Exploiting Competition Relationship for Robust Visual Recognition

Liang Du **Haibin Ling**

Center for Data Analytics and Biomedical Informatics
Department of Computer and Information Science
Temple University
Philadelphia, PA, 19122, USA
{liang.du, hbling}@temple.edu

Abstract

Joint learning of similar tasks has been a popular trend in visual recognition and proven to be beneficial. Between-task similarity often provides useful cues, such as feature sharing, for learning visual classifiers. By contrast, the competition relationship between visual recognition tasks (*e.g.*, content independent writer identification and handwriting recognition) remains largely under-explored. A key challenge in visual recognition is to select the most discriminating features and remove irrelevant features related to intra-class variations. With the help of auxiliary competing tasks, we can identify such features within a joint learning model exploiting the competition relationship. Motivated by this intuition, we propose a novel way to exploit competition relationship for solving visual recognition problems. Specifically, given a target task and its competing tasks, we jointly model them by a generalized additive regression model with a competition constraint. This constraint effectively discourages choosing of irrelevant features (weak learners) that support the auxiliary competing tasks. We name the proposed algorithm *CompBoost*. In our study, *CompBoost* is applied to two visual recognition applications: (1) content-independent writer identification from handwriting scripts by exploiting competing tasks of handwriting recognition, and (2) actor-independent facial expression recognition by exploiting competing tasks of face recognition. In both experiments our approach demonstrates promising performance gains by exploiting the between-task competition.

Introduction

Extensive studies have proved that sharing features between related visual recognition tasks is helpful (Torralba, Murphy, and Freeman 2007; Pan and Yang 2010). A basic assumption underlying these methods is that the tasks are positively correlated and can be learned in a synergic way. For example, the useful features might be shared between different tasks/classes, and regularizing classifiers to favor features used by other tasks might be beneficial (Torralba, Murphy, and Freeman 2007). Following this idea, a rich body of work has been done in recent years (Shalev-Shwartz, Wexler, and Shashua 2011; Argyriou, Evgeniou, and Pontil 2008; Torralba, Murphy, and Freeman 2007; Hwang, Sha, and Grauman 2011; Yao and Doretto 2010; Wang, Zhang, and Zhang 2009). By contrast, the role of competition relationship between tasks has received insufficient attention.

Roughly speaking, we say that tasks are competing with each other if there are competitions or conflicts between their goals. Such competitions are often reflected in feature selection for these tasks: features favored by different competing tasks are likely to be exclusive. To motivate our discussion, we use *content-independent writer identification* (CIWI) and *handwriting recognition* (HR) as an example. The objective of CIWI is to determine the identity of a person by handwriting script (Schomaker 2007). On the other hand, the objective of HR is to recognize the handwritten characters, regardless of the person who wrote it (Arica and Yarman-Vural 2001). Variances of different handwriting scripts can stem from both the writers' writing styles and the contents of the scripts. For CIWI, writing style is important factor while the script content is a distracting one. This observation is reversed for HR. We conjecture that it will be beneficial to exploit the competition relationship by encouraging feature exclusion rather than sharing to achieve robust visual recognition. Considering that tasks with competition relationship are universal, it is worthwhile to accommodate this intuition from an algorithmic perspective.

In this paper, we develop a general algorithm to utilize the between-task competition relationship for visual recognition tasks. We name the task we target on as the *target task* and its competing tasks as the *auxiliary task*. We jointly model the target and auxiliary tasks with a generalized additive regression model regularized by competition constraints. This model treats the feature selection as the weak learner (*i.e.*, base functions) selection problem, and thus provides a mechanism to improve feature filtering guided by task competition. More specifically, following a stepwise optimization scheme, we iteratively add a new weak learner that balances between the gain for the target task and the inhibition on the auxiliary ones. We call the proposed algorithm *CompBoost*, since it shares similar structures with the popular AdaBoost algorithm.

The proposed CompBoost algorithm can be applied

for various visual recognition tasks. In this paper we use two test beds for evaluation: (1) content-independent writer identification by exploiting competing tasks of handwriting recognition, and (2) actor-independent facial expression recognition by exploiting competing tasks of face recognition. In the experiments for both applications, our approach demonstrates promising performance gains by exploiting the between-task competition relationship. We will release the source code for the CompBoost algorithm at http://www.dabi.temple.edu/∼hbling/code/competing-task.htm.

The rest of the paper is organized as follows. In the next section, we review related works. After that, the setting and notations of the problem are presented. The details of the proposed algorithm are presented in the section afterwards, followed by the description of the experimental validation. Finally, we conclude this paper in the last section.

