Discriminant Additive Tangent Spaces for Object Recognition

Liang Xiong
Dept. of Automation
Tsinghua Univ., Beijing, China
xiongl@mails.thu.edu.cn

Jianguo Li
Intel China Research Center Ltd.
Beijing, China
jianguo.li@intel.com

Changshui Zhang
Dept. of Automation
Tsinghua Univ., Beijing, China
zcs@mails.thu.edu.cn

Abstract

Pattern variation is a major factor that affects the performance of recognition systems. In this paper, a novel manifold tangent modeling method called Discriminant Additive Tangent Spaces (DATS) is proposed for invariant pattern recognition. In DATS, intra-class variations for traditional tangent learning are called positive tangent samples. In addition, extra-class variations are introduced as negative tangent samples. We use log-odds *to measure the significance of samples being positive or negative, and then directly characterizes this log-odds using* generalized additive models *(GAM). This model is estimated to maximally discriminate positive and negative samples. Besides, since traditional GAM fitting algorithm can not handle the high dimensional data in visual recognition tasks, we also present an efficient, sparse solution for GAM estimation. The resulting DATS is a nonparametric discriminant model based on quite weak prior hypotheses, hence it can depict various pattern variations effectively. Experiments demonstrate the effectiveness of our method in several recognition tasks.*

1. Introduction

Recognizing objects is a basic task in computer vision and pattern recognition. Nevertheless, it is also a difficult one. Taking face recognition as an example, face images have inherent complex patterns and lie in high-dimensional feature spaces. Moreover, their features are very sensitive to the change of environment and the pose of subjects. In fact, pattern variation is one of the most critical factors that affects the performance of classifiers. People have been striving to develop invariant algorithms.

Manifold learning is very useful in developing such algorithms. Intuitively, small variations will not change a pattern's identity, and they lie "smoothly" in the feature space. Nonlinear dimensionality reduction methods like ISOMAP [22] and LLE [20] greatly help us inspect the structure of data. It is now commonly accepted that images of an object under changing conditions are on low-dimensional manifolds. This structural information can be used to derive effective invariant methods. Several manifold learning methods have been proposed in recent years, such as *Laplacian Eigenmaps* [3] and *manifold tangent* methods. Readers are referred to [4] for more details.

Manifold tangent are effective in describing the local structure of manifolds to facilitate classifications. Intuitively, manifold tangent refers to the direction in which the manifold lies, and this direction is represented by *tangent vectors*. As a forerunner, Simard *et al*. [21] proposed to use *tangent distance* as a new distance measure that is invariant to common transformations and achieved promising results in handwritten digit recognition. Hastie *et al*. [10] suggest the concept of *tangent subspace* as a compact representation of tangent vectors. In [13] Lee *et al*. presented a unified framework called *PTS* for tangent subspace learning. They proposed to approximate tangent vectors by pattern variations and model them by *Probabilistic PCA* [23]. Encouraging results were obtained using PTS.

However, existing manifold tangent learning methods only consider intra-class variations and ignore the extra-class information. In this paper, a discriminant model is proposed to integrate both of them. We first take the intra-class variations as the approximation of traditional tangent vectors, which we call positive tangent samples. In addition, extra-class variations are considered and called negative tangent samples. Provided both positive and negative samples, we use *log-odds* [5] to measure how significant a pattern variation lies in the tangent space of a reference

1-4244-1180-7/07/$25.00 ©2007 IEEE

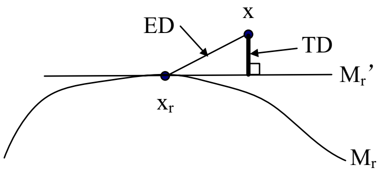


Figure 1. Tangent (TD) and Euclidean distance (ED). M_r is the reference manifold at x_r, and M_r' is its first order approximation.

sample. This log-odds is then characterized by an additive model which is trained to maximize its discriminant power. The proposed approach is thus called Discriminant Additive Tangent Space (DATS). DATS is a flexible model since it makes no hypothesis on the distribution of data, and estimates the model in a nonparametric way. Besides, the *curse of dimensionality* is seldom a problem in additive models.

Further, it is shown that this additive model for log-odds is essentially a *generalized additive model* (GAM [11]). In other words, we discriminatively model tangent spaces using GAM. When estimating the DATS model, we found that existing GAM fitting methods usually have difficulties in dealing with high-dimensional data. To solve this problem, we present a novel direct sparse fitting algorithm which is fast and memory efficient.

The rest of this paper is organized as follows. In section 2 we introduce tangent spaces and propose our discriminant additive tangent space model. Section 3 gives a detailed explanation on the direct sparse fitting procedure for GAM. Experiments are presented in section 4. Discussions are in section 5 and conclusions are drawn in section 6.

