

# Non-Linear Feature Extraction by Linear PCA Using Local Kernel

Kazuhiro Hotta

*The University of Electro-Communications)*

*hotta@ice.uec.ac.jp http://www.htlab.ice.uec.ac.jp/~hotta/*

## Abstract

*This paper presents how to extract non-linear features by linear PCA. KPCA is effective but the computational cost is the drawback. To realize both non-linearity and low computational cost simultaneously, the idea of local kernel is used. The mapped features of the polynomial kernel can be described explicitly. When input features are divided into some local features and the polynomial kernel is applied to each local features independently, the dimension of mapped features does not become so high. In addition, the inner product with all local mapped features corresponds to the local summation kernel. Thus, KPCA with the local summation kernel can be solved by linear PCA. The proposed approach is evaluated in object categorization problem which requires high non-linearity and computational cost. The proposed method gives much higher accuracy than linear PCA. The computational cost is lower than KPCA though the accuracy is slightly worse than KPCA.*

## 1. Introduction

In the last decade, the effectiveness of kernel-based methods for object detection and recognition have been reported [1, 2, 3]. In particular, Kernel Principal Component Analysis (KPCA) took the place of traditional linear PCA as the first feature extraction step in various research and applications. KPCA can cope with non-linear variations well. However, KPCA must solve the eigen value problem with the number of samples  $\times$  the number of samples. In addition, the kernel computations with all training samples are required to map a test sample to the subspace obtained by KPCA. Therefore, the computational cost is the main drawback. To reduce the computational cost of KPCA, sparse KPCA [4] and the use of clustering [5] were proposed. Ichino et al. [5] reported that KPCA of cluster centers is more effective than sparse KPCA. However, the computational cost becomes a big problem again when the number of classes is large and each category has one subspace. For example, KPCA of visual words (cluster centers of local features) [6] is effective in object categorization using Caltech 101 dataset [7] but the computational cost is high. In this method, each category has one subspace constructed by 400 visual words. Namely, 40, 400

(= 101 categorizes  $\times$  400 visual words) kernel computations are required to map a local feature to the subspace.

On the other hand, traditional linear PCA is independent of the number of samples when the dimension of input features is not so high. This is because the size of eigen value problem depends on the minimum number of dimension of input features and the number of samples. To map a test sample to the subspace, only inner products between basis vectors and the test sample are required. Therefore, in general, the computational cost of linear PCA is much lower than KPCA. In this paper, we propose how to use non-linearity of KPCA and computational cost of linear PCA simultaneously.

Kernel-based methods map training samples to high dimensional space  $\mathbf{x} \rightarrow \phi(\mathbf{x})$ . Non-linearity is realized by linear method in high dimensional space. The dimension of mapped feature space of the RBF kernel becomes infinity, and we can not describe the mapped features explicitly. However, the mapped feature  $\phi(\mathbf{x})$  of the polynomial kernel can be described explicitly. This means that KPCA with polynomial kernel is solved by linear PCA of mapped features. Unfortunately, in general, the dimension of mapped feature is too high to solve by linear PCA even when the polynomial kernel with 2nd degrees ( $K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T \mathbf{y})^2$ ) is used. The dimension of mapped features of the polynomial kernel with 2nd degrees becomes  $_{nd+2}C_2$  where  $nd$  is the number of dimension of input features. For example, the dimension of mapped features becomes 20, 301 even when the dimension of input features is 200. However, if polynomial kernel with 2nd degrees is applied to local features not whole features, the dimension of mapped features is not so high. For example, when the polynomial kernel with 2nd degrees is applied to each local 10 dimensional features of 200 dimensional input features without overlap, each local features  $\mathbf{x}_{li}$  is mapped to 66 dimensional features  $\phi(\mathbf{x}_{li})$  independently. Namely, the 200 dimensional input features is mapped to 1, 320 (= 66 dimensions  $\times$  20 local features) dimensional features  $(\phi(\mathbf{x}_{l1})^T, \dots, \phi(\mathbf{x}_{l20})^T)^T$ . In fact, this corresponds to the local summation kernel (the summation of local kernels) [3] because the inner product between 1, 320 dimensional features is the summation of the outputs of inner product between 66 dimensional features as  $\sum_i^{20} \phi(\mathbf{x}_{li})^T \phi(\mathbf{y}_{li})$ . This shows that KPCA with the local summation kernel can be solved by linear PCA.

