Scene Image Categorization and Video Event Detection using Naive Bayes Nearest Neighbor

Shiv N. Vitaladevuni, Pradeep Natarajan, Shuang Wu, Xiaodan Zhuang, Rohit Prasad and Premkumar Natarajan
Raytheon BBN Technologies
10 Moulton Street, Cambridge, MA 02138.
{svitalad, pradeepn, swu, xzhuang, rprasad, pnataraj}@bbn.com

Abstract

We present a detailed study of Naive Bayes Nearest Neighbor (NBNN) proposed by Boiman et al., with application to scene categorization and video event detection. Our study indicates that using Dense-SIFT along with dimensionality reduction using PCA enables NBNN to obtain state-of-the-art results. We demonstrate this on two tasks: (1) scene image categorization on the UIUC 8 Sports Events Image Dataset (obtaining 84.67%) and the MIT 67 Indoor Scene Image Dataset (obtaining 48.84%); and (2) detecting videos depicting certain events of interest on the challenging MED'11 video dataset with only 15 positive training videos per event. We present an extension referred to as sparse-NBNN that constrains the number of training images that can used to match with a given test image for the image-to-class distance computation. Experiments indicate that this improves upon NBNN for handling of imbalanced training data.

1. Introduction

Scene recognition deals with the task of analyzing an image or video's visual content to infer properties of the scene, with application such as image tagging, providing priors for object recognition, etc. We present a study on naive Bayes nearest neighbors (NBNN) that achieves state-of-the-art results on challenging image and video datasets, and an extension referred to as sparse-NBNN that constrains the number of training images that can used by NBNN for a given test image.

One of the most popular approaches to scene recognition is to use a bag-of-words model in which a set of features are extracted from a given image and compared with learned models of expected feature distributions in the scene categories of interest. The vast majority of bag-of-words based techniques involve three steps, coding to project feature vectors onto a codebook/dictionary, pooling to compute summary statistics of the projection coefficients, and learning classifiers to model the statistics for each category.

The coding-pooling-classification strategy requires careful design at each of these steps to tune for the task at hand, characteristics of the dataset and the type of features used. This is evidenced in the large number of studies published over the last few years in this line of research [5]. When shifting to a new dataset, e.g., learning categories on the fly, it is often not clear if the specific design of bag-of-words models being used is the best one available. Moreover, it is desirable to have a simple and reliable straw man that can be quickly applied to a new task, say on a new feature descriptor, to set a baseline for more complex approaches.

Boiman et al. proposed an approach referred to as naive Bayes nearest neighbor (NBNN) that uses an extremely simple formulation and still achieves good results for object categorization [2]. The approach relies on an image-to-class distance computed by directly comparing raw feature vectors of a test image with those of each category. Thus it avoids coding and pooling. The advantages include absence of parameter tuning and extremely simple training.

We show that with some modifications NBNN achieves state-of-the-art results in scene categorization. Relative to [2], we demonstrate NBNN on scene recognition and video analysis, and show that using principle component analysis (PCA) on Dense-SIFT (D-SIFT) improves performance. We test our approach on two standard datasets, UIUC 8 Sports Events Image dataset [10] and MIT 67 Indoor Scene dataset [15], and achieve state-of-the-art results compared to more complicated techniques. For the UIUC-8 dataset, we obtain 84.67% accuracy which is comparable to previously reported results of 84.4% [6]. On the MIT-67 dataset, we obtain 48.84% accuracy which is better than previously reported results of 47.01% [4]. To our knowledge, ours is the first study that applies NBNN to video analysis with state-of-the-art results. We demonstrate our

978-1-4673-5054-9/13/$31.00 ©2013 IEEE

approach on a challenging video event detection problem, namely the TRECVID MED'11 dataset, for detecting 10 video events. This task is especially challenging as there are only 15 positive training examples per event.