2. Additive Tangent Space

2.1. Tangent Methods Revisited

Letting x be a pattern and $t(x, \alpha)$ be an intra-class variation with parameter α, we have a class manifold $\mathcal{M}_x = \{t(x, \alpha)\}$ in the feature space. The tangent direction, represented by *tangent vector*, is $T = \partial t(x, \alpha)/\partial\alpha$. With above notions, Simard *et al.* [21] defined *tangent distance* (TD)

$$TD(x, x_r) = \min_{\alpha} \|x_r + \alpha T - x\| \tag{1}$$

where x is the test sample and x_r is the reference. An illustration of TD is shown in Figure 1. TD approximates the reference manifold $\mathcal{M}_r$ by its first order *Taylor expansion* and calculate the distance from x to M_r. In other words, TD is the distance from test samples to "reference classes" instead of single samples, thus it is insensitive to intra-class variations. However, it has to traverse all the tangent vectors so the computation load is high. TD has also been applied to other classifiers *e.g.* serves as a kernel for SVMs [19].

Traditional tangent algorithms rely on prior knowledge to obtain tangent vectors. For Optical Character Recognition (OCR) tasks, this knowledge is *affine transformations*. Yet other variations (*e.g.* change of perspective) are often very hard to model in prior. Lee *et al.* [13] proposed to use local pattern variations to approximate tangent vectors. These variations, which we call tangent samples, can be obtained by computing the difference between reference sample x_r and its intra-class neighbors, as:

$$T_r = \{t | t = x_i - x_r, c(x_i) = c(x_r), x_i \in \mathcal{N}_k(x_r)\} \tag{2}$$

where $c(x)$ is the label of x and $\mathcal{N}_k(x)$ is x's k nearest neighbors. In this paper we adopt the same approach to develop a general recognition algorithm.

Usually, obtained tangent vectors are redundant so a compact and effective representation is favorable. Hastie *et al.* [10] proposed the *tangent subspace* model, in which *singular value decomposition* (SVD) is performed on tangent vectors to reduce the dimensionality. Lee *et al.* [13] used *Probabilistic PCA* (PPCA [23]) to model the tangent vectors and construct a *probabilistic tangent subspace* (PTS). Then the projection residual of a pattern variation is calculated as TD. However, PTS assumes that the global tangent space can be depicted by a single Gaussian distribution. This is a rather strong assumption. Although they presented two kernel approaches to deal with complex cases, the model's clearance and efficiency are inevitably compromised. Thus a flexible and efficient method is desirable.

Another important point is that existing algorithms characterize manifold tangent only by tangent vectors that describe intra-class variations without considering samples from other classes. Put it another way, they are developed from a generative view. Thus, the quality of estimated model is heavily relied on the properness of model's hypothesis and the noise level.

2.2. Discriminant Additive Tangent Space

2.2.1 Discriminant Additive Modeling

Denote x_r as the reference sample. Tangent distance/subspace methods are essentially the same: they both calculate how far a test sample x_t is from x_r's tangent space. Traditionally, tangent spaces are derived using intra-class variations, which may be obtained either from knowledge or examples. However, in practice there are also extra-class samples near x_r. To obtain good performances, an algorithm should be able to discriminate intra-class and extra-class variations. Meanwhile, the latter can help model tangent spaces more precisely when noise exists. Therefore, we model the tangent space in a discriminative way incorporating both positive and negative tangent samples.

Denote t as a tangent sample obtained from the neighborhood of each reference sample, we assign positive tangent samples with label y=1, and negative tangent samples with y=0 *i.e.*

$$T^+ = \{(t, y = 1) | t = x_i - x_r, x_i \in \mathcal{N}(x_r), \text{if } c(x_i) = c(x_r)\}$$
$$T^- = \{(t, y = 0) | t = x_i - x_r, x_i \in \mathcal{N}(x_r), \text{if } c(x_i) \neq c(x_r)\}$$

where T^+ and T^- are positive tangent sample set and negative tangent sample set respectively.

Given a test sample x_t we need to determine whether x_t lies in the tangent space of x_r. In statistics, this yields the null hypothesis $H_0 : y = 1$. Denote $t = x_t - x_r$, the probability that H_0 is satisfied as $P(y = 1|t)$, and the probability that H_0 is violated as $P(y = 0|t) = 1 - P(y = 1|t)$. We adopt the *log-odds* [5] to measure how significant the null hypothesis is true:

$$\eta(t) = \log \frac{P(y = 1|t)}{P(y = 0|t)} = \log \frac{P(y = 1|t)}{1 - P(y = 1|t)}. \quad (3)$$

This significance actually serves as a probabilistic (inverse) distance measure. The larger η is, the higher possibility that x_t is in x_r's tangent space *i.e.* x_t and x_r belong to the same class.

To maximally discriminate intra-class and extra-class variations, our goal is to maximize the following objective:

$$\max \Big(\mathcal{L}(t) = \sum_{t_i \in T^+} \eta(t_i) - \sum_{t_j \in T^-} \eta(t_j) \Big) \quad (4)$$

Note that $\mathcal{L} = y\mathcal{L} + (1-y)\mathcal{L}$, it is easy to show that equation (4) is equivalent to

$$\max \prod_i P(y_i = 1|t_i)^{y_i} P(y_i = 0|t_i)^{1-y_i}, \quad (5)$$

which is in fact the maximization of the likelihood of *binomial* variables.