Related Works

Considering the increasing number of visual concepts needed to be recognized, jointly learning multiple prediction tasks has gained popularity in visual recognition recently (Torralba, Murphy, and Freeman 2007; Hwang, Sha, and Grauman 2011; Yao and Doretto 2010; Wang, Zhang, and Zhang 2009). In (Torralba, Murphy, and Freeman 2007), a method is proposed to encourage feature sharing across object classes and/or views. In (Hwang, Sha, and Grauman 2011), sharing common sparsity patterns across objects and their attributes is investigated for visual recognition. In (Yao and Doretto 2010), multiple sources of visual data are jointly learned by sharing data instances or weak learners in a boosting framework. In (Wang, Zhang, and Zhang 2009), face verification classifiers for multiple people are jointly learned by sharing a few boosting classifiers in order to avoid over-fitting. General studies on joint multi-task learning (Caruana 1997) have been a popular topic in machine learning, and a survey can be found in (Pan and Yang 2010).

Our study, while modeling multiple visual recognition tasks jointly, is different from the aforementioned works. The key innovation in our study is to address the task competition that is totally different from the task similarity studied before. More specifically, instead of seeking feature sharing among tasks, our method takes advantage of feature exclusion among tasks during the learning process.

There are a few works employing feature exclusion (Zhou, Jin, and Hoi 2010; Hwang, Grauman, and Sha 2011; Xiao, Zhou, and Wu 2011). However, our work differs from them in many aspects. Firstly, the underlying motivation and task competition is different. These papers explore feature exclusion in discriminating hierarchical category taxonomy, that is, in a top-down taxonomy, when distinguishing subclasses within the same superclass, the features used for the superclass should not be useful for subclasses. Secondly, the methods proposed in the three papers use linear models for feature exclusion. By contrast, our solution uses nonlinear models, which are more suitable for visual recognition. Finally, the visual recognition tasks studied in our paper are different than those in the three papers. We will provide more detailed discussion in future revision.

A recent work in (Romera-Paredes et al. 2012) also exploits feature exclusion between tasks, in which linear models for multiple tasks are jointly learned with an orthogonal regularization between model coefficient vectors of unrelated tasks. Our study differs from the work in two major aspects: (1) The work in (Romera-Paredes et al. 2012) focuses on *uncorrelatedness* while ours on *negatively correlation*. Consequently, different regularization terms are used. (2) Linear models are used in (Romera-Paredes et al. 2012) for classification, while non-linear ones in ours. Note that the introduction of nonlinearity is very important for many visual recognition tasks where observations often lie in a highly nonlinear spaces. In fact, as illustrated in the experiments, our approach demonstrate significant improvements over (Romera-Paredes et al. 2012) in the face expression recognition task.

The two visual recognition test beds used in our study are both important topics in computer vision. Content-independent writer identification from handwriting scripts is to determine the identity of a person by his/her handwritten script (Schomaker 2007). It is desirable to identify factors related to writing styles and avoid interferences of factors related to writing contents. On the contrary, for the task of handwriting recognition, tremendous efforts have been made in avoiding interferences from factors related to writers' writing styles (Zhang and Liu 2013). To the best of our knowledge, the two tasks have never been coupled together like in our study.

Facial expression recognition aims to determine facial emotion from an input image or video. A main challenge in the problem is the variability of facial image across individuals (Zeng et al. 2009). This makes the problem naturally conflicting with a widely studied problem: face recognition (Zhao et al. 2003). Again, these competing tasks have never been jointly modeled in previous studies.

Our study is the first attempt, for both problems, to exploit the competition priors for improving their solutions. In other words, our approach treats the competition relationship between visual recognition tasks as blessings rather than curses. We expect the idea can be generalized to more visual recognition tasks in the future.

Problem Setting

Though the competition relationship are symmetric between competing tasks, in practice we usually focus on one task, hereafter referred as *target task*, while treating its competitors as *auxiliary tasks*.

For visual recognition, we write the target classification function to be learned as $F(\mathbf{x}) : \Omega \rightarrow \mathcal{L}$, such that $\Omega \subset \mathrm{R}^p$ denotes the p-dimensional feature space and $\mathcal{L}$ denotes the label set which is set as $\mathcal{L} = \{-1, +1\}$ by default. In other words, $F(\mathbf{x})$ predicts the label given an input feature vector $\mathbf{x}$. Our goal is to learn $F(\cdot)$ given the training samples.

In the following we denote the n_T training samples for the target task by $\mathcal{D}^{(t)} = \{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_T}$, in which $\mathbf{x}_i^{(t)} \in \Omega$ is a sample feature vector and $y_i^{(t)} \in \mathcal{L}$ its corresponding label. Similarly, we denote the N auxiliary training sets by $\mathcal{D}^{(1)}, \cdots, \mathcal{D}^{(N)}$, such that $\mathcal{D}^{(a)} = \{(\mathbf{x}_i^{(a)}, y_i^{(a)})\}_{i=1}^{n_a}$, where

n_a is the number of training samples for the a-th auxiliary task. Note that the feature space Ω for the target task is used for the auxiliary tasks as well. This may not be "fair" for the auxiliary tasks, but it is strategically assumed since our true interest is the target task.

