2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A New Transfer Learning Boosting Approach Based on Distribution Measure with an Application on Facial Expression Recognition

Shihai Wang
School of Reliability and System Engineering
Science and Technology on Reliability and
Environmental Engineering Laboratory
Beihang University
Beijing, China
wangshihai@buaa.edu.cn

Zelin Li
School of Reliability and System Engineering
Science and Technology on Reliability and
Environmental Engineering Laboratory
Beihang University
Beijing, China
lizelin@dse.buaa.edu.cn

***Abstract*—In the machine learning community, most algorithms proposed, particularly for inductive learning, are based entirely on one crucial assumption: that the training and test data points are drawn or generated from the exact same distribution. If this condition is not fully satisfied, most learning algorithms or models are corrupted. In this paper, we propose a new instance based transductive transfer learning method based on Boosting framework by using a distribution measure approach. There follows a detailed description of this distribution measure approach. Subsequently, we describe our boosting transfer learning method in detail and report its performance in facial expression recognition tasks.**

***Keywords*—boosting; transfer learning; distribution measure; facial expression recognition**

I. INTRODUCTION ON TRANSFER LEAERNING

ThisIeeeparstartTraditional Traditional machine learning makes a vital assumption: the training and test data should have the same distribution. However, in practice, such an assumption may not always hold. There are several examples that reflect the value of transfer learning and lead to our interest in this research area.

Imagine there is website in a university; it hosts a huge amount of web documents previously labelled by experts. The task is to classify these web documents into several predefined categories. This is a typical real world application for web-document classification. Further, as we know, there are many new websites built every day, leading to another classification task. A newly built website needs to also perform web document classification. However, in this case, as the features or distributions may be slightly different, being from a different website, and as the experts have not labelled a new training dataset, we cannot use the previous model trained on the university website to perform classification in this new task. A new challenge is in front of us; how can we build a model to achieve acceptable classification performance on the new website by using the limited labelled training set collected from one website only? In such cases, transfer learning, which can transfer the classification knowledge into the new domain, can help.

Another real world example might involve the goal of providing automatic classification of customer review documents for a product, for instance a range of CD players, into two categories, positive and negative. In order to obtain good performance in this classification task, we need to collect and assign a label (positive or negative) to as many review samples as possible. Then we can build or train a model. However, the distributions of review data for various types of CD players are different. Therefore, one suggestion may be to collect the review data and manually assign a label to them on each individual product type separately, thus training the classifier for each product separately. Unfortunately, this will prove to be very expensive. In such cases, we want to train a classification model on the review data of some products and apply it to other products review data as well. Naturally this brings to mind the very concept of the transfer learning, which can save an immeasurable amount of effort on labeling work [1].

The initial motivation for the study of transfer learning comes from observing human behavior; human beings are able to intelligently apply previous knowledge to deal with new problems or challenges, creating better solutions without undergoing a re-learning process. The first discussion on transfer learning in the machine learning community was in the NIPS-95 workshop. Since 1995, and as a new research topic transfer learning has attracted a large amount of attention and been named in different ways; learning to learn, knowledge transfer, multi-task learning, knowledge consolidation, incremental learning [2]. Multi-task learning is the technique most closely related to transfer learning. In multi-task learning methods [3] the goal is to learn different tasks simultaneously.

This work was supported by a grant from the Major State Basic Research Development Program of China (973 Program) (No.2014CB744904), and also by a grant from the National Natural Science Foundation of China (No. 61300069).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

Transfer learning tries to extract the knowledge from one or more data sources and transformationally apply the knowledge yielded from previous tasks to new ones. It is worth noting, however, the difference between these two learning techniques. Multi-task learning aims to learn all the sources and tasks together. In contrast, transfer learning pays greater attention to the target task.

Before we make categories for transfer learning, there are three questions to deal with; what to transfer, how to transfer, and when to transfer. To answer the first question, we need to know which parts of the knowledge are available for transfer to another task. In some tasks there is specific knowledge not shared with other tasks, but some knowledge may be common among different tasks. If we can extract more information from this common knowledge, we may produce improved performance in the target task. The subsequent question is how to transfer. Here, learning algorithms should be able to learn the common knowledge and enable its transfer. The final question, when to transfer, must be answered in a reverse fashion; i.e. when should we not perform transfer learning? Obviously when the source task and target task have no relationship between them, using transfer learning will corrupt the performance of learning; we refer to this as negative transfer.

In traditional machine learning, based on different situations and learning tasks, transfer learning has been categorized as follows: inductive transfer learning, transductive transfer learning and unsupervised transfer learning. When the target task is different from the source task and there is some labelled data available in the target task, we perform inductive transfer learning. In cases where no labelled data is accessible in the target domain, moreover the target and source tasks are the same, but the feature space or the marginal probability distributions of the input data are different, we must perform transductive transfer learning. Finally, in unsupervised transfer learning, the target and source tasks are not the same but related. Moreover the labelling information is unavailable in both. The unsupervised learning targets, for instance clustering and density estimation, are concentrated on. In this paper, we propose a transductive transfer learning based on Boosting framework.

