A Novel Locally Linear KNN Model for Visual Recognition

Qingfeng Liu, Chengjun Liu
New Jersey Institute of Technology
Newark, NJ, USA
ql69, cliu@njit.edu

Abstract

This paper presents a novel locally linear KNN model with the goal of not only developing efficient representation and classification methods, but also establishing a relation between them so as to approximate some classification rules, e.g. the Bayes decision rule. Towards that end, first, the proposed model represents the test sample as a linear combination of all the training samples and derives a new representation by learning the coefficients considering the reconstruction, locality and sparsity constraints. The theoretical analysis shows that the new representation has the grouping effect of the nearest neighbors, which is able to approximate the "ideal representation". And then the locally linear KNN model based classifier (LLKNNC), which shows its connection to the Bayes decision rule for minimum error in the view of kernel density estimation, is proposed for classification. Besides, the locally linear nearest mean classifier (LLNMC), whose relation to the LLKNNC is just like the nearest mean classifier to the KNN classifier, is also derived. Furthermore, to provide reliable kernel density estimation, the shifted power transformation and the coefficients cut-off method are applied to improve the performance of the proposed method. The effectiveness of the proposed model is evaluated on several visual recognition tasks such as face recognition, scene recognition, object recognition and action recognition. The experimental results show that the proposed model is effective and outperforms some other representative popular methods.

1. Introduction

Visual recognition such as face recognition, object recognition, scene recognition and action recognition has received much attention over the past few decades. Many methods are developed and one of the most successful and well-studied method is the subspace method, whether the linear subspace [2], [27] or the non-linear ones [16], [35]. Recently, the sparse representation based method [31] is proposed to address the problem of robust representation and classification. Many variants [11], [34], [37] are further proposed to incorporate discriminative information for learning the discriminative dictionary and the sparse representation. For most methods, representation and classification are developed independently, which violates the need that the representation methods should serve and facilitate the subsequent classification methods for visual recognition. In addition, these methods bring other issues such as classifier restriction, computational complexity etc..

In order to address these issues, we propose a novel locally linear k nearest neighbors (LLKNN) model for robust visual recognition. The proposed method first learns a new representation for every test sample as a linear combination of all the training samples based on the criteria of reconstruction, locality, and sparsity. The new representation vector, which possesses the grouping effect of the nearest neighbors, is then provided as the input of a locally linear KNN model based classifier (LLKNNC) and a locally linear nearest mean classifier (LLNMC), respectively. The power of the proposed LLKNNC is guaranteed by establishing its connection to the Bayes decision rule for minimum error in the view of kernel density estimation. The shifted power transformation and a coefficients cut-off method are further applied for robustness and reliability.

The effectiveness of the proposed method is assessed on five representative data sets. In particular, for face recognition, the AR face database [17] is used; for scene recognition, the 15 scenes dataset [13] and the MIT-67 indoor scenes dataset [21] are applied; for object recognition, the Caltech 256 dataset [10] is utilized; and for action recognition, the UCF50 dataset [22] is used. The experimental results show the feasibility of the proposed method.

The system architecture is illustrated in figure 1. The pattern vector is first preprocessed by the the shifted power transformation. Then the dimension reduction method is applied. The proposed LLKNN model further derives a new representation vector v. Finally, the LLKNNC and the LLNMC are applied for classification.

The main contributions of the paper are as follows:

- First, we propose a novel locally linear KNN

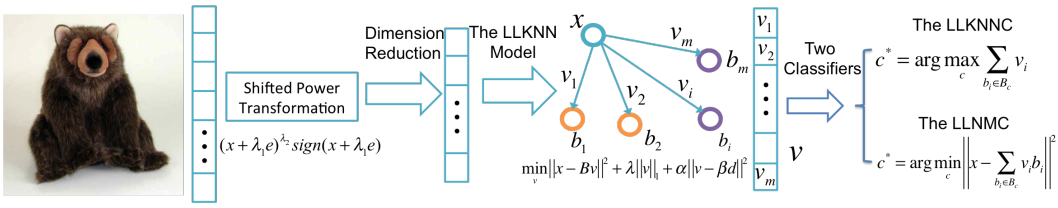

Figure 1. The system architecture of the proposed Locally Linear KNN model

(LLKNN) model, which derives representation that can approximate the "ideal representation" by exploiting the grouping effect of the nearest neighbors property of the proposed model.

- Second, we propose the locally linear KNN model based classifier (LLKNNC), which reveals its connection to the Bayes decision rule for minimum error in the context of kernel density estimation.

- Third, we propose the locally linear nearest mean classifier (LLNMC), whose relation to the LLKNNC is just like the relation between the nearest mean classifier and the KNN classifier.