Boosting by Competition Relationship

In this section, we present the formulation and derivation of the proposed CompBoost algorithm. We first introduce some notations used in the derivation. $\mathbb{E}$ denotes the expectation operation. $\mathbb{E}^{(a)}$ and $\mathbb{E}^{(t)}$ denote the expectation operation of the a-th auxiliary task (dataset) and target task (dataset) respectively. $\mathbb{E}_{\mathbf{w}}$ denotes the expectation operation according to the weighted samples with weight vector $\mathbf{w}$. In practice, the expectation is approximated by the empirical expectation over the training data.

Generalized additive regression model

Since our joint learning is built upon the generalized additive regression model, we first give a brief introduction of the model. Considering a regression problem mapping an input feature vector $\mathbf{x} \in \Omega$ to the output label $y \in \mathbb{R}$, and $\mathbf{x}, y$ have a joint distribution. We want to model the mean $\mathbb{E}[y|\mathbf{x}]$ by a function[1] $F : \Omega \to \mathbb{R}$. The generalized additive model has the form:

$$F(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x}),$$

where f_m's are the base predictors or weak learners and M is the number of base predictors selected for the regression. A back-fitting algorithm can be used to select weak learners to fit this model:

$$f_m(\mathbf{x}) \leftarrow \mathbb{E}[y - \sum_{k \neq m} f_k(\mathbf{x})|\mathbf{x}],\ m = 1, 2, \cdots, M. \quad (1)$$

Theoretically, any methods for function estimation can be used to estimate the conditional expectation in (1). In practice, by restricting $f_m(\mathbf{x})$ to be a simple parameterized function, one can solve for an optimal set of parameters through generalized back-fitting algorithms.

Additive models can be used to fit any form of functions (Hastie and Tibshirani 1990). In (Friedman, Hastie, and Tibshirani 2000), the classic AdaBoost algorithm is viewed as building an additive logistic model by minimizing loss function $\mathbb{E}[\exp(-yF(\mathbf{x}))]$. Our method follows this derivation but learns the target task classifier with regularizers that penalize the sharing of weak learners with auxiliary tasks.

Boosting by competition relationship

Now we formulate the proposed method as an additive modeling of exponential loss with constraints over competing auxiliary tasks.

For the target task, we define the the exponential loss for the target task as

$$\mathcal{E}_{\text{tar}}(F) = \mathbb{E}^{(t)}[e^{-yF(\mathbf{x})}]. \quad (2)$$

[1] $\mathcal{L}$ is relaxed to the real domain here.

To exploit the competing factors between the target task and its competing counterparts, we penalize those classifiers that perform well on both auxiliary and target tasks. For this purpose we introduce the following regularization term:

$$\mathcal{E}_{\text{aux}}(F) = \sum_{a=1}^{N} \mathbb{E}^{(a)}[e^{yF(\mathbf{x})}]. \quad (3)$$

Obviously, the more correct samples F predicts on the auxiliary tasks, the larger $\mathcal{E}_{\text{aux}}(F)$. This property implies that, in the minimization problem, the regularization term discourages F from performing well on the auxiliary tasks, which have competition relationship with the target task.

Combining $\mathcal{E}_{\text{tar}}$ and $\mathcal{E}_{\text{aux}}$, our regularized objective function is

$$\mathcal{E}(F) = \mathcal{E}_{\text{tar}}(F) + \lambda \mathcal{E}_{\text{aux}}(F), \quad (4)$$

where $\lambda > 0$ is a parameter to balance the effects of the two terms.

Following the derivation in (Friedman, Hastie, and Tibshirani 2000), we resort to a generalized additive regression model for the classification function. Consequently, F takes the form of $F(\mathbf{x}) = \sum_i f_i(\mathbf{x})$. The optimization is done by iteratively adding a weak learner f to current estimation of F.

In each iteration, suppose that we have a current estimate $\widehat{F}$ and are seeking for an improved estimate in the form of $\widehat{F} + f$. The cost function has the following form:

$$\begin{aligned} \mathcal{E}(\widehat{F} + f) = &\ \mathbb{E}^{(t)}[e^{-y(\widehat{F}(\mathbf{x}) + f(\mathbf{x}))}] \\ + &\ \lambda \sum_{a=1}^{N} \mathbb{E}^{(a)}[e^{y(\widehat{F}(\mathbf{x}) + f(\mathbf{x}))}]. \end{aligned} \quad (5)$$

While (5) entails expectation over the joint distributions of y and $\mathbf{x}$, it has been shown (Friedman, Hastie, and Tibshirani 2000) to be sufficient to minimize the criterion conditioned on $\mathbf{x}$, *i.e.*,

$$\begin{aligned} \mathcal{E}(\widehat{F} + f|\mathbf{x}) = &\ \mathbb{E}^{(t)}[e^{-y(\widehat{F}(\mathbf{x}) + f(\mathbf{x}))}|\mathbf{x}] \\ + &\ \lambda \sum_{a=1}^{N} \mathbb{E}^{(a)}[e^{y(\widehat{F}(\mathbf{x}) + f(\mathbf{x}))}|\mathbf{x}]. \end{aligned} \quad (6)$$