Regarding "what to transfer", different approaches in transfer learning can be assigned to the following cases. The first case is instance based transfer learning [4], [5], [6], [7], [8], [9], [10], [11], [12], in which we assume there is data in the source domain, which can be used as learning samples in the target domain. The second is referred to as the feature representation transfer approach [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. The basic idea behind this is that there are some common or good features in the feature space, shared by both the source domain and target domain. If the transfer learning method can learn the shared feature representation in the source domain, the performance in the target domain should be improved. The third case is to take a parameter transfer approach [23], [24], [25], [26], [27]. In this case we assume that the models generated from the source and target tasks have some common parameters or prior distributions of the hyper-parameters. By using these shared parameters between two models, the knowledge could be transferred. The fourth can be termed relational knowledge transfer problems [28]. Here it is believed that there are some relationships within data which is shared or similar between the source and target tasks. Here, transferring the knowledge amounts to discovering and transferring the relationships within the data. To sum up, in order to perform transfer learning there are assuredly certain common and shareable parts, e.g. common instances, common features, shareable parameters, and relationships within data. Finding these shareable or "good" parts between source and target tasks becomes one vital challenge in transfer learning.

In transductive transfer learning, the source and target tasks must be exactly the same, and some or all of the unlabeled samples in the target task must be available. The data distributions in two domains are allowed to be different. In [29] the authors initially propose transductive transfer learning. All unlabeled data are requested to be available to them at training time. Here is a formal description for transductive transfer learning. Based on two domains, source domain Ds and target domain Dt, corresponding to two tasks, learning task Ts and target task Tt, transductive transfer learning aims to yield a good performance by classification function Ft on the target task by transferring the knowledge learned in Ds and Ts, where $Ds \neq Dt$ and Ts=Tt, to Tt, with help from some unlabeled instances in Tt at training time. The method we propose later is related to this type of transfer learning. In [30], [9], [12] all the approaches proposed by the authors are related to instance based transductive transfer learning. The basic ideas behind these approaches are similar; by using different strategies to add different penalties to each instance in source task Ts a usable model is generated for target task Tt. Another type of transductive transfer learning is based on feature representation. A structural learning algorithm is proposed by Blitzer et al. [31]. In the first stage, some high valued features are selected based on the analysis of data from both domains. In the following stage, these valued features are used to generate a set of new augmented features by using singular value decomposition. Finally, a model suitable for the target task is built on these augmented features.

A. An Introduction on Boosting Learning for Transfer Learning

As a new research topic, transfer learning, especially transfer Boosting learning, has not yet attracted much attention. Dai et al. [32] propose an instance base inductive transfer learning approach based on the Boosting framework, TrAdaBoost. As an extension of the AdaBoost algorithm TrAdaBoost assumes that exactly the same feature set and label set are shared between the source and target data. The only difference is the distribution between the two. As an instance based transfer approach, TrAdaBoost also works with the assumption that certain data points in the source domain could be helpfully used when learning the target tasks. In TrAdaBoost data in the source domain are re-weighted for each Boosting iteration;

useful data are assigned high weights whilst "bad" data receive relatively low weights. The goal is that TrAdaBoost forces the Boosting algorithm to learn these "valuable" instances in the source data and ignore the "bad" ones. In comparison with the lack of need for labelled samples in the target domain for our method, as an inductive transfer learning method, TrAdaBoost asks for some labelled instances to be available in target task.

II. TRANSFER LEARNING ON FACIAL EXPRESSION RECOGNTION

A. The Motivation for Using Transfer Learning in Facial Expression Recognition

In the machine learning community, to obtain a good performance from a learning algorithm, people hope to have a training dataset as large as they can. It is hard to say how big a dataset is required for a particular learning task, it is decided by the complexity of the task, but as all researchers have agreed, a larger training dataset can offer a fuller coverage of the distribution of entire data and more inherent information for learning, consequently a better performance for any machine learning approach is definitely yielded. As a learning task with high complexity in a high dimension feature space, people desperately need a large dataset for learning facial expression recognition task. To collect and label such large amount of facial expression training images by human only, however, is always an extremely difficult, expensive, and time consuming job. This leads to that there is always a struggle with the lack of training data for learning facial expression recognition task.

Imagine this situation; a set of facial expression images with correct label information is available to us. Certainly it is not difficult for us to build a model and yield a good recognition performance on this set of images. However the target has been changed. We have another facial expression image set without labels and the data distribution is different from the previous one —perhaps this dataset is collected from a different ethnic group than the labelled facial expression dataset we already have. We are not able to spend a long time or make huge efforts in labelling these facial images, but we are then supposed to yield an acceptable classification performance on this dataset despite its having no labels. It is worth stating that this situation differs from one which would use semi-supervised learning. In semi-supervised learning, although there is also an unlabeled dataset, we assume that unlabeled data is drawn from the same distribution as the labelled data. Here, such an assumption is not necessarily held when using transfer learning.