We also present a novel extension referred to as Sparse NBNN. The main motivation is that in vanilla NBNN, all training images from a category can contribute to "explain" a given test image. However, in certain applications it might be useful to restrict the number of training images that can be used for a given test image. Notice that two different test images can use different subsets of training images, but each subset can have at most k training images. For instance, if a particular category has disproportionatly large number of training examples, then the classifier tends to be biased towards this category. Our experiments indicate that using sparse-NBNN to constrain the number of training images that can match a test image reduces the effects of such data imbalance. Moreover, the selected training images that match with a test image can be used to transfer annotation/tags to the test image. We formulate sparse-NBNN as a weighted Set Cover problem. Experiments on the UIUC 8 Sports Scene dataset indicate improved classification performance.

Our study would be useful to the vision community in presenting a simple, almost monolithic, algorithm that achieves state-of-the-art in scene image and video categorization. For problems in which it achieves state-of-the-art results, NBNN allows researchers to focus on improving the features and the training dataset. As the community develops better algorithms and more challenging datasets, our study would continue to provide an easy to apply straw man as a baseline.

1.1. Related Work

Recent years have witnessed a growing research interest in scene recognition. Here, we review a limited set of papers that present results on the two scene image datasets used in our experiments; please refer to [25] and [4] for more thorough literature reviews.

Li and Fei-Fei presented an approach to classify images of sports events using scene and object recognition, and collected the UIUC 8 Sports Scene dataset (UIUC-8) [10]. Quattoni and Torralba put forth the MIT 67 Indoor Scene dataset (MIT-67) presented results using local and global features, e.g. Gist and bag-of-words in ROIs [15]. Wu and Rehg presented scene recognition using histogram of intersection kernel [24]. Li et al. described an approach to use object detection cues for scene recognition in [11]. Boureau et al. employed densely sampled SIFT features (D-SIFT) with sparse projections and spatial pooling for scene recognition in [3]. Wu and Rehg presented CENTRIST, a visual descriptor for scene recognition, and demonstrated it on object and scene classification tasks [25]. Dixit et al. developed an adapted Gaussian mixture model to represent feature distributions and showed its utility for scene recognition in [6]. Pandey and Lazebnik presented an approach combining global image features and deformable parts-based models for scene recognition in [14].

Cakir et al. presented an approach to learn nearest neighbor based metric functions for indoor scenes by modifying NBNN to use codebook vectors instead of raw training feature vectors [4]. This speeds up computation for large image datasets. Tuytelaars presented a kernelized version of NBNN by constructing a feature vector from the image-to-class distances for a set of classes, and demonstrated state-of-the-art results for object and scene recognition [21]. McCann and Lowe presented an efficient extension of NBNN by simultaneously computing nearest neighbors of query features with all categories' features in one shot, and show that this provides significant speed up without affecting results [13].

The relative contributions of our study are: (1) experiments on both outdoor and indoor scene image and video datasets, (2) demonstration that using basic NBNN but combined with PCA on DSIFT gives state-of-the-art results, and (3) we present a novel sparsity constrained extension to NBNN. An interesting direction of research would be to study combination of approaches in [4], [21] and [13] with PCA and sparse-NBNN. For instance, NBNN-kernel can potentially be used to generate scene concept features to classify images and videos, and the approach in [13] can be used to speed-up the computation of sparse-NBNN. Shrivastava et al. present an approach to visual similarity that learns which parts of an image are informative for visual matching in a data-driven manner [17]. One possible future work is to use such approaches used to weight the contribution of the feature vectors in NBNN.

There has been extensive study of the task automatic video categorization, with a recent interest in unconstrained videos from the consumer domain, e.g. from websites such as the YouTube, e.g., [19, 16, 7, 20, 23, 26, 22]. There is a general consensus in the community that using multimodal features provides the best performance, e.g., [7, 20], etc. To highlight the utility of NBNN, we focus on a single visual feature, namely D-SIFT, that has been shown to be highly effective for both scene image and video categorization. A promising direction of future study is to build parallel NBNN models for multiple features including audio features such as MFCC, and then fuse their results.

2. Naive Bayes Nearest Neighbor

In this section, we give a brief review of Naive Bayes Nearest Neighbor (NBNN) approach proposed in [2], please refer to [2] for details. The overall idea in NBNN is to classify images based on an image-to-class distance rather than an image-to-image distance. The intuition is that comparing a query image with all the training images in a category allows us to compose new data by combining different parts of the training images. The experiments in [2] indicate that comparing raw feature vectors provides the best results.