To achieve this goal, traditional *Bayesian* learning approaches need to estimate the posterior probabilities $P(y = 1|t)$ and $P(y = 0|t)$ or their corresponding conditional probabilities. In this paper, we directly characterize the *log-odds* $\eta(t)$, and assume that it follows an additive model

$$\eta(t) = \log \frac{P(y = 1|t)}{1 - P(y = 1|t)} = \sum_{j=1}^{p} f_j(t_j) + \epsilon \quad (6)$$

where ϵ is a Gaussian noise with zero mean, p is the number of features, t_j is the j-th feature of t with f_j as the corresponding base function, whose form can be arbitrary.

A closer look at (6) shows that it is in fact a well-known statistic regression model: *generalized additive model* (GAM [11]). From the perspective of GAM, y is the *response* with binomial distribution, η in (6) is the *systematic component* and the *logit* function $logit(\mu) = (\mu/(1 - \mu))$ is the *link*. Then this model becomes a GAM regression problem for the posterior probability of binomial distributed responses. The estimation of this GAM model will be further studied in section 3. Since this approach is based on GAM and import discriminative information, it is named Discriminant Additive Tangent Space (DATS). The DATS method is summarized in Algorithm 1.

Algorithm 1: Discriminant Additive Tangent Space

```
Input: Training set $Train = \{x_i\} \subset R^N$. Size of neighborhood $k$.
Training:
    • Construct the tangent sample sets $T^+$ and $T^-$.
    • Combine $T^+$ and $T^-$ to obtain $T$ as the tangent training set.
    • Estimate binomial GAM model on $T$.
Testing:
    • Given test sample $x_t$, compute $\eta(t) = \eta(x_t - x_r)$ at each reference $x_r$.
    • Adopt nearest neighbor rule for classification.
```

For simplicity, we assume that tangent spaces are globally homogeneous as in [13]. This assumption inevitably compromises the model's accuracy. But hopefully the flexibility of GAM will reduce the loss as possible. The advantage is that computation is reduced and there are more tangent samples available for estimation.

2.2.2 Discussion on DATS

DATS has many advantages in modeling tangent spaces. Here, four points should be emphasized.

- It is discriminant. Unlike traditional manifold tangent methods which only consider intra-class variations, DATS incorporate extra-class variations to model tangent spaces more precisely.
- It is flexible. DATS does not impose any assumption on conditional probabilities, and can model the log-odds in a nonparametric way. Hence it is capable of representing various distributions, and to some extent, pattern variations.
- It is resistant to *the curse of dimensionality*. Under additive assumption, samples are projected onto each single dimension in the link function space. Usually, the number of samples is plenty for one-dimensional estimations, so the growth of dimensionality has little impact on the requirement of training samples.
- It can easily be interpreted. Its additive form allows us to inspect the impact of each feature to the problem. This information can further be utilized to do feature selection and visualization (details can be found in section 4.1).

Other efforts have also been made on modeling manifolds discriminatively. Chen *et al.* [7] proposed *local discriminant embedding* (LDE), in which not only is the local data structure preserved, but also the extra-class samples are kept far away. This embedding strategy is more suitable for classifications. Impressive results are obtained by LDE.

3. Direct Sparse GAM

This section presents a GAM fitting algorithm for DATS which can handle high-dimensionality efficiently.

3.1. GAM Revisited

As mentioned in section 2.2, the estimation of DATS is in fact the fitting of a GAM model. First we give a brief review on GAM. Details can be found in [11].

A GAM model consists of 3 parts: a *random response*, a *systematic component*, and a *link function* $g(\cdot)$ linking the above two. The response y are assumed to follow the exponential family density [16]

$$\rho(y, \theta, \phi) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\}, \tag{7}$$

where θ is the natural parameter, ϕ is the scale parameter, and a,b,c are terms for different densities. Many familiar distributions, such as *Gaussian*, *Poisson* and *binomial*, belong to this density family.

For example, in the DATS model (6), the label y is the response with binomial distribution (5). $\eta(t)$ is the systematic component. The expectation of response $\mu = P(y = 1|t)$ is related to $\eta(t)$ by the *logit* link function $g(\mu) = logit(\mu)$, which is the *canonical link* for binomial distributions. The use of canonical link leads to lemma (1) [16].

Lemma 1 *When using canonical link in GAM, we have:*

(1) $\theta = \eta$*;*

(2) $\frac{\partial b(\theta)}{\partial \theta} = \mu = P(y = 1|t)$*;*

(3) $1/a(\phi) = \frac{\partial \mu}{\partial \theta}\frac{1}{\mathrm{var}(y)}$, *where* $\mathrm{var}(y)$ *is the variance of* y*.*

The GAM model is estimated by fitting the systematic component according to *maximum likelihood* criterion. Specifically, we want to estimate the base functions f_j to maximize the likelihood. Various algorithms have been proposed for this task. However, according to [24] and our own experiments, they all have difficulties in handling high-dimensional problems. Therefore we develop a new GAM fitting method to estimate DATS.