When transfer learning is employed, although it allows for the training and testing datasets to be drawn from different distributions, a relationship between the two datasets remains necessary. After Ekman and Izards assertion that the 6 universal facial expressions —happiness, sadness, surprise, anger, fear, and disgust —are regardless of the difference in culture and nationality, using a transfer learning technique becomes a real possibility in facial expression recognition tasks. No matter from which nationality the facial expression images are collected, these 6 universal facial expressions can always be used as the classification information. Therefore, the same classification task can be applied to various facial expression datasets. A proper use of transfer learning in facial expression recognition tasks can dramatically reduce the amount of effort and time required for labelling a new facial expression image set.

B. An Introduction on Maximum Mean Discrepancy (MMD)

In [33] the concentrated issue that the authors address is whether or not the samples are drawn from different distributions and how great are the differences between them. For example two distributions p and q, after regenerating a Kernel Hilbert Space and mapping the distributions into the space, a smooth function is employed to justify the difference. When this smooth function produces a large and positive value the samples are drawn from different distributions. Contrarily when a small or negative value is produced by the smooth function the samples are collected from the similar distributions. A justification is made by the difference, a relatively large difference means the two samples possibly are from two different distributions; otherwise, they are from the similar distribution. This difference is called as Maximum Mean Discrepancy (MMD). Here, the key challenge is finding a properly smooth function F; a "rich enough" F can make the MMD be disappeared when p=q and be as large as possible when $p \neq q$. The authors selected F to be a unit ball in a universal RKHS H [45] with associated kernel k(.), where a Gaussian kernel is chosen. Here is a brief explanation of Reproducing Kernel Hilbert Space (RKHS). The RKHS provides a rigorous and effective framework for smooth multivariate interpolation of arbitrarily scattered data and for accurate approximation of general multidimensional functions. Let X be an arbitrary set and H a Hilbert space [46] of complex-valued functions on X. If a linear map from H to the complex numbers is continuous fro and x in X, we say that H is a reproducing kernel Hilbert space. After an F is selected the next is using the framework of statistical hypothesis testing. Two set of samples x and y are drawn from distributions p and q respectively. The test statistic MMD [F, x ,y] is produced with a particular threshold; x $\in$ X, $y \in Y$. When the threshold is exceeded the y is placed in rejecting set Q and is considered to be a sample drawn from a different distribution than was x. Finally, when the population is zero in Q the distributions p and q are justified as the same. In our boosting transfer learning method the test statistic MMD [F, x ,y] is used as a measure of the size of the difference between the distributions generating the samples x and y.

C. A Transfer Learning Boosting Approach Based on Distribution Measure

Promising applications of Transfer learning to facial expression recognition tasks have stirred great passion. Here, we propose a new instance based transductive transfer learning method based on Boosting framework by using a distribution measure approach. There follows a detailed description of our boosting transfer learning method, after then we demonstrate the performance on a synthetic dataset and facial expression recognition tasks.

By way of clarification for the proposal of our method; our method is related to an instance base transductive transfer learning method. When two sets of data points, A and B are available, A with labels and B without, the aim is to generate a classifier trained on dataset A which achieves acceptable performance on dataset B with aid from B. For such a situation, the original motivation is revealed at pervious section.

The basic idea behind our method is that after the test statistic MMD produces the measure of difference for the two distributions based on each pair of samples, the measure is employed to guide the Boosting method towards greater respect for data points with relatively small difference. A detailed description follows.

In our approach the first step is initialization, where the test statistic MMD produces the difference of each pair of samples, a and b where $a \in A \ and \ b \in B$. Afterwards, for each a, we summarize its difference with every point in set B.

$$SMMD_i = \sum_{j=1}^{N} MMD(a_i, b_j) \tag{1}$$

Where N is the size of set B. Subsequently, we employ a squeeze mapping function to adjust the factor, $SMMD_i$, which is inversely mapped into between [0],[1] interval non-linearly. The squeeze mapping function is defined as:

$$IMF(SMMD_i) = \frac{1}{1 + e^{\varepsilon \times SMMD_i}} \tag{2}$$

Where a is a sample in set A. ε is coefficient to adjust the shape of this mapping function. After we adjust the steepness of this non-liner function by choosing a range of ε, ε is set to 2 in our implementation. Here, we call the output the importance factor (IMF) where a large difference yielded from MMD will be inversely mapped to a relatively small value; conversely, a big importance factor is obtained when the difference is small.

After we yield IMF for each of the data points in set A, in the following step we modify the current Boosting approach, AdaBoost, to make it learn each instance in set A based on its IMF. In other words, training samples in set A with large IMF will be assigned large cost values in Boosting algorithm. Here, we propose a new Boosting approach called ITRBoost, which is related to instance based transductive transfer learning. We define a cost function as:

$$C(F) = \sum_{i=1}^{N} \beta_i \log(1 + e^{-y_i F(x_i)}) \tag{3}$$

Where, N is the size of set A, β_i is the IMF of sample i. y_i is the label. Subsequently, we can derive ITRBoost by the sequential minimization of this cost function (3).