Formally, suppose the problem is to classify a given image into one of N labels. Let T_i, $i = 1 \ldots N$, be the set of images for each label. During training, for each category i, all feature vectors for all images in that category are grouped into a set $C_i = \{\mathbf{x} | \mathbf{x} \in \mathcal{I} \text{ s.t. } \mathcal{I} \in T_i\}$. Given a query image, let $Q = \{\mathbf{d}_j\}$ be the set of its feature descriptors. NBNN computes the query's image-to-class distance for the i^{th} class as

$$h_i(Q) = \sum_{\mathbf{d} \in Q} \|\mathbf{d} - \text{NN}(\mathbf{d}, C_i)\|^2 \quad (1)$$

where $\text{NN}(\mathbf{d}, C)$ denotes the 1-nearest neighbor of $\mathbf{d}$ in the set C. The query image is assigned the label with the smallest distance, i.e. $\arg\min_i h_i(Q)$.

For the descriptors, the approach in [2] employed densely sampled SIFT descriptors [12] concatenated with spatial coordinates. More recent studies on scene recognition have also demonstrated the effectiveness of densely sampled SIFT descriptors, referred to as D-SIFT e.g., [3], etc. Therefore, for all experiments, we employ D-SIFT features concatenated with the spatial coordinates for the descriptors. For all results described in this paper, the spatial coordinates are normalized to between $[0, 1]$, and are assigned a weight of 0.5 relative to the D-SIFT features. The nearest neighbor search is performed using approximate nearest neighbor algorithm [1].

2.1. PCA for NBNN

A simple but important observation from our experiments is that applying principal components analysis (PCA) on the D-SIFT features significantly improves performance of NBNN in scene recognition. PCA-SIFT has previously been employed for image correspondence and object recognition, e.g., [9]. In order to explore the effects of dimensionality reduction, we studied two other variations of NBNN for scene categorization:

No PCA: Perform NBNN directly on the descriptors consisting of D-SIFT along with spatial-coordinates. Previous studies on approximate nearest neighbors have observed a worsening in performance of the kd-tree algorithm for high dimensions. To avoid confounding the fidelity of kd-tree's ANN computation, we employed a brute-force search for the no-PCA case.

Local PCA: It is generally accepted in the community that a single PCA on the entire feature space ignores the rich structure of the feature distributions. Kambhatla and Leen proposed a local version of PCA, in which PCA is performed separately in different regions of the feature space [8]. In our implementation, we compute local PCA on the fly for each feature vector from the query image. For the local PCA computation, we use all D-SIFT vectors from all images outside of the test set irrespective of their labels. In other words, we employ unlabeled data to improve the classification accuracy. The outline of the computation for each query feature is:

1. collect 100 nearest neighbors to the query feature vector among unlabeled data using ANN search.
2. compute local PCA on these neighbors.
3. for each category, collect 10 nearest neighbors to the query feature vector, project the neighbors and query feature using the local-PCA, and select the 1-NN to the query feature vector from among the local-PCA projected neighbors.
4. the distance to this 1-NN is used for the local PCA version of NBNN.

Approach	Classification Accuracy
D-SIFT+NBNN (no PCA)	81.48% ± 1.50%
D-SIFT+PCA+NBNN	84.67% ± 1.35%
D-SIFT+local-PCA+NBNN	85.42% ± 1.51%

Table 1. Classification results on the UIUC 8 Sports Scene Image dataset indicating the utility of PCA for NBNN based classification. For reference, the state-of-the-art is 84.4% using PH-SVM-SP [6]

Table 1 shows results obtained for D-SIFT+NBNN with global PCA, no PCA and local PCA on the UIUC 8 Sports Event image dataset (UIUC-8); for reference, the state-of-the-art is 84.4% using PH-SVM-SP [6]. We observe that applying PCA on the D-SIFT descriptors provides significant gains in NBNN's performace (+3%) compared to no PCA. Using local PCA within NBNN further improves the results.