3.1.1 Base Functions

Various base functions can be used in GAM. Here we focus on two types of base functions which can be estimated directly without the use of *back-fitting* [11].

(1) Linear coefficients. This choice degrade the GAM model to *generalized linear models* (GLM [16]):

$$\eta(\mathbf{t}) = \sum_{j=1}^{p} \beta_j t_j = \beta^T \mathbf{t} + \epsilon \tag{8}$$

In this case, (6) becomes *logistic regression* which can be efficiently solved by *iterated reweighed least square* [16].

(2) *B-spline* is another popular kind of base function in GAM. Notations used here are as follows. n is the number of samples, m is the number of coefficients for each spline, and p is the number of features/dimensions. Usually, $i = 1, \cdots, n$ indexes samples, $j = 1, \cdots, p$ indexes features/dimensions/splines, and $k = 1, \cdots, m$ indexes splines basis functions. For the j-th spline f_j, let $B_k^j(t_{ij})$ be the *basis functions*, $\mathbf{Z}_j^T = [\mathbf{z}_{1j}, \cdots, \mathbf{z}_{nj}]$ with $\mathbf{z}_{ij}^T = [B_1^j(t_j), \cdots, B_m^j(t_j)]$ be the $n \times m$ *collocation matrix*, and γ_j be the coefficients. Then $\mathbf{f}_j = \mathbf{Z}_j\gamma_j$ and

$$\eta = \sum_{j=1}^{p} \mathbf{f}_j = \sum_{j=1}^{p} \mathbf{Z}_j\gamma_j = \mathbf{Z}\gamma \tag{9}$$

where $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_p]$, $\gamma^T = [\gamma_1^T, \cdots, \gamma_p^T]$. This base function is advantageous because it achieves nonparametric behavior while retains the parametric form that is easy to manipulate. Readers are referred to [8] for more details.

3.2. Direct Sparse Solution

We aim at developing a GAM fitting method that can process a large number of features effectively. Here we focus on B-spline GAM. Based on previous notations, the penalized negative log-likelihood (deviance) of model (6) is

$$l_p(\gamma) = -\sum_{i=1}^{n} [(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)] + \frac{\lambda}{2}\gamma^T \mathbf{\Lambda}\gamma \tag{10}$$

where $\mathbf{\Lambda} = diag(\mathbf{\Lambda}_1, \cdots, \mathbf{\Lambda}_p)$ is the *ridge* matrix to ensure smoothness, and λ indicates the penalty strength. Then, γ is estimated to minimize the deviance l_p.

We propose to use the *Newton descend* method to directly optimize l_p. Using Lemma 1, it can be proved that the gradient is (see Appendix)

$$\begin{aligned} \mathbf{s}(\gamma) = \tfrac{\partial l}{\partial \gamma} &= -\mathbf{Z}^T\mathbf{D}\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{u}) + \lambda\mathbf{\Lambda}\gamma \\ &= -\mathbf{Z}^T\mathbf{W}\mathbf{D}^{-1}(\mathbf{y} - \mathbf{u}) + \lambda\mathbf{\Lambda}\gamma. \end{aligned} \tag{11}$$

where $\mathbf{\Sigma} = diag(\mathrm{var}(y_i))$, $\mathbf{D} = diag(\frac{\partial \mu_i}{\partial \eta_i})$. And $\mathbf{W} = \mathbf{D}\mathbf{\Sigma}^{-1}\mathbf{D}$ acts as the weight for each sample. Further, the second order derivative (*Hessian*) is

$$\mathbf{H}(\gamma) = \mathbf{Z}^T\mathbf{W}\mathbf{Z} + \lambda\mathbf{\Lambda} \tag{12}$$

Then, the descend direction $\mathbf{d}$ can be derived by solving the linear system

$$\mathbf{H}\mathbf{d} = -\mathbf{s}. \tag{13}$$

It is easy to see that $\mathbf{H}$ is *positive definite* (Λ is positive definite to be a valid ridge term), so the optimization problem is convex and global optimum is guaranteed.

However, $\mathbf{H}$ has size $mp \times mp$ so solving (13) by regular methods is expensive ($O(m^3p^3)$). To overcome this problem, we exploit the special structure of $\mathbf{H}$ to achieve an efficient solution, which is based on two observations. First, $\mathbf{H}$ can be naturally partitioned into blocks $\{\mathbf{H}_{ab}\}$ with

$$\mathbf{H}_{ab} = \frac{\partial^2 l(\gamma)}{\partial \gamma_a \partial \gamma_b^T} = \begin{cases} \mathbf{Z}_a^T \mathbf{W} \mathbf{Z}_a + \lambda \mathbf{\Lambda}_a & a = b \\ \mathbf{Z}_a^T \mathbf{W} \mathbf{Z}_b & otherwise \end{cases} \tag{14}$$

where γ_a is the parameter of the a-th splines and $\mathbf{\Lambda}_a$ is the corresponding ridge matrix. And second, the diagonal blocks $\mathbf{H}_{jj}$ is dominant in $\mathbf{H}$ *i.e.* elements in $\mathbf{H}_{jj}$ is much larger than those in $\mathbf{H}_{ij}, i \neq j$. This is because that 1) the correlation between $\mathbf{Z}_a$ and itself is much stronger than $\mathbf{Z}_b, b \neq a$, 2) $\mathbf{W}$ becomes very small as the iteration goes while the ridge matrices $\mathbf{\Lambda}_a$ remain unchanged. Above observations implies that parameters of different splines in (13) are *weakly coupled* to each other.