- Fourth, we address the issues of reliable density estimation by applying the shifted power transformation and the coefficients cut-off method from two aspects: the sensitiveness to the global bandwidth and the adverse impact of the distant neighbors.

2. Related Work

Sparse representation methods are broadly applied for visual recognition. Some methods [5], [6] seek to model the intra-class variations within the dictionary to improve the performance for face recognition. Recently, some discriminative dictionary learning methods are proposed for sparse representation. Zhang et al. [37] proposed an objective function and applied a discriminative singular value decomposition (D-KSVD) method to learn the discriminative dictionary and the classifier simultaneously. Jiang et al. [11] improved upon [37] by adding a label consistent regularization term. Zhou et al. [38] presented an Joint Dictionary Learning (JDL) method that jointly learns both a commonly shared dictionary and the class-specific sub-dictionaries to enhance the discrimination of the dictionaries. Yang et al. [33], [34] proposed the Fisher Discrimination Dictionary Learning (FDDL) method, which learns a structured dictionary that consists of a set of class-specific sub-dictionaries.

In comparison, our method improves upon these methods in the following ways. (i) The derivation of the dictionary of other methods is very time-consuming because it needs to iteratively update the dictionary and the sparse representation alternatively. (ii) Some methods are restricted to linear classifiers, which exclude nonlinear classifiers that may achieve better performance. (iii) The sub-dictionary based methods may lead to deteriorated performance when the number of the training samples for each class is small because the sub-dictionaries either are trained separately for each class or depend on the corresponding class too much. (iv) Our proposed method is capable of establishing the relation between a representation method and its classifiers in the sense of approximating the Bayes decision rule for minimum error.

3. The Locally Linear KNN Model

One common assumption in the literature is that a test sample is a linear combination of all the training samples [31], which captures the variation of the real datasets.

The ideal case is that only the coefficients of the training samples with the same class label as the test sample are non-zero, and otherwise zero. Mathematically, the idea representation is defined as follows.

Definition 3.1 ***The ideal representation***. *Given the test sample $\boldsymbol{x} \in \mathbb{R}^n$ from the class c, and all the training samples $\boldsymbol{b}_i \in \mathbb{R}^n (i = 1, 2, ..., m)$, the ideal representation of $\boldsymbol{x}$ is the coefficient vector $\boldsymbol{v} = [v_1, v_2, ..., v_m]^t \in \mathbb{R}^m$ so that*

$$\boldsymbol{x} = \sum_{i=1}^{m} v_i \boldsymbol{b}_i \tag{1}$$

where v_i is non-zero if $\boldsymbol{b}_i$ belongs to the c-th class and otherwise 0.

As a result, the ideal representation is highly sparse, which induces the development of the sparse representation based methods [31].

However, the following two issues inherent of such sparse representation based methods are still waiting for a solution. First, the sparsity constraint alone cannot guarantee that the expected coefficients are non-zero to approximate the ideal representation. Specifically, the training samples that are in the same class as the test sample are often highly correlated while the sparsity constraint often tends to select one of them with non-zero coefficient [39]. Second,

the sparse representation based classifier [31] is not directly related to the optimal classification rules, such as the Bayes decision rule for minimum error [7].

To address these two issues, we propose a new locally linear KNN (LLKNN) model based on the observation that the k nearest training sample neighbors of the test sample are highly probable to share the same class label with the test sample if they are robustly selected. Note that our method applies the L_1 norm for robustness and utilizes the nearest training sample neighbors of the test sample for locality property. As a result, our method reveals the grouping effect of the nearest neighbors (GENN), and consequently, the ideal representation is more likely to be realized by our proposed method.

Mathematically, the novel locally linear KNN (LLKNN) model is defined as follows:

$$\min_{\mathbf{v}} ||\mathbf{x} - \mathbf{B}\mathbf{v}||^2 + \lambda||\mathbf{v}||_1 + \alpha||\mathbf{v} - \beta\mathbf{d}||^2 \qquad (2)$$

where $\mathbf{x} \in \mathbb{R}^n$ is the test sample, $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_m] \in \mathbb{R}^{n\times m}$ is the training sample matrix and coefficients vector $\mathbf{v} \in \mathbb{R}^m$ is the derived representation. $||\cdot||$ is the L_2 norm and $||\cdot||_1$ is the L_1 norm. The vector $\mathbf{d} = [d_1, d_2, ..., d_m]^t \in \mathbb{R}^m$, and $d_i = exp\{-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{b}_i||^2\}$. The parameter σ is used for adjusting the decay speed. We can observe that d_i is larger if the training sample $\mathbf{b}_i$ is closer to the test sample $\mathbf{x}$.