Given that we have chosen f_{t+1}, we wish to choose ε_{t+1} to minimize the cost function:

$$C(F) = \sum_{i=1}^{N} \beta_i \log(1 + e^{-y_i[F_t(x_i) + \varepsilon_{t+1} f_{t+1}(x_i)]}) \tag{4}$$

Here the ε is defined as:

$$\varepsilon = \frac{1}{2}\ln(\frac{\sum\limits_{i=1\cdots N, f(x_i)=y_i} \beta_i e^{-y_i F(x_i)}}{\sum\limits_{i=1\cdots N, f(x_i)\neq y_i} \beta_i e^{-y_i F(x_i)}}) \tag{5}$$

By using the inner product $- < \nabla C(F), f >$, it is found that finding an f maximizing $- < \nabla C(F), f >$ is equivalent to finding an f minimizing the weighted error of samples, $\sum\limits_{i:f(x_i)\neq y_i} D(i)$ .More details are provided in the following.

$$\sum_{i=1, f(x_i)\neq y_i}^{N} \frac{\beta_i e^{-y_i F(x_i)}}{1 + e^{-y_i F(x_i)}}$$

Cost Function: $C(F) = \sum\limits_{i=1}^{N} \beta_i \log(1 + e^{-y_i F(x_i)})$

finding an f maximizing $- < \nabla C(F), f >$

$- < \nabla C(F), f > = \sum\limits_{i=1}^{N} \beta_i \frac{e^{-y_i F(x_i)}}{1 + e^{-y_i F(x_i)}} (y_i f(x_i))$

$= \sum\limits_{i=1, f(x_i)=y_i}^{N} \frac{\beta_i e^{-y_i F(x_i)}}{1 + e^{-y_i F(x_i)}} - \sum\limits_{i=1, f(x_i)\neq y_i}^{N} \frac{\beta_i e^{-y_i F(x_i)}}{1 + e^{-y_i F(x_i)}}$

$= 1 - 2 \sum\limits_{i=1, f(x_i)\neq y_i}^{N} \frac{\beta_i e^{-y_i F(x_i)}}{1 + e^{-y_i F(x_i)}}$

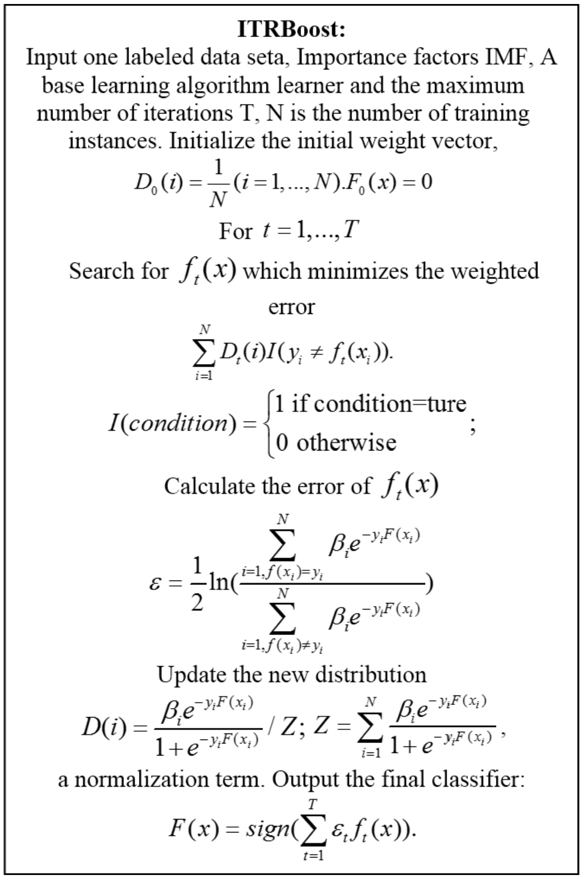
Thus finding an f maximizing $- < \nabla C(F), f >$

is equivalent to finding an f minimizing the weighted error

$$\sum_{i=1, f(x_i)\neq y_i}^{N} \frac{\beta_i e^{-y_i F(x_i)}}{1 + e^{-y_i F(x_i)}}$$

Let us define ITRBoost: an instance based transductive transfer learning. Assuming F is a linear combination of base hypotheses class, and f={-1,+1}. Y is the label, Y={-1,+1}, $y_i \in Y$.

A formal description of the Boosting transfer learning method we propose based on binary classification problems has been provided at follow. As can be seen from the algorithm, in each Boosting iteration round, the training instances with large IMF will relate to larger training weights. In other words, the training samples in dataset A with relatively larger differences from the distribution of dataset B will be weakly learned or ignored by our Boosting method. At the end, our method is trained on the data points with similar distribution to dataset B. Hence a reasonably good performance on dataset B is the result.