To exploit this structure, we propose to solve (13) by *block Jacobi iteration* [9]

$$\mathbf{d}^{(l+1)} = \hat{\mathbf{H}}^{-1}\mathbf{E}\mathbf{d}^{(l)} - \hat{\mathbf{H}}^{-1}\mathbf{s} \tag{15}$$

where $\hat{\mathbf{H}} = diag(\mathbf{H}_{11}, \cdots, \mathbf{H}_{pp})$ is a block diagonal matrix extracted from $\mathbf{H}$, and $\mathbf{E} = \mathbf{H} - \hat{\mathbf{H}}$ is the residual. $\hat{\mathbf{H}}^{-1}$ can be calculated very fast using block-wise inverse, so the iteration can be calculated in $O(p^2)$ time. Besides, since elements in $\mathbf{E}$ are small compared to $\hat{\mathbf{H}}$, the above iteration is supposed to converge fast [9]. The complexity is now reduced from cubic to near-quadratic.

To further simplify, we assume that (15) converges to the true solution fast enough so one iteration is adequate, resulting the solution

$$\hat{\mathbf{d}} = -\hat{\mathbf{H}}^{-1}\mathbf{s}, \tag{16}$$

where $\hat{\mathbf{d}}$ is the approximate solution. With a little sacrifice of precision, the complexity is now only linear. (16) suggests that $\hat{\mathbf{H}}$ can be used as a sparse approximation of $\mathbf{H}$. Thus we call this approach *sparse Newton* and the corresponding GAM fitting algorithm *direct sparse GAM* (DS-GAM). Using $\hat{\mathbf{H}}$, spline parameters are *decoupled* so we can decompose (16) into p small problems

$$\mathbf{H}_{jj}\mathbf{d}_j = -\mathbf{s}_j, j = 1, \cdots, p \tag{17}$$

each of which can be solved in constant time. Further, these small problems can be dispatched to multiple processors to achieve parallel computation. The algorithm is summarized in Algorithm 2. In DATS, binomial distribution are employed so $\mathbf{D} = \mathbf{\Sigma} = diag(\mu_1(1-\mu_1), ..., \mu_n(1-\mu_n))$.

Note that $\hat{H}$ is also positive definite so the convergence is still guaranteed. The convergence speed is determined by $\mathbf{E}$: the smaller elements in $\mathbf{E}$ is, the faster the optimization converges. In practice the performance of this sparse solution is comparable to pure Newton. Some empirical performances are presented in section 4.2. Detailed analysis on its convergence can be found in [25].

Algorithm 2: Direct Sparse GAM (DS-GAM)

```
Input: Training samples $\mathbf{x}_i$ and responses $y$. Form of distribution and the link function. Ridge term $\Lambda$.
Initialize: splines' collocation matrices $\{\mathbf{Z}_j\}$, parameters $\{\gamma_\mathrm{j}\}$.
While not convergent do:
  1. Estimate descend direction.
     For j = 1, ..., p, compute update $\mathbf{d}_j$ for each spline
  2. Line-search for optimum step length $l$.
  3. Update parameters.
     For j = 1, ..., p, update $\gamma_j = \gamma_j + l\mathbf{d}_j$
  4. Check convergence.
```

3.3. Related Work

Solving large scale linear systems by approximate decomposition of weakly coupled matrices has been applied in various problems *e.g.* [1]. In fact, algorithms have been developed to turn a normal matrix into its weakly coupled form [25]. Usually in GAM the *Hessians* are inherently weakly coupled so decompositions can be applied directly to accelerate the computation.

We compare DS-GAM to two similar GAM fitting methods: *direct penalized GAM* (DP-GAM) [15] and *GAMBoost* (GB) [24]. DP-GAM proposed the usage of B-spline with penalized likelihood. Then it applies normal Newton method to estimate the model. The convergence of DP-GAM is very fast once the descend function is computed. However, the storage of DP-GAM is m^2p^2, and the computational cost of each iteration is $O(p^2n + p^3m)$). In practice, the number of coefficient per spline m is usually 10 or above. Thus the above two requirements become prohibitive for common computers when dealing with hundreds of features (images larger than 16 × 16). On the other hand, DS-GAM decomposes the model and treats each spline separately. With a little sacrifice of precision, the storage problem is eliminated and the computational cost is now $O(pn + pm)$) ($O(n + m)$ if using p processors). This improvement is significant when p is large.