The first term maintains the reconstruction ability, the second term is the sparsity constraint that keeps the robustness property, and the third term represents the locality property so that the closer the training sample is to the test sample, the larger its coefficient will be. The model parameters: λ, α and β contribute to balancing each term. The proposed model thus emphasizes the nearest training sample neighbors of the test sample and assigns them large coefficients. Section 3.3 further shows that only the k largest coefficients of the nearest training neighbors in the same class as the test sample are necessary for achieving good performance.

The rationale of the proposed method is twofold.

- First, the k nearest training sample neighbors are more likely to be in the same class as the test sample if they are robustly selected. Besides, as shown in Section 3.1, the proposed model has the **grouping effect of the nearest neighbors (GENN)** — the training samples that are highly correlated and close enough to the test sample tend to have similar and large coefficients. As a result, the proposed model tends to derive the ideal representation.

- Second, the LLKNN model based classifier (LLKNNC) approximates the Bayes classifier in the sense of kernel density estimation in Section 3.2 based on the representation derived by the proposed model.

3.1. Representation

The new representation is derived by optimizing the criterion in equation 2. The FISTA (Fast Iterative Shrinkage Thresholding Algorithm) algorithm [1] is applied to solve it. The equation 2 can be decomposed into $f(\mathbf{v}) + g(\mathbf{v})$, where $f(\mathbf{v}) = ||\mathbf{x}-\mathbf{B}\mathbf{v}||^2+\alpha||\mathbf{v}-\beta\mathbf{d}||^2$ and $g(\mathbf{v}) = \lambda||\mathbf{v}||_1$. To ensure convergence, the maximal step size for the FISTA algorithm is selected as $\frac{1}{L}$, where $L = 2\lambda_{max}(\mathbf{B}^t\mathbf{B} + \alpha\mathbf{I})$, which means twice of the largest eigenvalue of the matrix $\mathbf{B}^t\mathbf{B} + \alpha\mathbf{I}$.

By optimizing the criterion in equation 2, the new representation possesses the grouping effect of the nearest neighbors (GENN) as shown below.

Theorem 3.1 *Given a L_2 normalized test sample $\boldsymbol{x}$ ($||\boldsymbol{x}||^2 = 1$), the L_2 normalized training sample matrix $\boldsymbol{B}$ ($||\boldsymbol{b}_i||^2 = 1, i = 1, 2, ..., m$) and the vector $\boldsymbol{d} = [d_1, d_2, ..., d_m]^t$, let $\boldsymbol{v}^* = [v_1^*, v_2^*, ..., v_m^*]^t$ be the solution to the LLKNN model defined in equation 2. Define the sample correlation ρ of two training samples $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$ as $\rho = \boldsymbol{b}_i^t\boldsymbol{b}_j$ and the difference between the coefficients v_i^* and v_j^* $(i, j = 1, 2, ..., m)$ as*

$$M(i, j) = |v_i^* - v_j^*| \qquad (3)$$

Then, if the signs of v_i^* and v_j^* are the same, we have

$$M(i, j) \leq \frac{C}{\alpha}\sqrt{2(1 - \rho)} + \beta|d_i - d_j| \qquad (4)$$

where $C = \sqrt{(1 + \alpha\beta^2||\boldsymbol{d}||^2)}$, which is a constant.

The equation 4 is called the grouping effect of the nearest neighbors (GENN), which means, if the training samples are highly correlated ($\rho \approx 1$) and close enough to the test sample ($d_i \approx d_j$ and d_i, d_j are large), then the coefficients of the training samples are similar ($v_i^* \approx v_j^*$). The experimental analysis in Section 5.5, implies the tightness of the bound in the equation 4 that the GENN property amplifies the coefficients of the training samples in the same class as the test sample while suppresses others. This is the case consistent with the ideal representation.

3.2. Classification

After we derive the representation $\mathbf{v}$ by the LLKNN model for the test sample $\mathbf{x}$ with the given training sample matrix $\mathbf{B}$, we first propose the LLKNN model based classifier (LLKNNC) to classify the test sample $\mathbf{x}$. The classification rule is defined as follows:

$$c^* = \arg\max_c \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i \qquad (5)$$

where $c = 1, 2, ..., w$ is the class label, $\mathbf{B}_c$ is the set of training samples in the c-th class. The LLKNNC thus assigns the test sample to the class c^* by a major soft vote of all the training samples in each class, which means that the test sample is classified to the class that owns the largest sum of the coefficients.

The effectiveness of the LLKNNC is stated in theorem 3.2. With some reasonable approximations, the classification rule defined in equation 5 approximates the Bayes decision rule for minimum error [7].