Therefore, our goal boils down to find f to minimize $\mathcal{E}(\widehat{F} + f|\mathbf{x})$, which can be expanded as

$$\begin{aligned} \mathcal{E}(\widehat{F} + f|\mathbf{x}) = &\ e^{-f(\mathbf{x})}\mathbb{E}^{(t)}[e^{-y\widehat{F}(\mathbf{x})}\mathbf{1}_{[y=1]}|\mathbf{x}] \\ + &\ e^{f(\mathbf{x})}\mathbb{E}^{(t)}[e^{-y\widehat{F}(\mathbf{x})}\mathbf{1}_{[y=-1]}|\mathbf{x}] \\ + &\ \lambda \sum_{a=1}^{N} e^{f(\mathbf{x})}\mathbb{E}^{(a)}[e^{y\widehat{F}(\mathbf{x})}\mathbf{1}_{[y=1]}|\mathbf{x}] \\ + &\ \lambda \sum_{a=1}^{N} e^{-f(\mathbf{x})}\mathbb{E}^{(a)}[e^{y\widehat{F}(\mathbf{x})}\mathbf{1}_{[y=-1]}|\mathbf{x}], \end{aligned} \quad (7)$$

where $\mathbf{1}_A = \begin{cases} 1, & \text{if } A \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$ is the indicator function.

Dividing (7) by $\mathbb{E}^{(t)}[e^{-y\widehat{F}(\mathbf{x})}|\mathbf{x}]$, we have the following minimization goal:

$$\begin{aligned} \mathcal{E}'(f) \triangleq &\ e^{-f(\mathbf{x})}\Big(\gamma^{+}(\mathbf{x}) + \lambda \frac{\eta^{-}(\mathbf{x})}{\mathbb{E}^{(t)}[e^{-y\widehat{F}(\mathbf{x})}|\mathbf{x}]}\Big) \\ &+ e^{f(\mathbf{x})}\Big(\gamma^{-}(\mathbf{x}) + \lambda \frac{\eta^{+}(\mathbf{x})}{\mathbb{E}^{(t)}[e^{-y\widehat{F}(\mathbf{x})}|\mathbf{x}]}\Big), \end{aligned}$$

where

$$\begin{aligned}\gamma^+(\mathbf{x}) &= \mathbb{E}^{(t)}[\mathbf{1}_{[y=1]}|\mathbf{x}],\\ \gamma^-(\mathbf{x}) &= \mathbb{E}^{(t)}[\mathbf{1}_{[y=-1]}|\mathbf{x}],\\ \eta^+(\mathbf{x}) &= \textstyle\sum_{a=1}^{N}\mathbb{E}^{(a)}[e^{y\widehat{F}(\mathbf{x})}\mathbf{1}_{[y=1]}|\mathbf{x}],\\ \eta^-(\mathbf{x}) &= \textstyle\sum_{a=1}^{N}\mathbb{E}^{(a)}[e^{y\widehat{F}(\mathbf{x})}\mathbf{1}_{[y=-1]}|\mathbf{x}].\end{aligned}$$

The boosting procedure can be interpreted as gradient descent in the functional space by using the functional derivatives and gradients (Frigyik, Srivastava, and Gupta. 2008). Set the derivative of $\mathcal{E}'$ w.r.t. $f(\mathbf{x})$ to zero, we get:

$$f(\mathbf{x}) = \frac{1}{2}\log\left(\frac{\gamma^+_{\mathbf{w}}(\mathbf{x}) + \lambda\eta^-_{\mathbf{w}}(\mathbf{x})}{\gamma^-_{\mathbf{w}}(\mathbf{x}) + \lambda\eta^+_{\mathbf{w}}(\mathbf{x})}\right), \tag{8}$$

where $\gamma^+_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}^{(t)}_{\mathbf{w}}[\mathbf{1}_{[y=1]}|\mathbf{x}]$, $\gamma^-_{\mathbf{w}}(\mathbf{x}) = \mathbb{E}^{(t)}_{\mathbf{w}}[\mathbf{1}_{[y=-1]}|\mathbf{x}]$ are the sample-weighted version of γ^+, γ^-; $\eta^+_{\mathbf{w}}, \eta^-_{\mathbf{w}}$ are the similar version of η^+, η^-; and the weight vector $\mathbf{w} = (w_1, w_2, \cdots, w_{n_T})^\top$ is defined by $w_i = \exp(-y_i F(\mathbf{x}_i))/A$, $i = 1, \cdots, n_T$, in which $A = \sum_{i=1}^{n_T}\exp(-y_i F(\mathbf{x}_i))$ Note that, while $\gamma^\pm_{\mathbf{w}}$ and $\eta^\pm_{\mathbf{w}}$ are originally defined based on conditional expectation, in practice we replace the expectation by corresponding conditional probability estimation.