GB was developed based on the idea of likelihood boosting. It also treats the base functions separately so runs fast. However, GB only update one of them each iteration. Essentially, GB is a *coordinate descend* method [6] with greedy direction search. Its convergence is much slower than Newton update. Besides, GB tends to focus on a few features while others are suppressed, so it is vulnerable to noises, especially when the training set is small.

In sum, DS-GAM integrates the optimality of DP-GAM and the efficiency of GAMBoost. Its computation of each iteration is significantly faster than DP-GAM and the descend speed is much higher than GB. In our experiments DS-GAM can handle images with thousands of features and always converges.

4. Experiments

To evaluate the performance of DATS in visual recognition tasks, experiments are conducted on several data sets. Specially, GLM models are applied to face data to gain an intuitive representation of tangent samples.

The following data sets are used in our experiments. They are all featured for their manifold structures.

- UMIST (cropped): A face data set of 20 persons. Each person has about 30 images with different poses. Images are resized to 28×23.
- COIL-20 (processed): This data set contains images of 20 objects. Each object has 72 images from different view points. Images are resized to 32×32.
- ORL: A face data set of 20 persons. Each person has 10 samples with different poses and expressions. Images are resized to 28×23.

Experiment settings are as follows. PCA (EigenFace) is used as the baseline. Linear Discriminant Analysis [12] (LDA, FisherFace [2]) is also used for comparison. To illustrate the effectiveness of DATS, results of PTS [13] are presented. As pointed out in [13], PTS is quite similar to Bayesian Face Model [18, 17] when the training set is small. In all experiments nearest neighbor classifiers are employed to test the effectiveness of different distance measures.

For DATS, we manually select the neighborhood size (between 3 and 7) to obtain tangent samples. The ridge penalty λ is roughly tuned by cross validation (we find that it has little impact on the performance in a rather large range). Each feature is normalized to 0-mean and 1-variance. DATS models are estimated by both GB and DS-GAM to compare their performance. 20 independent runs are performed. In each run, training samples are drawn randomly from each class, while the rest are used for testing, and training and testing data are kept the same for all algorithms. Mean performance of the 20 runs are reported.

4.1. Linear Model and Interpretation

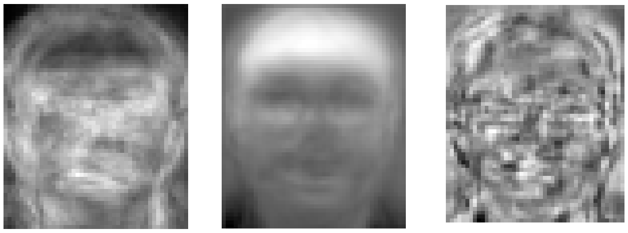
First, we use *generalized linear model* (GLM) described in (8) as a simplified alternative of B-spline GAM to model the tangent space of faces. We call this tangent space model *logistic discriminative tangent space* (LDTS). The advantage of LDTS is that its estimation is very fast and the model can easily be interpreted.

This experiment is conducted on the ORL face data. Performance of LDTS is compared to LDA and PTS(BFM). Mean performances are presented in Table 1.

Even with the loss of flexibility, promising result is obtained by LDTS. And if coefficients β of the derived LDTS models (see (8) for more details) are drawn as an image, it resembles a human face as shown in Figure 2. Since the coefficients lies in the log-odds space, we call this representation *LogFace*. LogFace is in fact an "typical" positive tangent sample for the data set, considering the fact that the correlation between LogFace and a test tangent sample determines the log-odds.

Training # per class	2	3	4	5
LDA	81.10	88.14	91.03	92.33
PTS	87.05	92.47	95.03	96.28
LDTS	86.75	92.14	95.39	97.24
DATS	**90.14**	**95.04**	**96.65**	**98.48**

Table 1. Recognition accuracy (%) on ORL data set.



LogFace EigenFace FisherFace

Figure 2. LogFace for ORL data set. It is actually an typical representation of positive tangent samples. EigenFace and FisherFace for the same data are also presented.

LDTS can also be used as a feature selection method. Features corresponding to small LDTS coefficients have little influence on the calculation of log-odds. Therefore they can be removed without much impact to the performance.

4.2. General Performance

Extensive experiments are conducted on UMIST, COIL and ORL. We use B-spline for DATS with 10 knots per spline. The penalty λ is 10. Results are presented in Figure 3, from which we can conclude that:

1. DATS using DS-GAM remarkably outperforms other methods on all the data sets. This result confirms the effectiveness of our discriminative and nonparametric modeling for manifold tangents.
2. PTS does not perform well on COIL data. This is probably because that COIL samples actually lie on a circle-shaped manifold [14]. Under the global homogeneous assumption, its tangents will most likely form a Gaussian model with no directional tendency. Therefore PTS is degraded to a naive nearest neighbor classifier in essence. On the other hand, results show that this complexity can be handled by DATS effectively.
3. DATS estimated by DS-GAM is superior to that estimated by GB. Typically in GB, after 100 iterations the models deviance is stable, with about 50 features utilized, while other features are discarded. This is not desired for visual tasks.