Theorem 3.2 *Given the test sample* $\boldsymbol{x}$*, the corresponding representation* $\boldsymbol{v}$*, the two transformations* $v_i = \frac{v_i - v_{min}}{v_{max} - v_{min}}$ *and* $v_i = \frac{v_i}{\sum_{i=1}^{m} v_i}$ *are applied first, where* v_{min} *and* v_{max} *is the minimal and maximal value among all the elements of the vector* $\boldsymbol{v}$*.*

Then, if the prior distribution $p(c)$ *is equal for all the classes, the Bayes decision rule is approximated by the proposed LLKNN model based classifier in the sense of kernel density estimation.*

$$\begin{aligned} c^* &= \arg\max_c \sum_{\boldsymbol{b}_i \in \boldsymbol{B}_c} v_i \\ &\approx \arg\max_c \sum_{\boldsymbol{b}_i \in \boldsymbol{B}_c} \beta d_i + const \\ &\propto \arg\max_c p(c|\boldsymbol{x}) \end{aligned} \tag{6}$$

Note that the transformations make $\sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i$ fall into $[0, 1]$ in order to establish a relation to the posterior probability. It is easy to see that the transformations do not affect the result of the LLKNNC.

Another classification method, the locally linear nearest mean classifier (LLNMC), is proposed as well. The relation between LLNMC and LLKNNC is similar to that between the nearest mean classifier and the KNN classifier.

We first define the "mean" of the c-th class as follows:

$$\mathbf{m}_c = \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i \mathbf{b}_i \tag{7}$$

The LLNMC is then defined as follows

$$\begin{aligned} c^* &= \arg\min_c ||\mathbf{x} - \mathbf{m}_c||_2^2 \\ &= \arg\min_c ||\mathbf{x} - \sum_{\mathbf{b}_i \in \mathbf{B}_c} v_i \mathbf{b}_i||_2^2 \end{aligned} \tag{8}$$

Note that the popular minimal residual classifier [31] is a special case of our proposed LLNMC when the locality information is discarded. The difference between our proposed LLNMC and the minimal residual classifier is that the LLNMC uses the new derived representation instead of the sparse representation used by the minimal residual classifier.

3.3. Reliable Kernel Density Estimation

Theorem 3.2 tells us that the power of LLKNNC comes from the kernel density estimation. Then there are two issues of the kernel density estimation that should be resolved to improve the reliability, namely the sensitiveness to the global window width denoted as the value of σ and the adverse impact of distant neighbors.

The first issue is that the kernel density estimation often suffers the global window width when the underlying density requires different amounts of smoothing at different locations, which means the value of σ should be different for different locations of features. As demonstrated in [28], kernel density estimation works well for densities that are not far from Gaussian in shape because of the uniformly used global window width. The shifted power transformation is able to transform data to a near Gaussian shape so that the new data can be well estimated. Therefore, we apply the following shifted power transformation to the pattern vector before applying the LLKNN model.

$$T(\mathbf{x}) = |\mathbf{x} + \lambda_1 \mathbf{e}|^{\lambda_2} sign(\mathbf{x} + \lambda_1 \mathbf{e}) \tag{9}$$

where $sign(\mathbf{x})$ denotes the sign vector of each element of the vector $\mathbf{x}$ with the value 0, 1 and -1, $0 < \lambda_1, \lambda_2 \leq 1$ and $e = [1, 1, ..., 1]^t$. Please note that all the vector operations are element-wise. Moreover, in the new transformed space, to further alleviate the sensitiveness to the parameter σ, we also propose to apply the L_2 normalization to the vector $\mathbf{d}$ defined in the model in the transformed space. As a result, we discover that the value of σ does not affect the performance much in practice (see Section 5).

Another issue is that in the new transformed space, we can discard some distant neighbors which have trailing coefficients that may have adverse impact on the performance. We therefore propose a coefficients cut-off method, where only the top k largest values of v_i for each class are kept so that the classifier can be computed more efficiently and the density estimation is more reliable.

The LLKNNC thus is defined as follows

$$c^* = \arg\max_c \sum_{\substack{(\mathbf{b}_i \in \mathbf{B}_c) \wedge \\ (v_i \in T(k))}} v_i \tag{10}$$

where T(k) is the set of top k largest values of v_i for each class. Similarly, the LLNMC is defined as follows:

$$c^* = \arg\min_c ||\mathbf{x} - \sum_{\substack{(\mathbf{b}_i \in \mathbf{B}_c) \wedge \\ (v_i \in T(k))}} v_i \mathbf{b}_i||_2^2 \tag{11}$$

In practice, the value of k is important to the performance (see Section 5).