Finally, a binary version of weak learner is achieved from $f(\mathbf{x})$ as

$$f'(\mathbf{x}) = \mathrm{sign}(f(\mathbf{x}))\ . \tag{9}$$

The above population version of weak learners can be flexibly replaced by a data version using trees and other classifiers. In our implementation, we use decision stumps as our weak learners such that $f(\mathbf{x})$ has the form $g(x_i, \theta) = \mathbf{1}_{[x_i > \theta]}, i \in \{1, \cdots, p\}$, in which $\theta \in \mathbb{R}$ is a thresholding parameter.

The above fitting process iterates until the preset maximum number of base functions are collected. The algorithm is summarized in Algorithm 1.

Discussion

If $\lambda = 0$, we have

$$f(\mathbf{x}) = \frac{1}{2}\log\left(\frac{\gamma^+_{\mathbf{w}}(\mathbf{x})}{\gamma^-_{\mathbf{w}}(\mathbf{x})}\right)\ .$$

This way the proposed *CompBoost* algorithm is reduced to the RealAdaboost algorithm (Friedman, Hastie, and Tibshirani 2000).

When $\lambda \neq 0$, the difference between the proposed *CompBoost* and the existing boosting methods is controlled by the terms $\eta^+_{\mathbf{w}}(\mathbf{x})$ and $\eta^-_{\mathbf{w}}(\mathbf{x})$. These two terms capture the competition relationship between the target task and the auxiliary tasks, and consequently render our algorithm in favor of weak classifiers that perform well for the target and bad on auxiliary tasks. This accords with our motivation that a classifier performs well for one task is unlikely to perform similarly for its competitors.

CompBoost differs from existing boosting-based methods only on the selection of weak learners during training, where additional overhead is to fit a weak learner for the auxiliary tasks. Therefore, the cost if (k+1)-times of the baseline boosting algorithm for k competing tasks. In the testing phase, the computational cost of our approach is the same as the baseline boosting algorithm.

Algorithm 1 CompBoost

Require: Target training dataset $\mathcal{D}^{(t)}$, auxiliary training datasets $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(N)}$, and the maximum number of iterations M.

1: Initialize the weight vectors $\mathbf{w}$ and $\mathbf{w}^{(a)}$:

$$\begin{aligned}w_i &= 1,\ i = 1, \cdots, n_T,\\ w_i^{(a)} &= 1,\ i = 1, \cdots, n_a; a = 1, \cdots, N\end{aligned}$$

2: **for** $t = 1, \cdots, M$ **do**

3: Normalize to 1 the target data weight vectors

$$\begin{aligned}\mathbf{w} &\leftarrow \mathbf{w}/\|\mathbf{w}\|_1,\\ \mathbf{w}^{(a)} &\leftarrow \mathbf{w}^{(a)}/\|\mathbf{w}^{(a)}\|_1,\ a = 1, \cdots, N\end{aligned}$$

4: Fit a classifier $f_t(\mathbf{x})$ according to (9)

5: Update the weight vectors

$$\begin{aligned}w_i &\leftarrow w_i e^{-y_i f_t(\mathbf{x}_i)},\ i = 1, ..., n_T.\\ w_i^{(a)} &\leftarrow w_i^{(a)} e^{y_i f_t(\mathbf{x}_i)},\ i = 1, ..., n_a; a = 1, ..., N.\end{aligned}$$

6: **end for**

7: **return** $F(\mathbf{x}) = \mathrm{sign}\big(\sum_{t=1}^{M} f_t(\mathbf{x})\big)$

Experimental Results

In this section, we evaluate the proposed method using two visual recognition applications: content-independent writer identification and facial expression recognition. In addition to the classic algorithms AdaBoost (Freund and Schapire 1997) and RealAdaBoost (Friedman, Hastie, and Tibshirani 2000), we also include two recently proposed jointly learning boosting variants briefly described below:

- **MultiSourceTrAdaBoost:** MultiSourceTrAdaBoost employs a mechanism that every weak learner is selected from ensemble classifiers learned from the auxiliary datasets which appears to be the most closely related to the target, at the current iteration (Yao and Doretto 2010). It is an extension of the seminal work of boosting based transfer learning TrAdaBoost (Dai et al. 2007).
- **TaskTrAdaBoost:** TaskTrAdaBoost is an instance of parameter-transfer approach which can be thought of as a task-transfer approach. During its learning phase, subtasks, coming from the various auxiliary tasks, can be reused, together with the target training instances(Yao and Doretto 2010). Instead of using the union of datasets for training weak learners as done in MultiSourceTrAdaBoost, TaskTrAdaBoost trains weak learners on source datasets and used as weak learner candidate pool for the target task.