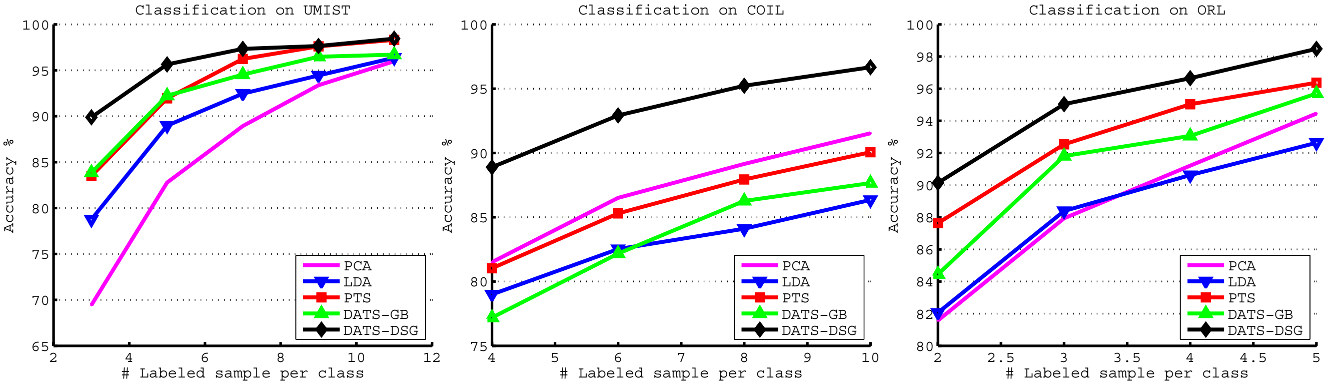


Figure 3. Experiment results of DATS on UMIST, COIL and ORL. DATS outperforms other algorithms remarkably. It is also shown that results obtained by DS-GAM are superior to those obtained by GAMBoost.

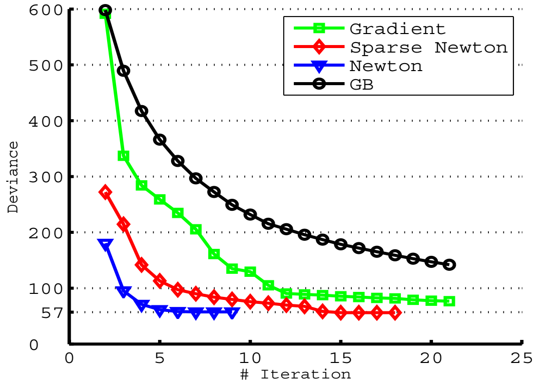


Figure 4. Convergence curves of optimization methods for DATS on COIL. For each iteration, the time consumption of *gradient*, *sparse Newton*, *GB* and *Newton* is about $1 : 2 : 2 : 90$ (implemented in Matlab$^{®}$).

Figure 4 shows the convergence speed of DATS with different optimization methods including *gradient descend*, *Newton* (DP-GAM), GB, and *sparse Newton* (DS-GAM). This experiment is conducted on COIL using 10 samples from each class as the training set. Images are resized to 16×16 (about 3000 variables in the GAM model) to make DP-GAM feasible. The convergence of GB is slow since it follows a coordinate descend strategy. DS-GAM descends very fast using near-optimum directions and the computation is very efficient. Although DP-GAM descends faster in each iteration, its computation is very slow (see Figure 4), and it rapidly grows infeasible as the dimensionality increases.

5. Discussion

5.1. Relaxation

The basic assumptions of DATS that tangent spaces are globally homogeneous can be relaxed to obtain stronger models. We can estimate separate DATS model for different samples or regions. In this way the performance is expected to be enhanced, while the drawback is that it will bring heavier computation burden and storage requirements.

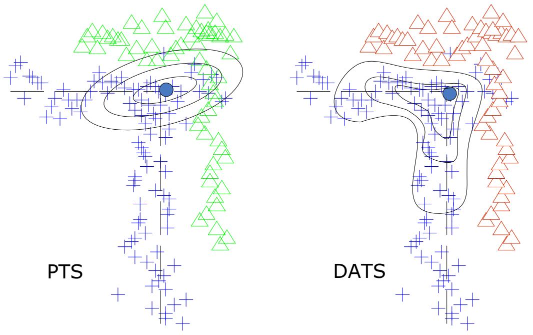


Figure 5. PTS and DATS for a noisy manifold. Gray circle is the reference sample. $+$ and Δ represent different classes. PTS is generative and ignores extra-class samples, while DATS considers them. PTS is Gaussian, while DATS is nonparametric. DATS can characterize this structure more precisely.