This approach is independent of the number of training samples. Subspace is obtained by solving the eigen value problem of mapped features (e.g. 1, 320 dimensions). To map a test sample to the subspace, only the inner products with basis vectors are required. In addition, it can represent non-linear distribution while the computational cost is low. Furthermore, it is reported that local summation kernel outperforms standard RBF kernel and polynomial kernel under partial occlusion [3].

We evaluate the proposed approach in object categorization problem using Caltech 101 dataset[7]. We demonstrate that the proposed approach gives much higher recognition rate than linear PCA. The computational cost is lower than KPCA while the accuracy is slightly worse than KPCA. The accuracy of our approach is also comparable with object categorization methods published in recent years [8, 9, 10, 7].

In section 2, the details of the proposed method are explained. Object categorization method using proposed approach is explained in section 3. Experimental results using Caltech 101 database are shown in section 4. Finally, conclusion and future works are described in 5.

## 2. KPCA with local summation kernel by linear PCA

This section explains how to solve KPCA with the local summation kernel by linear PCA. At first, the input features is divided into  $N$  local features  $\mathbf{x} = (\mathbf{x}_{11}^T, \mathbf{x}_{12}^T, \dots, \mathbf{x}_{1N}^T)^T$ . In the following experiments, the division is performed without overlap. We consider that structural local features (e.g. local parts of image) are more effective than meaningless local features selected randomly. Each local features  $\mathbf{x}_{li}$  is mapped to high dimensional features  $\phi(\mathbf{x}_{li})$ . All mapped local features  $\phi(\mathbf{x}_{11}), \dots, \phi(\mathbf{x}_{1N})$  are connected and used as a new feature vector  $\mathbf{x}' = (\phi(\mathbf{x}_{11})^T, \dots, \phi(\mathbf{x}_{1N})^T)^T$ . The inner product of new feature vectors corresponds to the summation of local kernels.

$$\begin{aligned} \mathbf{x}'^T \mathbf{y}' &= (\phi(\mathbf{x}_{11})^T, \dots, \phi(\mathbf{x}_{1N})^T)(\phi(\mathbf{y}_{11})^T, \dots, \phi(\mathbf{y}_{1N})^T)^T, \\ &= \sum_i^N \phi(\mathbf{x}_{li})^T \phi(\mathbf{y}_{li}) \end{aligned} \quad (1)$$

Therefore, if local mapped features  $\phi(\mathbf{x}_{li})$  can be described explicitly, KPCA with the local summation kernel can be solved by linear PCA of new features  $\mathbf{x}'$ . Basis vectors are obtained by solving the eigen value problem of the covariance matrix  $C = X X^T$  where  $X = (\mathbf{x}'_1, \dots, \mathbf{x}'_L)$ . The dimension of basis vectors is the same as  $\mathbf{x}'$ .

Note that mapped features of the polynomial kernel can be described explicitly. In the case of the polynomial kernel with 2nd degrees, the dimension of mapped features  $\phi(\mathbf{x})$  becomes  $nd+2C_2$  where  $nd$  is the number of dimension of input feature  $\mathbf{x}$ . For example, a input feature with 2 dimensions  $\mathbf{x} = (x_1, x_2)$  is mapped to 6 dimensional features as  $x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1$ . Thus, when the dimension of local features to which kernel is applied is not large, KPCA with the local summation kernel can

be solved by linear PCA. Of course, if the dimension of local features is large, the dimension of mapped features becomes large. However, the size of covariance matrix in linear PCA depends on the minimum value of dimension of features and the number of samples. When the number of samples is smaller than dimension of features, we solve the eigen value problem of  $D = X^T X$  where  $X = (\mathbf{x}'_1, \dots, \mathbf{x}'_L)$ , and eigen vectors can be computed from the eigen vectors  $B$  of  $D$  as  $A = X B \Lambda^{-1/2}$  where diagonal elements of  $\Lambda$  is the eigen values.

