in the feature space and $y \in \{1, +1\}$ is the binary label representing expression presence/absence.

For person-specific facial expression recognition, the goal is to learn a classifier $f_T : \mathbf{x}_T \rightarrow y_T$ from the target data D_T. However, since the size of target data (N_T) is very small, learning from D_T only would suffer serious overfitting problems. Transfer learning can improve the learning of f_T by transferring knowledge from the abundant source data D_S.

3.1. Inductive Transfer Learning Algorithm

In this section, we use the boosting-based inductive transfer learning in [14] to learn the person-specific model. This framework consists of two phases. In the first phase, the knowledge of the source data is represented by a large collection of weak classifiers. In the second phase, some of the weak classifiers are selected to boost the target classifier on the target data.

This algorithm is summarized in Algorithm 1. Notice that it transfers the knowledge from multiple sources. The total number of the source data is $N_S = \sum_{m=1}^{M} N_m$. Compared to the transfer learning from a single source [4], this multi-source transfer learning can identify and take advantage of the sources that closely related to the target, making it less vulnerable to *negative transfer* from unrelated sources.

Algorithm 1 Inductive transfer learning for person-specific model

input: Source data of M subjects $\mathbf{D}_1, ..., \mathbf{D}_M$, where $\mathbf{D}_m = \{(\mathbf{x}_{m,1}, y_{m,1}), ..., (\mathbf{x}_{m,N_m}, y_{m,N_m})\}$. The target data of a new subject $\mathbf{D}_T = \{(\mathbf{x}_{T,1}, y_{T,1}), ..., (\mathbf{x}_{T,N_T}, y_{T,N_T})\}$.
output: A person-specific classifier for the target subject $y = f_T(\mathbf{x})$.
Phase-I Learning a weak classifier set $\mathcal{H}$ from source data $\mathbf{D}_1, ..., \mathbf{D}_M$.
for $m = 1$ to M **do**
  Initialize the weight vector $\mathbf{w}_m^{(1)} = (w_{m,1}^{(1)}, ..., w_{m,N_m}^{(1)})$
  for $k = 1$ to K **do**
    Normalize the weight vector $\mathbf{w}_m$ to 1.
    Find the weak classifier $h_m^{(k)}$ that minimizes the weighted classification error ε over the data set $\mathbf{D}_m$.
    Compute the error $\varepsilon = \sum_{i=1}^{N_m} w_{m,i}^{(k)} [y_{m,i} \neq h_m^{(k)}(\mathbf{x}_{m,i})]$.
    $\alpha = \frac{1}{2} \ln \frac{1-\varepsilon}{\varepsilon}$.
    Update the weights

$$w_{m,i}^{(k+1)} = w_{m,i}^{(k)} \exp\{-\alpha y_{m,i} h_m^{(k)}(\mathbf{x}_{m,i})\}.$$

    $\mathcal{H} \leftarrow \mathcal{H} \bigcup h_m^{(k)}$.
  end for
end for
Phase-II Learning a target classifier on target data $\mathbf{D}_T$.
Initialize the weights $\mathbf{w}_T^{(1)} = (w_{T,1}^{(1)}, ..., w_{T,N_T}^{(1)})$.
for $k = 1$ to K **do**
  Normalize the weight vector $\mathbf{w}_T$ to 1.
  Select one weak classifier $h_T^{(k)}$ from $\mathcal{H}$ that minimizes the weighted classification error ε over the data set $\mathbf{D}_T$.
  Compute the weighted error $\varepsilon = \sum_{i=1}^{N_T} w_{T,i}^{(k)} [y_{T,i} \neq h_T^{(k)}(\mathbf{x}_{T,i})]$.
  $\alpha_T^{(k)} = \frac{1}{2} \ln \frac{1-\varepsilon}{\varepsilon}$.
  Update the weights $w_{T,i} = w_{T,i} \exp\{-\alpha_T^{(k)} y_{T,i} h_T^{(k)}(\mathbf{x}_{T,i})\}$.
  $\mathcal{H} \leftarrow \mathcal{H} \setminus h_T^{(k)}$.
end for
return $f_T(\mathbf{x}) = sign(\sum_k \alpha_T^{(k)} h_T^{(k)}(\mathbf{x}))$.

Phase-I is the standard Adaboost algorithm run for each of the source data. The Adaboost classifier includes the weak classifiers that best discriminate the positive and negative data for that source. All the weak classifiers learned from source data constitute a large classifier set $\mathcal{H}$. Phase-II is a variation of Adaboost on the target D_T. In contrast to the traditional Adaboost which learns weak classifiers from the target data, we pick the weak classifiers from the source classifier set $\mathcal{H}$. Since only the classifiers with the lowest classification rate on D_T are selected, it can ensure the *positive transfer* of the knowledge from source to target.

3.2. Transductive Transfer Learning Algorithm

In this section, we apply the transductive transfer learning algorithm in [10] to the facial expression recognition. This approach is attractive because it can learn the target classifier without knowing the target labels $\{y_{T,1}, ..., y_{T,N_T}\}$, so that the labeling work for a new subject can be entirely avoided.