5.2. Comparison to PTS

From the perspective of GAM regression, DATS model (6) yields the posterior probability $P(y = 1|t)$ with logit link function. This notion coincides with PTS[13], which uses PPCA to estimate a density for tangent samples. Considering the fact that in general tangent samples are obtained by centralizing reference points' neighbors, both methods are actually modeling the local distribution of data. There are mainly two differences between them: (1) PTS is generative while DATS is discriminative. PTS aims at estimating a Gaussian model for positive tangent samples. On the other hand, the goal of DATS is to maximize the model's discriminative power. In training, negative tangent samples are utilized to obtain more accurate model. (2) PTS restricts the local distribution to be a single Gaussian while DATS assumes no prior knowledge. For complex manifold structures PTS is obviously not an adequate model, whereas DATS's nonparametric model allows for maximum adaptability. These two points are illustrated in Figure 5.

6. Conclusion

In this paper, we propose a novel manifold tangents modeling method called discriminant additive tangent space (DATS), and apply it to object recognition tasks as an invariant classification algorithm.

To model manifold tangents effectively, DATS propose to utilize both positive and negative tangent samples and measure the significance of their difference by log-odds. Then the log-odds are directly modeled by GAM which is optimized to maximize the discriminant power. In this way, a discriminant nonparametric model for manifold tangents is derived. In order to apply GAM to visual tasks, we also presented an efficient fitting algorithm called DS-GAM for high-dimensional tasks. This algorithm is highly efficient and possesses good convergence properties.

DATS is very flexible to model various pattern variations and is resistant to the curse of dimensionality. Experiments on several object recognition tasks demonstrate its effectiveness.

Acknowledgement

Funded by Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList)

Appendix

Using canonical link and Lemma (1), we have

$$\begin{aligned}
&\theta_i = \eta_i = \sum_j f_j(x_i) = \sum_j \mathbf{z}_{ij}^T \gamma_j, \\
&\mathbf{s}(\gamma_j) = \tfrac{\partial l_p}{\partial \gamma_j} = \tfrac{\partial l_p}{\partial \theta} \tfrac{\partial \theta}{\partial \eta} \tfrac{\partial \eta}{\partial \gamma_j} + \lambda \Lambda_j \gamma_j \\
&= -\sum_{i=1}^{n} [(y_i - \tfrac{\partial b(\theta_i)}{\partial \theta_i}) \tfrac{1}{a(\phi)} \tfrac{\partial \theta_i}{\partial \eta_i} \tfrac{\partial (\sum_{j=1}^{p} z_{ij}^T \gamma_j)}{\partial \gamma_j}] + \lambda \Lambda_j \gamma_j \\
&= -\sum_i [(y_i - u_i) \tfrac{\partial h(\eta_i)}{\partial \eta_i} \tfrac{1}{\mathrm{var}(y_i)} z_{ij}] + \lambda \Lambda_j \gamma_j \\
&= -Z_j^T \mathbf{D} \mathbf{\Sigma}^{-1} (y - \mu) + \lambda \Lambda_j \gamma_j
\end{aligned} \tag{18}$$

in which $h(\cdot)$ is the inverse function of the link $g(\cdot)$.

References

[1] M. Amano, A. I. Zecevic, and D. D. Siljak. An improved block-parallel newton method via epsilon decompositions for load-flow calculations. *IEEE Trans. Power Systems*, 11:885–895, 1996.

[2] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. PAMI*, 19(7):711–720, 1997.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[4] Y. Bengio and M. Monperrus. Non-local manifold tangent learning. In *Advances in NIPS-05*, pages 129–136, 2005.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge, UK, 2004.

[7] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *CVPR-05*, 2005.

[8] P. Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, Oxford, 1993.

[9] L. A. Hageman and D. M. Young. *Applied Iterative Methods*. Academic Press, San Diego., 1981.

[10] T. Hastie, P. Simard, and E. Sackinger. Learning prototype models for tangent distance. In *Advances in NIPS-95*, 1995.

[11] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[13] J. Lee, J. Wang, C. Zhang, and Z. Bian. Probabilistic tangent subspace: A unified view. In *ICML-04*, 2004.

[14] J. Lee, J. Wang, C. Zhang, and Z. Bian. Visual object recognition using probabilistic kernel subspace similarity. *Pattern Recognition*, 38(1):997–1008, 2005.

[15] B. D. Marx and P. H. Eilers. Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28:193–209, 1998.

[16] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, 1983.

[17] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans. PAMI*, 24:780 – 788, 2002.

[18] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. PAMI*, 19:696 – 710, 1997.

[19] A. Pozdnoukhov and S. Bengio. Tangent vector kernels for invariant image classification with svms. In *ICPR-04*, 2004.

[20] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[21] P. Simard, Y. L. Cun, and J. Denker. Efficient pattern recognition using a new transformation distance. In *Advances in NIPS-93*, pages 50–58, 1993.

[22] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2322, 2000.

[23] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, Series B, 61, Part 3:611–622, 1997.

[24] G. Tutz and H. Binder. Generalized additive modeling with implicit variable selection by likelihood based boosting. *Biometrics*, 62:961–971, 2006.

[25] A. I. Zecevic and D. D. Siljak. A block-parallel newton method via overlapping epsilon decompositions. *SIAM Journal on Matrix Analysis and Applications*, 15:824–844, 1994.