ITRBoost:

Input one labeled data seta, Importance factors IMF, A base learning algorithm learner and the maximum number of iterations T, N is the number of training instances. Initialize the initial weight vector,

$$D_0(i)=\frac{1}{N}(i=1,...,N).F_0(x)=0$$

For $t=1,...,T$

Search for $f_t(x)$ which minimizes the weighted error

$$\sum_{i=1}^{N} D_t(i)I(y_i \neq f_t(x_i)).$$

$$I(condition)=\begin{cases}1 \text{ if condition=ture}\\ 0 \text{ otherwise}\end{cases};$$

Calculate the error of $f_t(x)$

$$\varepsilon=\frac{1}{2}\ln(\frac{\sum_{i=1,f(x_i)=y_i}^{N}\beta_i e^{-y_iF(x_i)}}{\sum_{i=1,f(x_i)\neq y_i}^{N}\beta_i e^{-y_iF(x_i)}})$$

Update the new distribution

$$D(i)=\frac{\beta_i e^{-y_iF(x_i)}}{1+e^{-y_iF(x_i)}}/Z;\ Z=\sum_{i=1}^{N}\frac{\beta_i e^{-y_iF(x_i)}}{1+e^{-y_iF(x_i)}},$$

a normalization term. Output the final classifier:

$$F(x)=sign(\sum_{t=1}^{T}\varepsilon_t f_t(x)).$$

Fig. 1. ITRBoost

III. EXPERIMENTS

In this section, in order to demonstrate the effectiveness of ITRBoost on transfer learning we first generate two synthetic datasets with the same task and different distributions, subsequently we extend our method to multi-class problems with a one-against-rest strategy [34] and apply it to facial expression recognition tasks with two facial expression datasets. The effectiveness of this method has been validated.

A. Experiments on Synthetic Data Classification

Here, Fig. 2 and Fig. 3 illustrate two synthetic datasets for source and target tasks respectively generated by two different sets of Gaussian distributions with the same means and different covariance. In order to perform transfer learning on the two datasets, the same classification task, two categories, is provided in both. In source domain 400 data points are generated, 200 points for each category; target domain has 200 points, 100 point for each. In the following we perform the transfer learning by using AdaBoost and ITRBoost, training classifier on source task and testing on target task; two decision boundaries are generated by them respectively, the MLP with 3 hidden neurons was employed as base learner, and the maximum number of iteration is set as 20.

In the following the decision boundary on the target domain generated by AdaBoost trained on the source domain is shown in Fig. 4. Clearly this decision boundary is suitable for the source task but is poor on the target task.

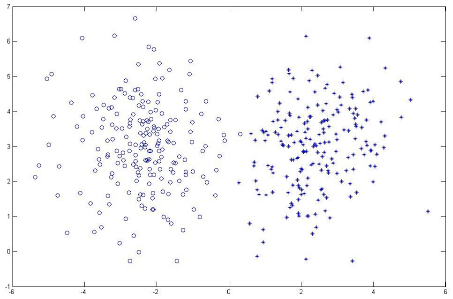


Fig. 2. Synthetic Dataset for Source Task

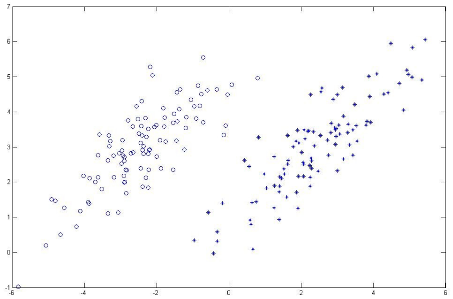


Fig. 3. Synthetic Dataset for Target Task

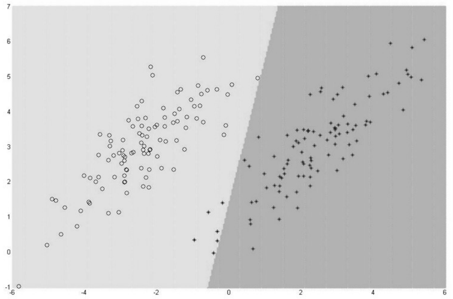


Fig. 4. Decision Boundary of AdaBoost on Target Task

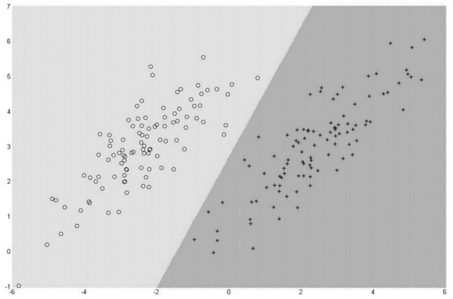


Fig. 5. Decision Boundary of AdaBoost on Target Task

From Fig. 5 it is clear that ITRBoost successfully answers the question “what is transferred”. These points in the source domain, which is close to the distribution of the target domain, are found out and employed as transfer training samples in ITRBoost. A decision boundary with better performance on the target task is built.

B. Experimental Facial Expression Datasets and the Representation Method Details

In this paper, to demonstrate and evaluate the performance of the approaches we propose for facial expression recognition and analysis problems, we mainly conduct our experiments using two facial expression databases; the Japanese Female Facial Expression (JAFFE) dataset [35] and the AR face database [36]. Here, we give some samples and details of each of the datasets. A powerful representation method, Gabor wavelet representation, is chosen as the feature extraction method in our experiments. A detailed description for using this method is given in this section as well.