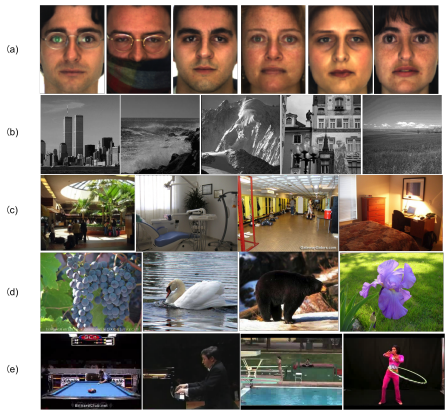

Figure 2. Sample images of the datasets: (a) the AR face database, (b) the 15 scenes dataset, (c) the MIT-67 indoor scenes dataset, (d) the Caltech 256 dataset and (e) the UCF50 dataset.

4. Experiments

In this section, we evaluate the performance of the proposed LLKNN model and two classifiers: the LLKNNC and the LLNMC on several visual recognition databases: face recognition on the AR face database [17]; scene recognition on the 15 scenes dataset [13] and the MIT-67 indoor scenes dataset [21]; object recognition on the Caltech 256 dataset [10]; and action recognition on the UCF50 dataset [22]. Some sample images are shown in figure 2.

The image or video is first represented as a pattern vector. In order to conduct fair comparison or achieve comparable results to the state-of-the-art methods, we use different pattern vectors for different datasets. Please see more details in the corresponding sub-sections. And the marginal Fisher analysis (MFA) with the principal component analysis (PCA) is used to reduce the dimension and extract features.

4.1. Face Recognition

We evaluate the proposed method on face recognition by using the AR face database, which is composed of over 4000 frontal view images for 126 individuals each with 26 pictures taken in two separate sessions. A subset of the data [17], which consists of 50 male and 50 female images with dimension 165*120, is chosen from the original database. Then, we follow the experimental settings defined in [31] and [34] to make fair comparisons that 14 images with only illumination change and expressions are selected for each person: the seven images from session 1 for training and the other seven from session 2 for testing. Before applying the proposed method, the dimension of the face vector is reduced to 180.

The model parameters are selected as $\sigma = 1$, $\lambda = 0.02$, $\alpha = 0.1$, and $\beta = 1.5$ for the proposed model. For the shifted power transformation, $\lambda_1 = 0.0$ and $\lambda_2 = 0.9$. For the LLKNNC, the value of $k = 5$ and for the LLNMC, the value of $k = 7$. The results that are presented in table 1 show that the proposed method is able to improve upon the other popular methods significantly.

Experimental setting 1	Accuracy %
D-KSVD [37]	85.40
LC-KSVD [11]	89.7
JDL [38]	91.7
FDDL [34]	92.00
SRC [31]	94.99
The proposed LLNMC	**96.14**
The proposed LLKNNC	**97.00**

Table 1. Comparisons between the proposed LLKNNC, LLNMC and the other popular methods on AR face database.

Methods	Accuracy %
KSPM [13]	81.40 ±0.50
ScSPM [32]	80.28 ±0.93
LLC [30]	80.57 ±–
KC [9]	76.67 ±0.93
D-KSVD [37]	89.10
LC-KSVD [11]	90.40
LaplacianSC [8]	89.7
The proposed LLNMC	**97.45**±0.27
The proposed LLKNNC	**93.54**±0.45

Table 2. Comparisons between the proposed LLKNNC, LLNMC and the other popular methods on the 15 scenes dataset

4.2. Scene Recognition

4.2.1 The 15 Scenes Dataset

The 15 scenes dataset [13] contains totally 4485 images from 15 scene categories, each with the number of images ranging from 200 to 400. Following the experimental protocol defined in [13] and [32], we randomly select 100 images per class for training and the remaining for testing for 10 iterations. First, we use the spatial pyramid representation provided by [11] to represent the image as a vector with the dimension of 3000. The representation is obtained by using a four-level spatial pyramid and a codebook with a size of 200. Then the image vector is further reduced to dimension 500. For the shifted power transformation, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$. The model parameters are selected as $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 1.0$. For the LLKNNC, the value of $k = 1$ and for the LLNMC, the value of $k = 2$. It can be concluded from the results in table 2 that our proposed method is able to achieve much better results than the non-linear or linear kernel based support vector machine, which is used by the compared methods.