The local PCA approach is of interest in applications having a small number of feature vectors in each category. In such scenarios, the fidelity of the nearest neighbor search is adversely affected by the sparse distribution of features in high-dimensional feature space (e.g. 128D for SIFT). Classification results can be improved by exploiting unlabeled data using techniques such as local PCA. For instance, when using only 10 training images in the UIUC-8 dataset, global-PCA+NBNN gives 67.98%±1.96%, where as local-PCA+NBNN gives 69.96% ± 2.33%.

The performance improvement with PCA is truly fortuitous because using PCA significantly reduces the space and time complexity. In particular, a study by Silpa-Anan and Hartley [18] shows that applying PCA on the feature vectors improves the efficiency of the kd-trees algorithm, which is used for the approximate nearest neighbor search.

3. Sparse Naive Bayes Nearest Neighbor

In this section we present a sparsity constrained version of NBNN. The basic version of NBNN computes an image-to-class distance, thus all training images within a category can contribute to matching a query image's features. In certain scenarios, it would be advantageous to limit the number of training images that can be involved for a given query image. Notice that we allow for a different subset of training images to match with each query image, restricting the cardinality of the subset to be at most k. This is akin to a sparsity constraint on matching, and we refer to it as sparse-NBNN. Motivating scenarios include:

- Handle imbalanced training data. Low-level features such as D-SIFT have limited specificity. As the number of training images in a category grows, they would be able explain almost any image, not just from the category in question. Intuitively, one would think that if the query image is from a category, then a limited number of training images from that category should be able to explain the query image's features.

- Transfer tags and annotations from the training images to the query image. Quality of the tagging will be improved by limiting the number of training images that are allowed to match with the query image. Such tagging will also be able exploit the fact that NBNN inherently provides a fine grained correspondence between the training images' features and the query image's features through the nearest neighbor matching.

The objective of sparse-NBNN is to choose the best subset of k training images from a category to give the lowest image-to-class distance. We pose this as a weighted Set Cover problem. The query image's features correspond to the set that must be covered. The training images constitute the subsets, and we must choose at most k of them. When a training image, $\mathcal{I}$, is selected, we may get a benefit for each of the query feature, $\mathbf{d}$, based on the NN distance $\|\mathbf{d} - \mathrm{NN}(\mathbf{d}, \mathcal{I})\|$. The constraint is that each query feature may be covered at most once.

Suppose we want to compute sparse-NBNN based distance between a query image and a category. Let the query image have features $Q = \{d_i\}_{i=1}^m$. Let the category consist of images $T = \{\mathcal{I}_j\}_{j=1}^n$. The nearest neighbor of a query feature within each training image is denoted by $\mathrm{NN}(\mathbf{d}_i, \mathcal{I}_j)$. We define the benefit of covering $\mathbf{d}_i$ with image $\mathcal{I}_j$ as $\delta - \|\mathbf{d}_i - \mathrm{NN}(\mathbf{d}_i, \mathcal{I})\|^2$, where δ is the maximum possible squared distance in the feature space. The goal is to compute a subset S of the training images set T to optimize

$$\max_{S \subseteq T} \sum_{\mathbf{d} \in Q} \max_{\mathcal{I} \in S} (\delta - \|\mathbf{d} - \mathrm{NN}(\mathbf{d}, \mathcal{I})\|^2) \quad \text{such that } |S| \leq k \tag{2}$$

Approach	Classification Accuracy
D-SIFT+PCA+NBNN	84.67% ± 1.35%
D-SIFT+PCA+sparse-NBNN $k = 20$	86.04% ± 1.14%

Table 2. Classification results on the UIUC 8 Sports Scene Image dataset indicating the utility of sparse-NBNN relative to basic NBNN.

We formulate an integer linear program (ILP) for sparse-NBNN. One ILP is constructed for each category and query image pair, and optimized. For categorization, the label is assigned based on the ILP resulting in maximal objective function value. Let $b_{ij} = \delta - \|\mathbf{d}_i - \mathrm{NN}(\mathbf{d}_i, \mathcal{I}_j)\|^2$; by definition we have $b_{ij} \geq 0$. The ILP for sparse-NBNN is as follows:

$$\begin{aligned} \max \quad &: \quad \sum_{i,j} b_{ij} z_{ij} \\ \text{subj. to} \quad &: \quad \sum_{j} z_{ij} \leq 1 \\ & \quad\; z_{ij} \leq y_j \\ & \quad \sum_{j} y_j \leq k \\ & \quad\; y_j \in \{0, 1\} \;\; \text{(binary)} \end{aligned} \tag{3}$$

Here, y_j's are binary variables indicating the training images selected by sparse-NBNN to match with a given query image. z_{ij}'s indicate if the i^{th} query feature is covered by the j^{th} training image. Due to the inequality constraints and the fact that $b_{ij} \geq 0$, exactly one of z_{ij}'s for i^{th} query feature will be set to 1.