In recent years, it is reported that the normalized polynomial kernel outperforms standard polynomial kernel and RBF kernel after setting optimal parameters [11]. Therefore, in the following experiments, we use the normalized polynomial kernel with 2nd degrees as local kernel function. The dimension of mapped features  $\phi(\mathbf{x})$  is the same as the standard polynomial kernel. The difference is the norm of mapped features. In standard polynomial kernel, the norm of mapped features  $\|\phi(\mathbf{x})\|$  is not normalized. However, the normalized polynomial kernel normalizes the norm of mapped features as  $\phi(\mathbf{x})/\|\phi(\mathbf{x})\|$ . By normalizing the norm of mapped features, the inner product  $\frac{\phi(\mathbf{x})^T \phi(\mathbf{y})}{\|\phi(\mathbf{x})\| \|\phi(\mathbf{y})\|}$  is between 0 and 1 such as RBF kernel.

In this paper, KPCA with the summation kernel of local normalized polynomial kernels is solved by linear PCA of new features  $\mathbf{x}'$ . By this approach, we can treat many samples easily which are not able to treat in KPCA by the computational cost and memory required. In addition, it can represent non-linear distribution while linear PCA can not. In addition, the computational cost for mapping a test sample to the subspace is much lower than KPCA because kernel computations with all training samples are not required in the proposed approach. The test sample is mapped to the subspace by inner products with basis vectors whose dimension is the same as new features  $\mathbf{x}'$ . Therefore, it is effective for multi-class classification problem such as object categorization. In this paper, the proposed approach is evaluated in object categorization problem using Caltech 101 dataset. We compare it with conventional method based on KPCA and linear PCA of visual words [6]. The comparison with other methods is also done. In section 3, object categorization method using the proposed approach is explained.

## 3. Object categorization method

In object categorization problem, we can not know the position of object in advance. Therefore, characteristic local features (regions) are selected automatically from training images of each category. Visual words of each category are made by applying clustering to the ensemble of local features of each category. In [6], KPCA with the normalized polynomial kernel with 5 degrees is used to represent the ensemble of visual words of each category. After extracting features specialized for each category, linear SVM is used. In this paper, KPCA with the

local summation kernel which is solved by linear PCA is used, and we evaluate whether our approach can represent non-linear variations with low computational cost.

First, the characteristic local regions are selected automatically by using Harris operator. The orientation histogram of multi-resolution Gabor features is used to describe each local region. Figure 1 shows how to make orientation histogram. Concretely, Gabor filters of 8 different orientations with 3 scales are used. Gabor features of 9 (height)  $\times$  9 (width)  $\times$  3 (scales)  $\times$  8 (orientations) dimensions are extracted from a local region. Orientation histogram is computed in non-overlap local region of 3  $\times$  3 pixels of each scale independently. As a result, we obtain 216 (= 3  $\times$  3  $\times$  3  $\times$  8) dimensional features. These are used as the descriptor of a local region.

In the proposed approach, local kernel is applied to each 8 dimensional orientation histogram. Since the normalized polynomial kernel with 2nd degrees is used, each 8 dimensional orientation histogram  $\mathbf{x}_{li}$  is mapped to 45 dimensional features  $\phi(\mathbf{x}_{li})$ . Thus, the dimension of new features  $\mathbf{x}'$  becomes 1,215 dimensions (= 45 dimensions  $\times$  3 (height)  $\times$  3 (width)  $\times$  3 (scales)).

After describing the local features, k-means is adopted to obtain visual words (cluster centers) of each category. In the following experiments, 400 visual words for each category are used. Since the visual words include various kinds of local features, the distribution becomes non-linear. To represent the ensemble of visual words of each category well, non-linearity of KPCA is required. Therefore, linear PCA of visual words will not give good accuracy. In the proposed approach, non-linearity is realized by local mapped features (local summation kernel). In our object categorization method, each category has one subspace. In standard KPCA, kernel computations with all visual words in certain category to compute the covariance matrix. On the other hand, the proposed approach can compute the covariance matrix by the inner product of 1,215 dimensional visual words. The maximum benefit of our approach in object categorization based on KPCA of visual words is obtained in mapping a test sample to the subspace. In standard KPCA, 40,400 (= 101 categories  $\times$  400 visual words) kernel computations is required to map a local feature to all subspaces. However, our approach does not need to compute 40,400 kernel computations. Only inner products with basis vectors are required to map a local features to the subspace. Therefore, the computational cost is much lower than conventional method.