The basic idea of transfer learning is to re-use the source data that is close to the target. Given the labeled source data $D_S = \{(\mathbf{x}_{S,i}, y_{S,i})\}_{i=1...N_S}$ and the unlabeled target data $D_T = \{\mathbf{x}_{T,j}\}_{j=1...N_T}$, transductive transfer learning reweights every sample $(\mathbf{x}_{S,i}, y_{S,i})$ in the source data using the probability ratio $w(\mathbf{x}_{S,i}) = \frac{p_S(\mathbf{x}_{S,i})}{p_T(\mathbf{x}_{S,i})}$, where $p_S(\mathbf{x})$ and $p_T(\mathbf{x})$ are the marginal distributions of the source and the target, and then the reweighted source data are used to train the target model.

Here, the sample weight $w(\mathbf{x})$ is approximated by a linear model $\hat{w}(\mathbf{x}) = \sum_{l=1}^{b} \alpha_l \phi_l(\mathbf{x})$, where $\phi_l(\mathbf{x})$ is a basis function such that $\phi_l(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. α_l is the parameter to be estimated.

Thus, the target distribution can be approximated by the weighted source distribution: $\hat{p_T}(\mathbf{x}) = \hat{w}(\mathbf{x}) p_S(\mathbf{x})$. Transductive transfer learning minimizes the KL divergence between $\hat{p_T}(\mathbf{x})$ and $p_T(\mathbf{x})$:

$$KL[p_T(\mathbf{x})||\hat{p_T}(\mathbf{x})] = \int p_T(\mathbf{x}) \log \frac{p_T(\mathbf{x})}{\hat{w}(\mathbf{x}) p_S(\mathbf{x})} d\mathbf{x}$$
$$= \int p_T(\mathbf{x}) \log \frac{p_T(\mathbf{x})}{p_S(\mathbf{x})} d\mathbf{x} - \int p_T(\mathbf{x}) \log \hat{w}(\mathbf{x}) d\mathbf{x}$$

. Given the training data, the first term is a constant, we just need to maximize the second term with respect to $\hat{w}(\mathbf{x})$. (For more details of this algorithm please refer to [10]).

Finally, we use the weighted source data to train an Adaboost classifier for the target subject, i.e. the sample weights of the source data are initialized as $\{\hat{w}(\mathbf{x}_{S,i})_{i=1...N_S}\}$ in the AdaBoost learning algorithm.

4. EXPERIMENTAL RESULTS

We tested the transfer learning algorithms on the PAINFUL database [8], which contains video sequences (totally 48,398 frames) of 25 subjects with shoulder injuries.

Local Binary Pattern (LBP) is used as the facial image feature in our experiments. We first use the eye locations provided in the PAINFUL database to crop and warp the face region to a 128×128 image. Following the method in [1], this face image is divided into 8×8 small regions and a 59-dimensional $LBP^{u2}_{8,1}$ feature is extracted from each region. Superscript $u2$ reflects the use of uniform patterns. $(8, 1)$ represents 8 sampling point on a circle of radius of 1. These LBP features are concatenated into a single, spatial enhanced feature with $8 \times 8 \times 59 = 3776$ dimensions (Figure 2).

Similar to [8], we perform a leave-one-subject-out evaluation on 25 subjects. For the target subject, we use the first N_T frames of his/her video sequence to learn the person-specific model via transfer learning and test on the second half of the sequence. The number of frames in the sequence varies from 518 to 3360.

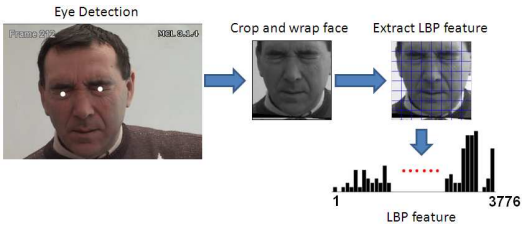

Fig. 2. LBP vector is used as the facial image feature.

Table 1. Performance comparison: AUC for different number of transfer learning data (N_T). 'half' represents the result using the first half of the target data for transfer learning.

N_T	10	25	50	100	half
Traditional Model-A	0.557	0.684	0.786	0.862	0.893
Traditional Model-B	0.786	0.816	0.819	0.835	0.878
Inductive Transfer	0.782	0.821	0.880	0.891	0.895
Transductive Transfer	0.756	0.755	0.765	0.756	0.760
Generic Model (Baseline)	0.769				

We compare the generic model and four different person-specific models as follows. *Generic model* is the Adaboost classifier learned from the source data of 24 subjects. This is our baseline algorithm. *Traditional person-specific Model-A* is a Adaboost classifier learned only from the target data without transfer learning. *Traditional person-specific Model-B* is a Adaboost classifier learned from a combined dataset of both the source data and the target data. *Inductive transfer model* is learned using Algorithm 1. *Transductive transfer model* is learned using the algorithm in Sec. 3.2. Each of the above models consists of 50 weak classifiers. When the number of training samples is 50, the ROCs of these models are shown in Figure 3. The results with different number of training samples are summarized in Table 1.