Table 2 shows results of using sparse-NBNN for scene categorization on the UIUC-8 dataset. Using sparse-NBNN improves classification accuracy over basic NBNN.

Figure 1 shows a visualization of the constraints imposed by sparse-NBNN on the matching between features from the test image and from training images of a category. The top panel shows a correctly classified image from the UIUC-8 dataset's Rowing class, and a color map rendering of coverage of this test image's features by training images selected by sparse-NBNN, for different values of k. We can see that the mapping from training images to the test image is smooth for small values of k. Compare this with the "noisy" coverage map obtained by basic NBNN using all 70 training images. We also show the two training images from Rowing class that were selected for $k = 2$. The left (blue) image contributes water-features, and the right (green) training image predominantly contributes to the boat.

The bottom panel of Figure 1 shows an image from Rowing class that was incorrectly classified as Croquet, along with the coverage map rendering for $k = 2$ sparse-NBNN, and the two selected Croquet training images. The extreme

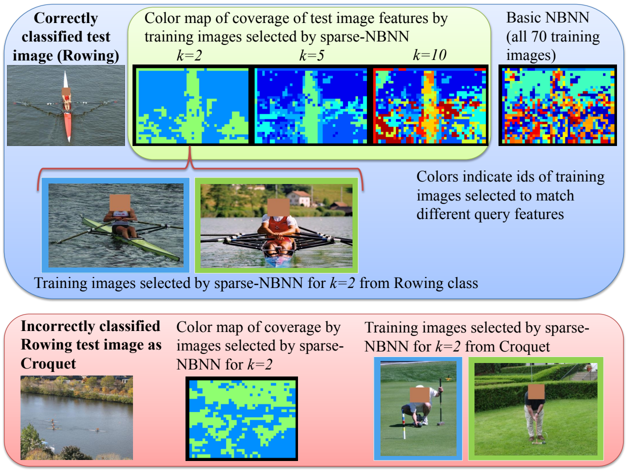
Correctly classified test image (Rowing)
Color map of coverage of test image features by training images selected by sparse-NBNN
k=2
k=5
k=10
Basic NBNN (all 70 training images)
Colors indicate ids of training images selected to match different query features
Training images selected by sparse-NBNN for k=2 from Rowing class
Incorrectly classified Rowing test image as Croquet
Color map of coverage by images selected by sparse-NBNN for k=2
Training images selected by sparse-NBNN for k=2 from Croquet

Figure 1. Visualization of constraints imposed by sparse-NBNN on feature matching between test and training images taken from UIUC 8 Sports Event image dataset. Colors indicate ids of training images selected to match with different query features. Top panel: A correctly classified image. For sparse-NBNN, the coverage of test image features by training images is spatially smooth for small values of k. The colors indicate indices of selected training images matching different query features. The coverage map for basic NBNN is highly noisy as a large number of training images contribute small pieces to the test image. Of the two training images selected by sparse-NBNN for $k = 2$, one contributed features to water areas, and the other to the boat. Bottom panel: An incorrectly classified image along with 2 training images selected by sparse-NBNN for $k = 2$. The extreme mismatch in semantic content indicates the limitations of low-level features. *Best viewed in color.*

mismatch in the semantic content of the test image and the selected training images indicates the limitations of low-level texture-like features such as D-SIFT. This points to an potential direction of research to combine high-level reasoning with NBNN. Essentially, sparse-NBNN can be used to “force” a test image to select a small set of training images to explain its hypothesized label. We can then apply high-level reasoning such as object concepts [11], etc. to verify the hypothesis. Use of sparse-NBNN would focus the application of high-level reasoning to a small set of “evidence” images.