JAFFE: This database contains 213 images of Japanese female facial expression. Ten female facial expressers are involved in this dataset. Each of the expressers poses 3 or 4 times for each of the 7 universal facial expressions (anger, disgust, fear, happiness, neutral, sadness, and surprise). There are some samples from the JAFFE dataset in Fig. 6. The original image size is 256256 pixels. As our work is essentially focused on the classification step, we have manually cropped every face image to remove the influence of the background. The resized image is 120160 pixels. This is, however, not an absolute necessity if all the subjects are located at roughly the same position and detected well in every image.

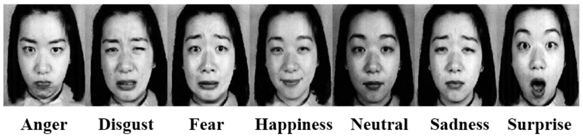


Fig. 6. Facial Expression Samples from Japanese Female Facial Expression (JAFFE) Dataset

AR: In the AR face dataset there are 56 female and 70 male facial expressers. Each expresser shows four expressions: neutral, smile, anger, and scream. Some examples selected from the AR database are shown in Fig. 7. For each persons expression, two images are taken from two different sessions. Thus in all we have a total of 1008 pictures of 768×576 pixels each. We also reduce each image to 230×250 pixels. All the pictures are collected in situations involving different illumination conditions, different backgrounds, with the subject wearing glasses and scarf or not. All these diverse situations make the difficulties in making accurate facial expression recognition on this database far higher than with other datasets. However, this makes the recognition task much closer to a true situation in the real world.

Representation of Facial Expression Datasets: In [37] there is a comparison of a range of feature extraction methods for facial expression recognition. At the end, the authors point out that the best recognition performance is obtained when using Gabor representation as Gabor representation provides more discriminatory power.

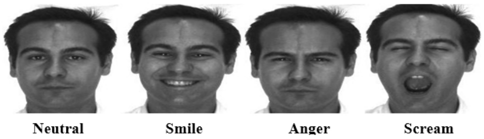


Fig. 7. Facial Expression Samples from AR Dataset

The 2-D Gabor wavelet filters have been proven to be a very useful tool in computer vision and image analysis, [38], [39], [40], [41], [42], [43]. One of benefit is the lack of sensitivity to differences in illumination. Use of 2-D Gabor wavelet representation in computer vision was pioneered by Daugman in the 1980s [44]. The Gabor wavelet representation allows description of spatial frequency structure in the image while preserving information about spatial relations.

In our work, after an original image is transformed by the Gabor filter bank, in a Gabor feature space there are 24 Gabor feature images. For each Gabor filter, the output is a Gabor component-feature. Finally, we concatenate all the outputs of every Gabor filter and derive a Gabor feature vector. Apparently the dimensionality of a Gabor feature vector is extremely high. Thus, for receiving a better generalized performance and to decrease the computational cost, we employ the principal component analysis (PCA) to reduce the dimensions.

To conclude, we keep the top 40 and 100 PCA coefficients to form a feature vector for the JAFFE and AR datasets respectively. Now in a feature space, an original facial expression image is represented by a vector with 40 elements for JAFFE or 100 elements for AR.

C. Experimental Results and Analysis on Facial Expression Recognition Tasks

In this section, we describe a series of simulation studies which demonstrate the performances of the method we propose for facial expression recognition tasks.Two facial expression datasets are involved, JAFFE and AR. The JAFFE dataset is a collection of female facial expression images from Japan. The AR dataset involves a set of facial expression images collected from Europeans of different genders. Obviously, the differences in nationality, ethnic group and gender make the distributions of the two datasets dissimilar. There is one condition, involving the same task in different datasets, which must be met before we can perform transfer learning. Thus, we select samples with the same facial expressions between the two datasets - anger, happiness and natural - and we keep the top 40 PCA coefficients to form a feature vector for both of the datasets. Now, in feature space both have the same dimensions. In order to achieve fair experimental results, we balance the sizes between the two datasets, selecting around 90 images from each. In the experiments, MLP is employed as base learner; and the number of hidden neurons is 30, and 50 is used as the maximal number of iterations.

Experiment 1: In this experiment, the JAFFE dataset is used as the source training set, and the performances are yielded from testing on the AR data.

Figure 8 shows the performances of AdaBoost and our method used to do transfer learning from JAFFE to AR. A significant improvement has been produced by our method.

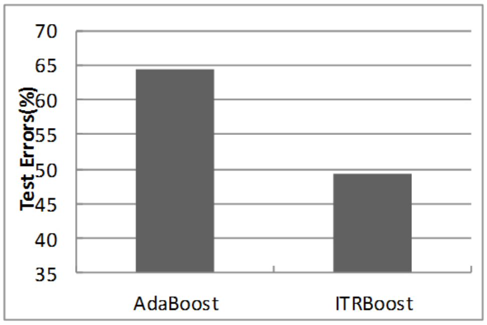


Fig. 8. Transfer learning from JAFFE to AR

Experiment 2: Another experiment is carried out in reverse. We swap the source training set and target set. In other words we transfer the knowledge learned in the AR dataset to the JAFFE dataset.