Methods	Mean Accuracy %
ROI + Gist [21]	26.1
DPM [19]	30.4
Object Bank [14]	37.6
miSVM [15]	46.4
D-Parts [25]	51.4
DP + IFV [12]	60.8
CNN-SVM no Aug [24]	58.4
The proposed LLNMC	59.12
The proposed LLKNNC	58.18

Table 3. Comparisons between the proposed LLKNNC, LLNMC and the other popular methods on the MIT-67 indoor scenes dataset

4.2.2 The MIT-67 indoor scenes Dataset

The MIT-67 indoor scenes dataset [21] is a very challenge scene recognition dataset, which contains 67 indoor categories with 15620 images. We follow the commonly used experimental setting [21], wherein 80*67 images are used for training and 20*67 images for testing. The performance is measured as the average classification accuracy over all the categories. We consider the Fisher vector feature [26] for representation. The SIFT feature is first projected to 80 dimension and a codebook with 256 visual words is computed, then the dimension of the Fisher vector is 2*256*80 = 40960. Then we further reduce the dimension to 2000. For the shifted power transformation, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$. The model parameters are selected as $\lambda = 0.01$, $\alpha = 0.1$, and $\beta = 1.5$ for the LLNMC while $\beta = 0.5$ for the LLKNNC for the best performance. For both the LLKNNC and the LLNMC, the value of $k = 20$. The results in table 3 shows that the proposed method is able to achieve comparable results to the support vector machine, which is used by the compared methods. And by borrowing the power of the Fisher vector feature, we are able to achieve very competitive results on the challenge MIT-67 indoor scenes dataset.

Please note that we learn the Fisher vector directly from the SIFT features of the images instead of learning the part detectors in [12]. Moreover, we reduce the dimension of Fisher vector 40960 to 2000, which saves much storage space. And no data augmentation technique is used. However, we can still achieve very competitive results to the state-of-the-art methods [12].

4.3. Object Recognition

The Caltech 256 dataset [10] contains 30607 images divided into 256 object categories and a clutter class. We follow the common experimental settings [30] that 15, 30, 45, 60 images per category are selected randomly for training and no more than 25 images for testing for 3 iterations. In order to achieve comparable results to the state-of-the-art methods, the proposed method is built upon the 4096 dimension vector that is extracted by using a pre-trained convolutional neural network CNN-M [4]. For shifted power transformation, $\lambda_1 = 0.0$ and $\lambda_2 = 0.5$. We further reduce the dimension to 1000. The model parameters are selected as $\sigma = 1.5$, $\lambda = 0.01$, $\alpha = 0.1$, and $\beta = 1.5$. For both the LLKNNC and the LLNMC, the value of $k = 15$. The results that are shown in table 4, demonstrate the proposed method is able to be comparable to the other popular methods with the support vector machine for classification in all the training image sizes.

training images	15	30	45	60
ScSPM [32]	27.73	34.02	37.46	40.14
LLC [30]	34.36	41.19	45.31	47.68
IFV [20]	34.70	40.80	45.00	47.90
Bo et al. [3]	40.50	48.00	51.90	55.20
Zeiler [36]	65.70	70.60	72.70	74.20
The LLNMC	68.32	71.89	74.13	75.47
The LLKNNC	68.55	72.09	74.07	75.36

Table 4. Comparisons between the proposed method and the other popular methods on the Caltech 256 dataset.

Please note that the proposed method does not make use of the data augmentation and fine-tuning techniques [4]. However we can still achieve the comparable results to the state-of-the-art methods [36], [4].

4.4. Action Recognition

We use the action recognition dataset: the UCF50 dataset [22], which is a large scale video dataset for action recognition collected from YouTube, for assessing the proposed method. It consists of 50 action categories with a total of 6676 videos and with a minimum of 100 videos for each action class. We follow the experimental setting provided in [23], which divides the dataset into 5 groups with similar size of data and uses the 5-fold group-wise cross-validation. Please note that this setting is more challenge than the leave-one-out-cross-validation with 25 folds proposed in [22]. The action bank feature [23] is applied for representing the video data with dimension 14965 that is further reduced to 500. For the shifted power transformation, $\lambda_1 = 0.01$ and $\lambda_2 = 0.8$. The model parameters are selected as $\sigma = 1.5$, $\lambda = 0.05$, $\alpha = 0.1$, and $\beta = 1.5$. For both the LLKNNC and the LLNMC, the value of $k = 10$. The results in table 5 demonstrate that the proposed method can improve upon other popular methods a lot.

5. Analysis

In this section, we provide more comprehensive analysis of the proposed method concerning about the performance.