For the traditional person-specific Model-A, we learn the Adaboost classifier only from the person-specific target data. This model suffers serious over-fitting when the target data is limited (AUC is 0.557 when the number of target data is 10). Its performance can be improved by adding more training data, but it is always worse than inductive transfer learning. When using adequate training data (i.e. half of the target data), its performance is close to inductive transfer learning.

For the traditional person-specific Model-B, the classifier is learned from a combined data-set that consists of both the source and the target data. Because we have a large amount of training data, over-fitting problem can be avoided. However, because this classifier focuses on the combined data-set, its performance on the target data is not as good as Model-A when the target data is sufficient. Furthermore, since the training data size is very large, the learning process is very time consuming. We list the average training time for person-specific models in Table 2.

The inductive transfer learning achieves the best perfor-

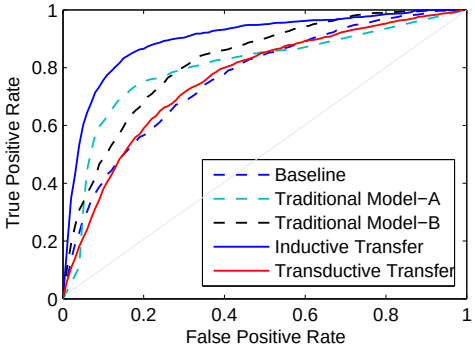

Fig. 3. ROCs when $N_T = 50$.

Table 2. Average time for training a person-specific model.

Traditional Model-A	2.6 min
Traditional Model-B	14.3 min
Inductive Transfer Model	0.16 min
Transductive Tranfer Model	17.6 min

mance among person-specific models. It out-performs the baseline with a small number of target training data (AUC is improved from 0.769 to 0.782 with only 10 samples) and its performance increases significantly when adding more training samples (AUC=0.891 with 100 target samples). Furthermore, because inductive transfer learning does not need to train new weak classifiers, it is the fastest algorithm in Table 2, which makes it suitable for rapid retraining for a new target.

For the transductive transfer learning, we didn't observe any improvement even with adequate training data. A possible reason is that the boosting classifier is not sensitive to the marginal distribution change. In [15], the classifiers are grouped into two categories: *local classifiers*, which depend only on $P(y|\mathbf{x})$, and *global classifiers*, which depend on both $P(y|\mathbf{x})$ and $P(\mathbf{x})$. In our transductive transfer learning, we only reweight the source data to approximate the target marginal distribution $P_T(\mathbf{x})$. Since the AdaBoost classifier tends to be a local learner, this tranductive transfer may not work.

The training and testing time of an Adaboost classifier is proportional to its number of weak classifiers. An efficient algorithm can learn a good AdaBoost classifier with fewer weak classifiers. Figure 4 depicts the performance of different algorithms with different number of weak classifiers. It shows that inductive transfer learning can achieve good performance (AUC=0.818) with only five weak classifiers, which further confirms its efficacy.

5. CONCLUSION

This work exploits the idea of learning a person-specific model to improve facial expression recognition. In order to learn a robust person-specific model with minimal new data input, we propose to use the transfer learning, which can mitigate the overfitting in the target domain by transferring the informative knowledge from similar source domains. We deploy and evaluate different transfer learning algorithms within the context of pain expression recognition. Compared to the traditional methods, the experiment shows that *inductive transfer learning* can significantly improve the recognition performance with a limited number of target samples.

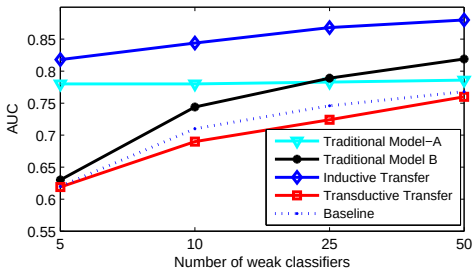

Fig. 4. AUC of different models with different number of weak classifiers when $N_T = 50$.

6. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE T-PAMI*, 28(12):2037 –2041, 2006.

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

[3] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *CVIU*, 91(1-2):160 – 187, 2003.

[4] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *ICML*, pages 193–200, 2007.

[5] A. Gawande. *The checklist manifesto: how to get things right*. Metropolitan Books, 2009.

[6] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. In *FG*, pages 46 – 53, 2000.

[7] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE T-SMC, Part B: Cybernetics*, 41(3):664–674, 2011.

[8] P. Lucey, J. F.Cohn, K. M. Prkachin, P. E.Solomon, and I. Matthews. Painful data : The UNBC-McMaster shoulder pain expression archive database. In *FG*, pages 57–64, 2011.

[9] S. J. Pan and Q. Yang. A survey on transfer learning. *T-KDE*, 22(10):1345–1359, 2010.

[10] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2007.

[11] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE T-PAMI*, 32(2):258–273, 2010.

[12] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *FG*, pages 921–926, 2011.

[13] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE T-PAMI*, 31(11):2106–2111, 2009.

[14] Y. Yao and G. Doretto. Boosting for transfer learning with multiple sources. In *CVPR*, pages 1855–1862, 2010.

[15] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML*, pages 903–910, 2004.

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T-PAMI*, 31(1):39–58, 2009.