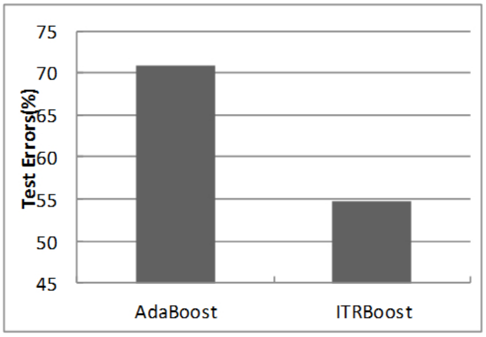


Fig. 9. Transfer learning from AR to JAFFE

Figure 8 and 9 provide the results based on testing performances of transfer learning by using two facial expression datasets. It can be clearly observed that AdaBoost as a successful supervised learning method cannot yield a good performance in the case of training and testing sets drawn from different distributions. As ITRBoost uses the difference of distribution between the two dataset, a significant improvement is achieved.

IV. SUMMARY AND FURTHER WORK

In this paper, we have made an investigation into transfer learning and presented a transfer learning method based on a Boosting framework by using the method of measuring the difference of distributions. After comparison between the testing performances of the baseline method, AdaBoost, and our method on two facial expression datasets with obviously dissimilar distributions, promising results are yielded. Recently, another Boosting transfer learning approach [32] has been proposed, namely TrAdaboost. This method is related to instance base inductive transfer learning. Therefore, labels in the source (training) dataset are necessary; moreover a number of samples in the target (testing) dataset must have label information as well. This is in contrast with our method, where this effort is unnecessary.

As a pilot study of transfer learning in facial expression recognition tasks, there is a lot of ongoing work for us. Our experiments thus far have involved only two facial expression datasets. Clearly, more facial expression image sets are needed to further evaluate the effectiveness of our method. In the current stages, the measure of difference between distributions is used to guide a Boosting algorithm to selectively learn the training samples. However it is not comprehensive and a large amount of unlabeled data in target domain has not been involved in the Boosting learning process. To further improve the performance, we believe that when using transfer learning to transfer the knowledge, we should not ignore the unlabeled data which contains rich learning information in the target domain. A new Boosting transfer learning method, in which both labelled data in source domain and unlabeled data in the target domain is able to be involved in the learning process and a closer distribution of target data can be discovered, is among our ongoing work.

REFERENCES

[1] J. Blitzer, M. Dredze, and F. “Pereira. Biographies, Bollywood, boomboxes and blenders: Domain adaptation for sentiment classification,” In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 432–439. 2007.

[2] S. Thrun, and L. Pratt. Learning to learn. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[3] R. Caruana, “Multitask learning,” Machine Learning, vol. 28(1), pp. 41–75, 1997.

[4] W. Dai, G. Xue, Q. Yang, and Y. Yu, “Transferring naive bayes classifiers for text classification,” In Proceedings of the 22rd AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, pp. 540–545. 2007.

[5] J. Q. Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. Dataset Shift in Machine Learning. The MIT Pres, 2009.

[6] J. Jiang and C. Zhai, “Instance weighting for domain adaptation in NLP,” In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics, pp. 264–271. 2007.

[7] X. Liao, Y. Xue, and L. Carin, Logistic regression with an auxiliary data source. In Proceedings of the 21st International Conference on Machine Learning, Bonn, Germany, pp. 505–512 2005.

[8] B. Zadrozny, Learning and evaluating classifiers under sample selection bias. In Proceedings of the 21st International Conference on Machine Learning, Banff, Alberta, Canada. 2004.

[9] J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf, Correcting sample selection bias by unlabeled data. In: Proceedings of the 19th Annual Conference on Neural Information Processing Systems 2007.

[10] S. Bickel, M. Bruckner, and T. Scheffer, Discriminative learning for differing training and test distributions. In Proceedings of the 24th international conference on Machine learning. New York, NY, USA: ACM, pp. 81–88. 2007.

[11] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, Direct importance estimation with model selection and its application to covariate shift adaptation. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada 2008.

[12] W. Fan, I. Davidson, B. Zadrozny, and P. S. Yu, An improved categorization of classifiers sensitivity on sample selection bias. In: Proceedings of the 5th IEEE International Conference on Data Mining 2005.

[13] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, Self-taught learning: Transfer learning from unlabeled data. In Proceedings of the 24th International Conference on Machine Learning, Corvalis, Oregon, USA, pp. 759–766. 2007.

[14] W. Dai, G. Xue, Q. Yang and Y. Yu, Co-clustering based classification for out-of-domain documents. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, 2007.

[15] R. K. Ando, and T. Zhang, A high-performance semi-supervised learning method for text chunking. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, pp. 1–9. 2005.

[16] J. Blitzer, R. McDonald, and F. Pereira, Domain adaptation with structural correspondence learning. In Proceedings of the Conference on Empirical Methods in Natural Language, Sydney, Australia, pp. 120–128. 2006.