Imbalanced data: To observe the relative utility of sparse NBNN to handle imbalanced data, we conduct an experiment on the UIUC 8 Sports event dataset by increasing the training data to include all images other than the test images. This resulted in imbalance in the number of training images in the 8 classes, ranging from 77 to 190 images; the number of testing images was kept at 60. There are two counteracting forces here: increasing training data typically improves classification accuracy; however, imbalance in training data typically results in decrease in performance. Table 3 shows the results. We observe that basic NBNN shows a *loss* in performance with increase in training data, likely due to data imbalance. In contrast, Sparse-NBNN is able to exploit the additional training data despite the imbalance, showing gains of $\approx 3\%$. Moreover, the performance improves upon making sparisity more stringent, from $k = 20$ to 10. This makes intuitive sense, with more training data, assuming diversity one would expect a smaller number of training images to be adequate to explain a test image. Please note that although better than state-of-the-art, these results use larger amounts of training data and hence cannot be compared with previous studies that use only 70 training images. To our knowledge, ours is the study to address imbalanced data issues for NBNN.

Approach	Classification Accuracy
D-SIFT+PCA+NBNN (70 training images)	$84.67\% \pm 1.35\%$
D-SIFT+PCA+NBNN $(77-190$ training images)	$79.17\% \pm 1.89\%$
D-SIFT+PCA+sparse-NBNN $k = 20$ $(77-190$ training images)	$86.85\% \pm 1.2\%$
D-SIFT+PCA+sparse-NBNN $k = 10$ $(77-190$ training images)	$87.71\% \pm 1.49\%$

Table 3. Classification results on the UIUC 8 Sports Scene Image dataset with larger and imbalanced training data indicating utility of sparse-NBNN vs. basic NBNN.

4. Scene Image Categorization Experiments

In this section we present experimental results on indoor and outdoor scene categorization. We use the same PCA projection matrix for all scene and video datasets. For the scene recognition experiments, we use 30 principal components. Thus, together with the two spatial coordinates, we obtain a set of 32-dimensional feature vectors for the images. All images were scaled so that the smallest dimension was at most W pixels, with 4 scales used by setting $W = \{150, 200, 256, 300\}$. The feature vectors from all scaled versions of an image were put in one “bag” to enable combinations of local features from different scales during matching.

4.1. UIUC 8 Sports Events Image Dataset

The UIUC 8 Sports Events Image dataset [10] consists of 8 sports events: rowing, badminton, polo, boccie, snowboarding, croquet, sailing and rock climbing. Following [10], we use 70 randomly selected images from each class for training and 60 images for testing, with 10 random iterations.

Table 4 shows the results of our approach and state-of-the-art in literature. The results indicate that D-SIFT+PCA+NBNN achieves state-of-the-art results despite

Approach	Classification Accuracy
D-SIFT+PCA+NBNN	$84.67\% \pm 1.35\%$
PH-SVM-SP [6]	84.4%
AGMM-SP [6]	82.9%
HIK-CBK [24]	81.17%
CENTRIST [25]	$78.25\% \pm 1.27\%$
Object-Bank [11]	76.3%
Li & Fei-Fei [10]	73.4%

Table 4. Classification results on the UIUC 8 Sports Scene Image dataset

Approach	Classification Accuracy
D-SIFT+PCA+NBNN	$48.84\% \pm 1.08\%$
NNbMF [4] (Involves manual tuning)	47.01%
NNbMF [4] (Automatic parameter tuning)	45.22%
DPM + GIST-color + SP [14]	43.1
CENTRIST [25]	$36.88\% \pm 1.10\%$
Object-Bank [11]	37.6%
Quattoni & Torralba [15]	≈ 28

Table 5. Classification results on the MIT 67 Indoor Scene Image dataset

its simplicity.

4.2. MIT 67 Indoor Scene Image Dataset

The MIT 67 Indoor Scene Image dataset [15] consists of 15620 images from 67 indoor scene categories. Following [15], we use 80 images from each category for training and 20 images for testing, with 10 randomized sampling iterations. Table 5 shows the results of our approach and state-of-the-art in literature.