We emphasize that MultiSourceTrAdaBoost and TaskTrAdaBoost are compared because they share similar jointly learning mechanism with us, but with totally different motivation. Our experimental results show that, when their assumptions on task similarity are violated, they even perform worse than baseline algorithms. In addition to the aforementioned methods, for expression recognition, we also include the results reported from (Romera-Paredes et al. 2012).

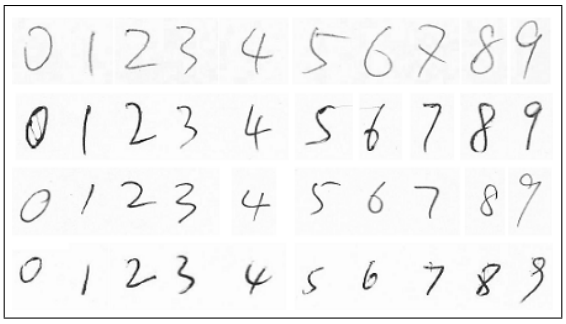

Figure 1: Handwriting samples from our handwriting dataset: each row shows sample images from one subject.

In all experiments, we use 100 weak learners in our algorithm, *i.e.*, $M = 100$. Note that although identical feature sets are used for both auxiliary and target tasks, they are virtually different tasks, since each task is associated with different labels. Training samples are randomly selected in all experiments. Five-fold cross validation strategy is applied in the training data to automatically determining the parameter λ in (4) in the candidate set of $\{\lambda = m \times 10^k : k \in \{-2, -1, 0\}, m \in \{1, 2, 5, 8\}\}$. Different numbers of training samples are tested.

Content-Independent Writer Identification Competing with Handwriting Recognition

Data set. We collected a dataset consisting of 14 writers' writing scripts. Each writer was asked to write digits from 0 to 9 ten times. This leads to a handwriting dataset with $1,400$ writing samples. Our experiments are based on the visual features extracted from these handwritten scripts. We regard classifying each writer from the rest as a binary target task. In addition, classifying each digit from the rest generates the ten auxiliary tasks. Note that here the auxiliary and target datasets are actually using the same images. However, since each task have different labels, they still can be regarded as different. Thus, the number of auxiliary datasets $N = 10$. Some writing samples are shown in Figure 1.

Experimental setup. The scripts in the dataset are first digitalized and then preprocessed as follows: each digit is manually segmented and normalized into an image of size 40×20. After that, PCA is applied to the data to reduce the dimension of the original data and the first 140 PCA coefficients (preserving 99% of the total energy) are used as features, denoted as PCA Coef. In addition, local binary patterns (LBP) (Ojala, Pietikainen, and Maenpaa 2002) ($116 = 58 \times 2$ dimensions, which is the combination of the upper and lower halves of the original images) and the projected features (60 dimensional) are also extracted from the images. The projected features, denoted by Projection, are obtained by projecting the character intensities along the horizontal and vertical axis respectively, which amount to a feature dimension of 60 ($= 40 + 20$ dimensions). In addition, the gray level cooccurrence matrix feature (GLCM) (Haralick, Shanmugam, and Dinstein 1973) is extracted (1024 dimensions). The concatenation of these features leads to a final feature vector of length 1340 ($= 1024 + 140 + 116 + 60$). We intentionally construct a large heterogeneous feature pool to test the ability of *CompBoost* in selecting the most relevant features for the target tasks (Results shown in Figure 2).

Table 1: Performances (%) on the writer identification tasks

Training samples	400	600	800	1000
MultiSourceTrAdaBoost	89.80	88.29	87.93	87.69
TaskTrAdaBoost	92.54	92.52	92.54	92.54
AdaBoost	92.20	92.75	93.04	93.31
RealAdaBoost	92.25	92.80	92.94	93.15
CompBoost	**93.09**	**93.39**	**93.73**	**93.83**

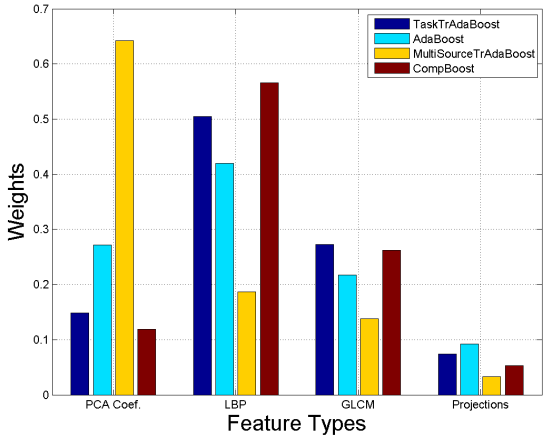

Figure 2: Distribution of selected features for writer identification.

Since for each task, the dataset is unbalanced, weights which are inversely proportional to the number of training instances are assigned as initial weights in the training process. For evaluation of experimental results, each experiment is randomly repeated 100 times and the mean accuracies are reported as results for comparison.