[17] H. Daume, Frustratingly easy domain adaptation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 256–263. 2007.

[18] A. Argyriou, T. Evgeniou, and M. Pontil, Multi-task feature learning. In Proceedings of the 19th Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, pp. 41–48. 2007.

[19] A. Argyriou, A. Micchelli, M. Pontil, and Y. Ying, A spectral regularization framework for multi-task structure learning. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, pp. 25–32. 2008.

[20] S.I. Lee, V. Chatalbashev, D. Vickrey and D. Koller, Learning a meta-level prior for feature relevance from multiple related tasks. In: Proceedings of the 24th International Conference on Machine Learning. Corvalis, Oregon, USA: ACM, pp. 489–496. 2007.

[21] T. Jebara, Multi-task feature and kernel selection for svms. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Alberta, Canada: ACM. 2004.

[22] C. Wang and S. Mahadevan, Manifold alignment using procrustes analysis. In Proceedings of the 25th International Conference on Machine learning. Helsinki, Finland: ACM, pp. 1120–1127. 2008.

[23] N. D. Lawrence, and J. C. Platt, Learning to learn with the informative vector machine. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Alberta, Canada: ACM. 2004.

[24] E. Bonilla, K. M. Chai and C. Williams, Multi-task gaussian process prediction. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, pp. 153–160. 2008.

[25] A. Schwaighofer, V. Tresp, and K. Yu, Learning gaussian process kernels via hierarchical bayes. In Proceedings of the 17th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, pp. 1209–1216. 2005.

[26] T. Evgeniou and M. Pontil, Regularized multi-task learning. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington, USA: ACM, pp. 109–117. 2004.

[27] J. Gao, W. Fan, J. Jiang, and J. Han, Knowledge transfer via multiple model local structure mapping. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada: ACM, pp. 283–291. 2008.

[28] L. Mihalkova, T. Huynh, and R. J. Mooney, Mapping and revising markov logic networks for transfer learning. In Proceedings of the 22nd AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada, pp. 608–614. 2007.

[29] A. Arnold, R. Nallapati, and W. W. Cohen, A comparative study of methods for transductive transfer learning. In Proceedings of the 7th IEEE International Conference on Data Mining Workshops. Washington, DC, USA: IEEE Computer Society, pp. 77–82. 2007.

[30] B. Zadrozny, Learning and evaluating classifiers under sample selection bias. In Proceedings of the 21st International Conference on Machine Learning, Banff, Alberta, Canada. 2004.

[31] J. Blitzer, R. McDonald, and F. Pereira, Domain adaptation with structural correspondence learning. In Proceedings of the Conference on Empirical Methods in Natural Language, Sydney, Australia, pp. 120–128. 2006.

[32] W. Dai, Q. Yang, G. Xue, and Y. Yu, Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning, Corvalis, Oregon, USA, pp. 193–200. 2007.

[33] A. K. Gretton, M. Borgwardt, B. Schoelkopf, and A. Smola, A Kernel Method for the Two-Sample-Problem. In Advances in Neural Information Processing Systems. 2006.

[34] R. Rifkin and A. Klautau, "In Defense of One-vs-All Classification," J. Machine Learning Research, vol. 5, pp. 101–141. 2004.

[35] JAFFE Database. [Online]. Available: http://www.kasrl.org/jaffe.html. time accessed: May 2013.

[36] A. Martinez, and R. Benavente, The AR face database. CVC Technical Report 24, Purdue University 1998.

[37] I. Buciu, C. Kotropoulos and I. Pitas, ICA and Gabor Representation for Facial Expression Recognition. In: Proceedings of International Conference on Image Processing, pp. 855-C858 2003.

[38] L.L. Huang, A. Shimizu, and H. Kobatake, Classification-based face detection using gabor filter features. In Sixth IEEE International Conference on Automatic Face and Gesture Recognition Seoul, Korea, pages 397–402. 2004.

[39] V. Kyrki, J.K. Kamarainen, and H. Kalviainen, "Simple gabor feature space for invariant object recognition," Pattern Recognition Letters, 25(3):31 1–318. 2004.

[40] T. S. Lee, "Image representation using 2d gabor wavelets." IEEE Transaction on Pattern Analysis Machine Intelligent, 18(10):959–971. 1996.

[41] C. Liu, and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminate model for face recognition," IEEE Transactions on Image Processing, 11(4):467–476. 2002.

[42] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, Coding facial expressions with gabor wavelets. In Third International Conference on Face & Gesture Recognition, Nara, Japan, pages 200–205. 1998.

[43] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, Comparison Between Geometry-based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perception. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara Japan. 1998.

[44] J.G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two dimensional visual cortical filters," Journal of the Optical Society of America A, vol. 2, pp. 1160–1169. 1985.

[45] I. Steinwart "On the influence of the kernel on the consistency of support vector machines," Journal of Machine Learning Research. 2:67–93. 2002.

[46] B.M. Levitan, Hilbert space, in Hazewinkel, Michiel, Encyclopaedia of Mathematics, Kluwer Academic Publishers, ISBN 978–1556080104.2001.