Methods	Accuracy %
GIST [18]	38.8
Wang et al. [29]	47.9
Action bank [23]	57.9
SRC [31]	59.6
D-KSVD [37]	38.6
LC-KSVD [11]	53.6
JDL [38]	53.5
FDDL [34]	61.1
The proposed LLNMC	**62.42**
The proposed LLKNNC	**62.66**

Table 5. Comparisons between the proposed LLKNNC, LLNMC and the other popular methods on the UCF 50 action recognition dataset

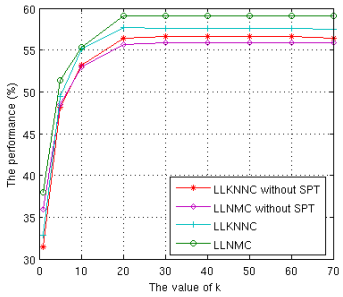

Figure 3. The assessment of the effectiveness of shifted power transformation under different values of the parameter k on the MIT-67 indoor scenes dataset

Particularly, we evaluate the proposed method on the following critical issues: (1) the effectiveness of the shifted power transformation, (2) the sensitiveness to the parameter σ, (3) the sensitiveness of the parameter k, (4) the comparison to plain KNN classifier, and (5) the evaluation of grouping effect of the nearest neighbors. And all the experimental settings in this section are the same as the ones used in the above experimental section unless otherwise specified.

5.1. The effectiveness of the shifted power transformation

First, we evaluate the effectiveness of the shifted power transformation (SPT) by comparing the results of "LLKNNC" (already with SPT), "LLKNNC without SPT", "LLNMC" (already with SPT) and "LLNMC without SPT" under different values of the parameter k, which is defined in Section 3.3. Without specification, when we denote the method as "LLKNNC" or "LLNMC", the shifted power transformation is already applied. As shown in figure 3, the performance is indeed improved by using the shifted power transformation under all the values of k. Note that the other parameters are fixed.

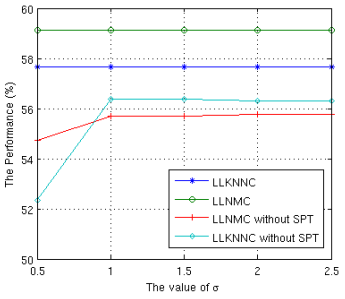

Figure 4. The assessment of the sensitiveness to different values of the parameter σ on the MIT-67 indoor scenes dataset

5.2. The sensitiveness to the parameter σ

As proposed in Section 3.3, the shifted power transformation and the L_2 normalization are able to alleviate the sensitiveness of the parameter σ, which represents the global window width for kernel density estimation. As shown in figure 4, we demonstrate the results of "LLKNNC" and "LLNMC" under different values of the parameter σ. The parameter σ truly does not affect the performance of classification much since we have applied both the shifted power transformation and the L_2 normalization to the vector $\mathbf{d}$. However, without the the shifted power transformation, the performance relies on the value of σ much. We also discover that the L_2 normalization to the vector $\mathbf{d}$ is necessary to guarantee the performance, otherwise the performance will drop below 10%.

To further show that the sensitiveness to the parameter σ can be alleviated by our proposed method, some extreme values of σ such as 20, 30 are also evaluated. The performance of LLKNNC is 58.40 (even better than that reported above) and the performance of LLNMC is 58.49 for both values.

5.3. The sensitiveness to the parameter k

As proposed in Section 3.3, the coefficients cut-off method is able to discard the longer distance neighbors which contribute trailing coefficients that may have adverse impact on the performance. Thus, we evaluate the performance when the value of the parameter k varies and all the other parameters are fixed. The results shown in figure 3 demonstrate the importance of the value of k that we cannot use too small value of k and also there is no need to use larger value of k.

5.4. Comparison to KNN Classifier

We compare our proposed classifiers with the KNN (K = 3) classifier as the amount of training data varies in this section. Results in table 6 demonstrate that our proposed classifiers can improve upon the plain KNN classifier significantly.

training images	15	30	45	60
KNN	62.22	65.68	67.30	68.59
The LLNMC	68.32	71.89	74.13	75.47
The LLKNNC	68.55	72.09	74.07	75.36

Table 6. Comparisons between the proposed method and the plain KNN classifier on the Caltech 256 dataset.

5.5. The Grouping Effect of the Nearest Neighbors

In this section, we evaluate the tightness of the bound of the grouping effect of the nearest neighbors with the goal of showing that the proposed LLKNN model is able to derive the representation that approximates the ideal representation in terms of two measurements: the true activation ratio (TAR) and the false activation ratio (FAR).

Given the number of test samples N_{test} and the number of the expected non-zero coefficients t_i for the i-th test sample, the first measurement: the true activation ratio (TAR), is defined as follows

$$TAR = \frac{\sum_{i=1}^{N_{test}} t_i}{N_{test}} \tag{12}$$

which means the average numbers of the expected coefficients that are activated (non-zero) for all the test samples. Ideally, the TAR is the size of the training samples of the c-th class if the test sample is from class c.