Results. The classification rates of writer identification using different approaches are summarized in Table 1. It shows that the proposed method consistently outperforms other boosting-based alternatives. Interestingly, we found that TaskTrAdaBoost and MultiSourceTrAdaBoost perform worse even than the AdaBoost baseline. This is because the task similarity assumption for them is directly violated in current scenario. Figure 2 shows the comparison of feature weights for each feature channel of the weak learners when using 1000 training samples. We can see that, for CompBoost, more LBP and GLCM features are selected. This could be explained by the fact that writer identification is determined by the writing styles (*e.g.*, stroke structure's statistics) whose features can be extracted by LBP and GLCM. In comparison, the other two features are more related to the writing content.

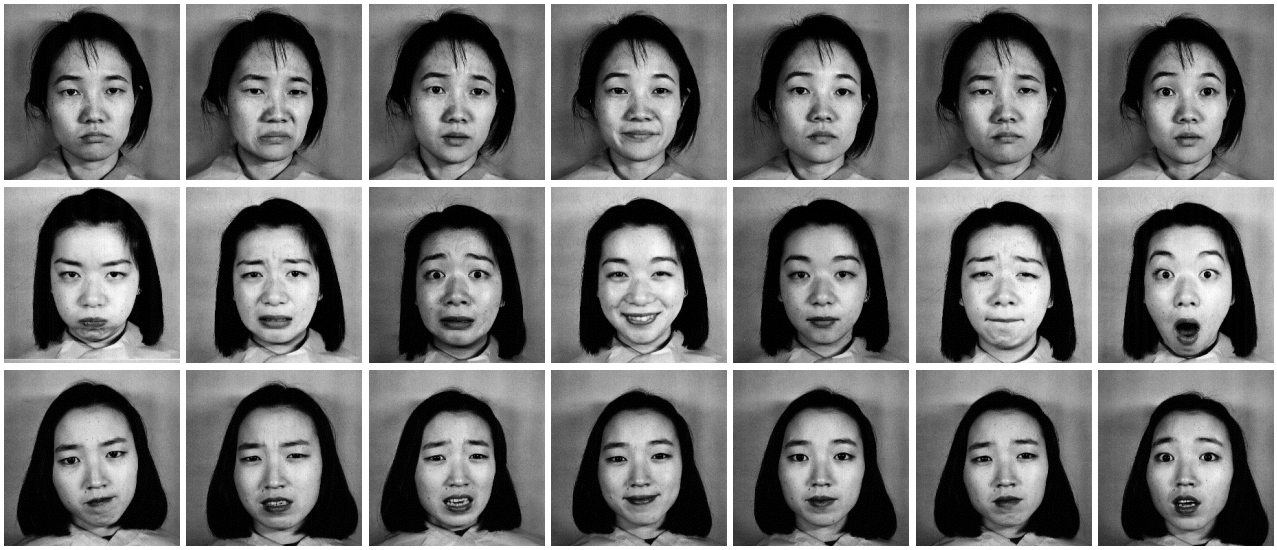

Figure 3: Sample images from JAFFE dataset: each row shows sample images of seven different expressions of a subject, with the expression labels shown in the bottom row.

Actor-Independent Facial Expression Recognition Competing with Face Recognition

Data set. For actor independent expression recognition, we use the Japanese Female Facial Expression (JAFFE) database (Lyons et al. 1998) as our testbed. The dataset is composed of 213 images of 10 subjects displaying seven mutually exclusive facial expressions, as illustrated in Figure 3. We treat classifying each expression from the rest as a binary target task and classifying each subject from the other subjects as a binary auxiliary task. Therefore, there are ten auxiliary tasks ($N = 10$).

Experimental setup. We follow the same experimental setup as in (Romera-Paredes et al. 2012) for fair comparison. For each face image, the face and eyes are first extracted by using the OpenCV implementation of the Viola-Jones face detector (Viola and Jones 2004). After that, we rotated the face so that the eyes are horizontally aligned. Finally, the face region is normalized to the size of 200×200. In order to obtain a descriptor of the textures of the image we used the Local Phase Quantization (LPQ) (Ojansivu and Heikkil 2008). Specifically, we divided every image into 5×5 non-overlapping grids. We computed the LPQ descriptor for each region and we created the image descriptor by concatenating all the LPQ descriptors. Finally, Principal Component Analysis is applied to extract component coefficients that retain 99% of the data variance energy. After the preprocessing, we obtained a descriptor of 203 dimensions for each image.

For evaluation, each experiment is randomly repeated 200 times and the mean accuracies are reported for comparison. Different numbers of training samples are tested.

Results. Table 2 shows the classification rates of different approaches for various numbers of training samples. We can see that the CompBoost performs the best compared with the other alternatives. Note that CompBoost outperforms the linear model with orthogonal regularization as in (Romera-Paredes et al. 2012)[2] by a large margin. We attribute the large improvement partly to the nonlinearity in CompBoost and partly to the exploitation of competition priors in our algorithm. This suggests that *CompBoost* is more suitable for visual recognition problems with nonlinear observations or representations.