After mapping a local features into the subspace specialized for each category, linear SVM is used to classify object categories. Since non-linear features are extracted by KPCA, linear SVM is sufficient. In this paper, the projection length to the subspace is used as the features for linear SVM. However, the projection values to the principal component axes with lower rank become small. To compensate it, the projection value to each axis is normalized by standard deviation of training samples at the axis. The normalized projection length of  $q$ -th local features to

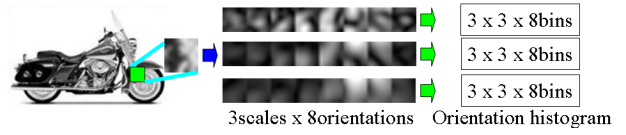


Figure 1. How to describe a local feature

the  $p$ -th axis of the subspace of category  $k$  is defined as  $z'_{kq} = \left( A_{kp}^T \mathbf{x}'_q / \sqrt{\sigma_{kp}^2} \right)^2$  where  $\sigma_{kp}^2$  is the variance of  $p$ -th principal component axis of category  $k$ . Since  $M$  local features are obtained from a test image, we must integrate  $M$  projection length to be robust to the order and the number of local features. We use the mean projection length to integrate all local features. Namely,  $p$ -th feature in the subspace of category  $k$  is computed as  $z'_k = \frac{1}{M} \sum_q z'_{kq}$ . Of course, the mean projection length is invariant to the number and the order of local features. These features are used in linear SVM. We use one-against-all SVM to treat multi-categories. A test image is fed into all  $NC$  SVMs, and it is classified to the category given maximum output.

#### 4. Evaluation using Caltech 101 database

The proposed approach is evaluated using Caltech 101 database [7] which is frequently used in recent papers [8, 9, 12, 13, 14, 10]. The number of images in each category is different. The minimum is 31 and the maximum is 800. Since many conventional methods evaluate the performance when 30 images are used in training, we also use 30 images selected randomly in training. All remaining images of all categories are used for evaluation. To reduce the bias of the different number of test images in each category, the mean of the classification rate of each category is used. This evaluation is repeated 3 times with different initial seeds for random function, and the mean classification rate of 3 runs is used as a final result.

In this paper, all images are transformed to gray-level images. The image size is normalized so that all images have the nearly same area. Serre et al. [8] and Mutch et al. [9] also normalize the image size because the parameters of Gabor filters are fixed. The number of local features cropped from an image is set to 500 empirically. When the number of visual words becomes large, the computational cost and required memory of standard KPCA is very large. Therefore, the number of visual words is set to 400. Although the proposed approach is independent of the number of visual words, the same number of visual words is used for fair comparison.

The result of KPCA (the normalized polynomial kernel with 5 degrees of input features  $\mathbf{x}$  in [6]) of visual words, linear PCA of visual words and the proposed method are shown in Table 1. The accuracy of linear PCA of visual words is very low. This means that non-linearity of visual words of each category is high, and linear PCA can not represent it well. The proposed method gives much higher

**Table 1. Comparison result**

Method	Recognition rate
<b>Proposed method (K=400)</b>	54.8%
linear PCA of visual words (K=400)	36.2%
KPCA of visual words (K=400) [6]	60.0%
Wang et al. (CVPR06) [10]	63%
Grauman et al. (ICCV05) [12]	58.23%
Mutch et al. (CVPR06) [9]	56%
Serre et al. (CVPR05) [8]	42%

accuracy than linear PCA. This shows that the proposed method can represent non-linear distribution though each subspace is constructed by linear PCA. However, the accuracy is worse than that of KPCA of visual words while the computational cost is lower than KPCA<sup>1</sup>. The one reason is that KPCA of visual words [6] uses the normalized polynomial kernel with 5 degrees. The non-linearity of the proposed method based on the normalized polynomial kernel with 2nd degrees may not be slightly enough. However, the proposed approach has a chance for improving the accuracy further with low computational cost. One direction is to increase the number of visual words. In this paper, the number of visual words is set to 400 because of the computational cost and memory required of KPCA of visual words [6]. However, the proposed method is independent of the number of visual words. In general, better subspace will be constructed when the number of training samples is large. This is a subject for future works.