The results indicate that our approach achieves state-of-the-art results in both outdoor and indoor scene recognition relative to more complex algorithms. NNbMF [4] obtains 40.75% accuracy on MIT-67 dataset using basic NBNN, whereas we obtain 48.84% by include PCA. We observe that sparse-NBNN did not produce improvement over basic NBNN for the MIT-67 dataset. One potential reason is that given the large number of target categories in MIT-67, all 80 training images in each category are needed for modeling a given test image.

The study in [21] shows results on the UIUC 15 Scene dataset (UIUC-15), obtaining 75% for NBNN and 79% for NBNN-kernel. We obtain 79% accuracy on the UIUC-15 dataset using PCA on D-SIFT with basic NBNN.

5. Video Event Detection Experiments

In this section, we describe application of D-SIFT+NBNN for video event detection. We focus on the TRECVID Multimedia Event Detection task, which aims at detecting videos that depict certain events of interest from a large set of query videos, many of which are irrelevant to the desired events. The MED 2011 evaluation consisted of 10 events, e.g. "Making a sandwich," "Grooming an animal," etc. The problem is complex because of high intra-class variance in terms of scenes, actors, camera view, appearance, etc.

We test our approach for the challenging Ad Hoc event scenario in which training set consists of only 15 positive example videos per event, along with several training videos from the background. In our experiments, the training set consisted of 15 videos for each of the 10 events of interest, and 4900 videos from the background. The testing set consisted of between 23 to 44 videos from each event of interest, and 2450 videos from the background.

Under the MED task, each video is scored for each event independently, i.e. it is a detection task and there is no comparison across the event models. In addition to collecting D-SIFT features for the training videos of each event, we construct an NBNN model for the background videos by collecting D-SIFT vectors from a randomly sampled subset of videos from the training background set. Let us denote the NBNN distance of query V to this set by $h_B(V)$ (see eq.(1)). The score for a video having a set of features V to belong to an event i is defined as

$$s(V,i) = -h_i(V) + \lambda h_B(V) \tag{4}$$

where λ is a relative weight parameter.

We compare D-SIFT+NBNN with two state-of-the-art coding and pooling techniques applied on D-SIFT features for video categorization:

- D-SIFT+SQ+SP+SVM: project D-SIFT on a 4000 element codebook using soft quantization followed by spatial pooling and using SVMs for classification [7].
- D-SIFT+sparsity+SVM: project D-SIFT on a 2048 element codebook using sparse constraints, followed by α-histogram pooling and using SVMs for classification [22].

The approaches are evaluated based on their probability of missed detections (pmd) and probability of false acceptance (pfa). A criterion defined by NIST, referred to as NDC is used to measure the performance. By definition, a lower NDC score indicates better performance. Therefore, the goal is to minimize

$$\min_t \left[\mathrm{pmd}(t) + 12.49 \times \mathrm{pfa}(t)\right]$$

while achieving at least 25% detection rate if possible.

The λ weighting parameter are selected to minimize NDC, with discrete choice of $\lambda \in \{1,2\}$ for simplicity. Table 6 shows the results of NBNN and the SVM-based approaches. As can be observed, the result using NBNN is within standard deviation of the average NDC measure for the two state-of-the-art algorithms using D-SIFT for video categorization. This is an important result given the complexity of the event detection task, and the simplicity of the D-SIFT+PCA+NBNN approach. Despite its simplicity, our approach matches the results of two state-of-the-art coding and pooling techniques.

Importance of modeling the background class: NBNN is a generative model. Therefore, it is possible to compute score for a video belonging to an event based solely on that event's NBNN model, ignoring the background training video data. This is equivalent to setting $\lambda = 0$ in the computation of $s(V, i)$. We observed that ignoring the background class when using NBNN led to deep loss in performance; the average NDC score rose to 1.702 ± 0.048. Compare this with the average NDC score of 1.196 ± 0.346 obtained by NBNN when background video data is modeled, i.e. $\lambda \neq 0$. In light of this result, it is important to train background models for NBNN even for detections tasks in which each category is detected independently.