Table 2: Performances (%) on expression recognition tasks

Training samples	60	80	100	120
MultiSourceTrAdaBoost	83.32	83.74	83.96	84.09
TaskTrAdaBoost	83.37	84.02	84.15	84.24
AdaBoost	84.71	84.97	85.70	86.02
RealAdaBoost	84.23	84.55	85.60	85.95
ORTHOMTL-EN (Romera-Paredes et al. 2012)	64.0	69.0	71.0	n/a
CompBoost	**85.52**	**86.00**	**86.92**	**87.04**

Conclusions

We have shown that exploiting between-task competition can be beneficial for robust visual recognition. The idea is implemented by harnessing a generalized additive regression model with a competition-regularization term, which inhibits weak learner (or feature) sharing between competing tasks. Experimental validations on content independent writer identification and actor independent facial expression recognition show the effectiveness of the proposed method. Since between-task competition exists in many visual recognition tasks, we expect the study to be broadly generalized to other applications in the future.

[2]Results reported in (Romera-Paredes et al. 2012) are error rates, we converted them into accuracies here.

Acknowledgement

We thank the anonymous reviewers for valuable comments and suggestions. This work is partly supported by the US NSF Grant IIS-1218156 and the US NSF CAREER Award IIS-1350521.

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Arica, N., and Yarman-Vural, F. 2001. An overview of character recognition focused on off-line handwriting. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 31(2):216–233.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Dai, W.; Yang, Q.; Xue, G.-R.; and Yu, Y. 2007. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, 193–200. New York, NY, USA: ACM.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1):119 – 139.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting. *Annals of Statistics* 28(2):337–407.

Frigyik, B.; Srivastava, S.; and Gupta., M. 2008. An introduction to functional derivatives. *Technical Report (University of Washington)*.

Haralick, R.; Shanmugam, K.; and Dinstein, I. 1973. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on* SMC-3(6):610–621.

Hastie, T., and Tibshirani, R. 1990. Generalized additive model. *Chapman Hall, London*.

Hwang, S. J.; Grauman, K.; and Sha, F. 2011. Learning a tree of metrics with disjoint visual features. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F. C. N.; and Weinberger, K. Q., eds., *NIPS*, 621–629.

Hwang, S. J.; Sha, F.; and Grauman, K. 2011. Sharing features between objects and their attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1761–1768.

Lyons, M.; Akamatsu, S.; Kamachi, M.; and Gyoba, J. 1998. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 200–205.

Ojala, T.; Pietikainen, M.; and Maenpaa, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24(7):971–987.

Ojansivu, V., and Heikkil, J. 2008. A method for blur and affine invariant object recognition using phase-only bispectrum. In Campilho, A., and Kamel, M., eds., *Image Analysis and Recognition*, volume 5112 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 527–536.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on* 22(10):1345–1359.

Romera-Paredes, B.; Argyriou, A.; Berthouze, N.; and Pontil, M. 2012. Exploiting unrelated tasks in multi-task learning. In Lawrence, N. D., and Girolami, M., eds., *AISTATS*, volume 22 of *JMLR Proceedings*, 951–959. JMLR.org.

Schomaker, L. 2007. Advances in writer identification and verification. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, 1268–1273.

Shalev-Shwartz, S.; Wexler, Y.; and Shashua, A. 2011. Shareboost: Efficient multiclass learning with feature sharing. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F. C. N.; and Weinberger, K. Q., eds., *NIPS*, 1179–1187.

Torralba, A.; Murphy, K.; and Freeman, W. 2007. Sharing visual features for multiclass and multiview object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29(5):854–869.

Viola, P., and Jones, M. 2004. Robust real-time face detection. *International Journal of Computer Vision* 57(2):137–154.

Wang, X.; Zhang, C.; and Zhang, Z. 2009. Boosted multi-task learning for face verification with applications to web image and video search. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 142–149.

Xiao, L.; Zhou, D.; and Wu, M. 2011. Hierarchical classification via orthogonal transfer. In Getoor, L., and Scheffer, T., eds., *ICML*, 801–808. Omnipress.

Yao, Y., and Doretto, G. 2010. Boosting for transfer learning with multiple sources. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 1855–1862.

Zeng, Z.; Pantic, M.; Roisman, G.; and Huang, T. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(1):39–58.

Zhang, X.-Y., and Liu, C.-L. 2013. Writer adaptation with style transfer mapping. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(7):1773–1787.

Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A. 2003. Face recognition: A literature survey. *ACM Comput. Surv.* 35(4):399–458.

Zhou, Y.; Jin, R.; and Hoi, S. C. H. 2010. Exclusive lasso for multi-task feature selection. In Teh, Y. W., and Titterington, D. M., eds., *AISTATS*, volume 9 of *JMLR Proceedings*, 988–995. JMLR.org.