Finally, we compare our method with conventional methods published in recent years. The objects in Caltech 101 database are located at the almost same position. Therefore, the use of absolute position information much improves the accuracy [13, 14]. Since our method does not use absolute position information, the comparison with only conventional methods without absolute position information is performed. Table 1 also shows the comparison result when the number of training images is 30. The proposed method is superior to linear SVM of biological inspired features [8]. It is comparable with [12, 9] though it is worse than Wang’s approach [10]. This result shows the effectiveness of modeling of ensemble of visual words by KPCA with the local summation kernel.

## 5. Conclusions and future works

We propose how to solve KPCA with the local summation by linear PCA. In the classification process, KPCA must compute kernel functions with all training samples (e.g. cluster centers), and the computational cost and memory required are high. This is the drawback. In this paper, input features are divided into some local features, and local features are mapped to higher dimensional space

<sup>1</sup>The computational time of [6] and the proposed method is 80 and 20 seconds respectively on a standard PC with Xeon 2.0 GHz CPU.

by  $\phi(\mathbf{x}_{li})$ . In this formulation, the dimension of new feature vector  $(\phi(\mathbf{x}_{l1})^T, \dots, \phi(\mathbf{x}_{lN})^T)^T$  is not so high. Thus, we can use linear PCA of new features directly. In fact, linear PCA of new features corresponds to the KPCA with the local summation kernel. Effectiveness of the proposed method is shown in object categorization problem. In the experiment, the number of visual words is set to small value for fair comparison with conventional method while the proposed is independent of the number of visual words. Evaluation using the larger number of visual words is a subject for future works.

The proposed method is a general framework which is independent of recognition tasks. It can be apply to other recognition problem. In addition, the proposed idea is also applicable to SVM and Kernel Fisher Discriminant Analysis without any changes. This means that kernel-based methods with local summation kernel can be solved by linear method.

## References

- [1] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.
- [2] M.-H. Yang, “Face recognition using kernel methods,” in *Advances in Neural Information Processing Systems 14*, pp. 215–220, 2002.
- [3] K. Hotta, “Robust face recognition under partial occlusion based on support vector machine with local gaussian summation kernel,” *Image and Vision Computing*, in press.
- [4] M. Tipping, “Sparse kernel principal component analysis,” in *Advances in Neural Information Processing Systems 13*, pp. 633–639, 2001.
- [5] M. Ichino, H. Sakano, and N. Komatsu, “A study on speed up of kernel mutual subspace method,” in *Proc. Meeting on Image Recognition and Understanding*, pp. 1035–1042, 2005.
- [6] K. Hotta, “Object categorization based on kernel principal component analysis of visual words,” in *Proc. IEEE Workshop on Applications of Computer Vision*, 2008.
- [7] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE Trans. Pattern Analysis and Machine Intelligence* **28**(4), pp. 594–611, 2006.
- [8] T. Serre, L. Wolf, and T. Poggio, “Object recognition with features inspired by visual cortex,” in *Proc. IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 994–1000, 2005.
- [9] J. Mutch and D. Lowe, “Multiclass object recognition using sparse, localized features,” in *Proc. IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 11–18, 2006.
- [10] G. Wang, Y. Zhang, and L. Fei-Fei, “Using dependent regions for object categorization in a generative framework,” in *Proc. IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, 2006.
- [11] R. Debnath and H. Takahashi, “Kernel selection for the support vector machine,” *IEICE Trans. Info. & Syst.* **E87-D**(12), pp. 2903–2904, 2004.
- [12] K. Grauman and T. Darrell, “Discriminative classification with sets of image features,” in *Proc. International Conference on Computer Vision*, pp. 1458–1465, 2005.
- [13] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.
- [14] H. Zhang, A. Berg, M. Maire, and J. Malik, “Svm-knn: Discriminative nearest neighbor classification for visual category recognition,” in *Proc. IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 2126–2136, 2006.