6. Conclusion

We presented a thorough study using NBNN for scene recognition and video event detection achieving results comparable to state-of-the-art techniques using more complex algorithms. An important observation is to apply PCA on D-SIFT vectors to significantly improve scene recognition accuracy. The D-SIFT+PCA+NBNN approach forms a strong and simple straw man that can be quickly applied by researchers to new image categorization problems to create baselines.

We described an approach to use local PCA to improve NBNN's performance, and demonstrated its significance for limited training data.

We presented a novel sparsity constrained version of NBNN. Experiments indicate that sparse-NBNN improves scene recognition accuracy, and is better able to handle imbalance in data compared to basic NBNN. Promising directions for future work include using sparse-NBNN for annotation/tag transfer from training to test images, and combining PCA and sparse-NBNN with [21] and [13] to efficiently compute object and scene concepts for images and videos.

Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC or the U.S. Government.

References

[1] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45:891–923, November 1998. 3

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, june 2008. 1, 2, 3

[3] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Proc CVPR*, 2010. 2, 3

[4] F. Cakir, U. Gdkbay, and zgr Ulusoy. Nearest-neighbor based metric functions for indoor scene recognition. *Computer Vision and Image Understanding*, 115(11):1483 – 1492, 2011. 1, 2, 6

[5] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011. 1

[6] M. Dixit, N. Rasiwasia, and N. Vasconcelos. Adapted gaussian models for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 937–943, june 2011. 1, 2, 3, 6

[7] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010. 2, 6

[8] N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural Comput.*, 9:1493–1516, October 1997. 3

[9] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia*, 2004. 3

[10] L.-J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *IEEE Intern. Conf. in Computer Vision (ICCV)*, 2007. 1, 2, 5, 6

[11] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Proceedings of the Neural Information Processing Systems (NIPS), 2010*, 2010. 2, 5, 6

Event	D-SIFT+PCA+ NBNN	D-SIFT+SQ+SP+ SVM	D-SIFT+sparsity+ SVM
Birthday party	1.592	1.459	1.772
Changing a vehicle tire	1.332	0.850	0.932
Flash mob gathering	0.743	0.499	0.543
Getting a vehicle unstuck	0.874	1.107	1.435
Grooming an animal	1.347	1.573	1.573
Making a sandwich	1.615	1.569	1.617
Parade	1.146	1.264	1.137
Parkour	1.594	0.960	0.947
Repairing an appliance	0.796	1.014	1.063
Working on a sewing project	0.922	0.924	1.072
Average across events	1.196 ± 0.346	1.122 ± 0.328	1.209 ± 0.361

Table 6. NDC results on the Ad Hoc Multimedia Event Detection Task. Lower NDC score indicates better performance.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004. 3

[13] S. McCann and D. G. Lowe. Local naive bayes nearest neighbor for image classification. Technical Report TR-2011-11, Department of Computer Science, University of British Columbia, 2011. 2, 7

[14] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *Proc ICCV*, 2011. 2, 6

[15] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420, june 2009. 1, 2, 6

[16] G. Schindler, L. Zitnick, and M. Brown. Internet video category recognition. In *Proc. CVPRW '08*, 2008. 2

[17] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph.*, 30(6), 2011. 2

[18] C. Silpa-Anan and R. Hartley. Optimised kd-trees for fast image descriptor matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, june 2008. 3

[19] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc ICCV*, 2003. 2

[20] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In *Proc CVPR*, 2010. 2

[21] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The nbnn kernel. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1824–1831, Nov. 2011. 2, 6, 7

[22] S. Vitaladevuni, P. Natarajan, R. Prasad, and P. Natarajan. Efficient orthogonal matching pursuit using sparse random projections for scene and video classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2312–2319, nov. 2011. 2, 6

[23] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtubecat: Learning to categorize wild web videos. In *Proc. CVPR*, 2010. 2

[24] J. Wu and J. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 630–637, 29 2009-oct. 2 2009. 2, 6

[25] J. Wu and J. Rehg. Centrist: A visual descriptor for scene categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1489–1501, aug. 2011. 2, 6

[26] H. Zhou, T. Hermans, A. Karandikar, and J. M. Rehg. Movie genre classification via scene categorization. In *Proc Multimedia*, 2010. 2