The second measurement is the false activation ratio (FAR)

$$FAR = \frac{\sum_{i=1}^{N_{test}} f_i}{(c-1)N_{test}} \tag{13}$$

where c is the number of classes and f_i is the number of the non-expected non-zero coefficients for the i-th test sample. The FAR represents the average numbers of the non-expected coefficients that are activated over all the other false classes $(c-1)$ for all the test samples. We expect the proposed model to keep a higher value TAR that is close to the size of the training samples in each class and a lower value of FAR that is close to 0.

As shown in table 7, we present the accuracy, the value of FAR and the value of TAR when the parameter λ, α and β changes. It can be observed that sometimes increasing (decreasing) the value of TAR will also result in the increasing (decreasing) of the value of FAR, which may degrade the performance. Therefore a trade-off between the value of FAR and the value of TAR is necessary to achieve the best performance.

6. Conclusion

This paper presents a novel locally linear KNN model for robust visual recognition. The theoretical analysis shows that the derived representation has the grouping effect of the nearest neighbors, which is able to approximate the "ideal representation". And then the locally linear KNN based classifier (LLKNNC), which is proved to approximate the Bayes classifier in the view of kernel density estimation, is proposed for classification. Besides, the locally linear nearest mean classifier (LLNMC), is also proposed. Furthermore, the shifted power transformation and the coefficients cut-off method are used to improve the performance of the proposed classifiers. The effectiveness of the proposed model is evaluated on several representative visual recognition databases, and the experimental results show that the proposed model outperforms some other representative popular methods.

λ	α	β	LLKNNC	LLNMC	TAR	FAR
0.01	0.1	1.5	57.65	59.12	30.60	15.35
0.05	0.1	1.5	57.13	56.06	11.05	1.44
0.10	0.1	1.5	49.15	48.60	4.35	0.16
0.01	0.1	1.5	57.65	59.12	30.60	15.35
0.01	0.3	1.5	56.91	58.52	35.92	18.33
0.01	0.5	1.5	55.95	57.88	39.91	20.63
0.01	0.1	0.5	58.18	58.60	30.45	15.34
0.01	0.1	1.0	57.80	58.82	30.52	15.34
0.01	0.1	1.5	57.65	59.12	30.60	15.35
0.01	0.1	2.0	57.95	59.12	30.66	15.36

Table 7. The accuracy, the value of FAR and the value of TAR when the parameter λ, α and β changes on the MIT-67 indoor scenes dataset

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009. 3

[2] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, 1997. 1

[3] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–667, 2013. 6

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014. 6

[5] W. Deng, J. Hu, and J. Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1864–1870, 2012. 2

[6] W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In *CVPR*, pages 399–406, 2013. 2

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000. 3, 4

[8] S. Gao, I. W.-H. Tsang, and L.-T. Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):92–104, 2013. 5

[9] J. C. Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, pages 696–709, 2008. 5

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. 1, 5, 6

[11] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2651–2664, 2013. 1, 2, 5, 7

[12] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 923–930, 2013. 6

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006. 1, 5

[14] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, pages 1378–1386, 2010. 6

[15] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 851–858, 2013. 6

[16] C. Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(5):725–737, 2006. 1

[17] A. M. Martínez and A. C. Kak. Pca versus lda. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2):228–233, 2001. 1, 5

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 7

[19] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 1307–1314, 2011. 6

[20] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, pages 143–156, 2010. 6

[21] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 413–420, 2009. 1, 5, 6

[22] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.*, 24(5):971–981, 2013. 1, 5, 6

[23] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6, 7

[24] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014. 6

[25] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, pages 3400–3407, 2013. 6

[26] J. Snchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013. 6

[27] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991. 1

[28] R. D. Wand M.P., Marron J.S. Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353, 1991. 4

[29] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009 - British Machine Vision Conference*, pages 124.1–124.11, 2009. 7

[30] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010. 5, 6

[31] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009. 1, 2, 3, 4, 5, 7

[32] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801, 2009. 5, 6

[33] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, pages 543–550, 2011. 2

[34] M. Yang, L. Zhang, X. Feng, and D. Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, pages 1–24, 2014. 1, 2, 5, 7

[35] D. You, O. C. Hamsici, and A. M. Martínez. Kernel optimization in discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):631–638, 2011. 1

[36] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014*, pages 818–833, 2014. 6

[37] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, pages 2691–2698, 2010. 1, 2, 5, 7

[38] N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *CVPR*, pages 3490–3497, 2012. 2, 5, 7

[39] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005. 2