2016 15th IEEE International Conference on Machine Learning and Applications

Robust Kernel Embedding of Conditional and Posterior Distributions with Applications

Muhammad Zeeshan Nawaz and Omar Arif
School of Electrical Engineering and Computer Science
National University of Sciences and Technology
H-12 Sector, Islamabad
Email: 13mscsmnawaz@seecs.edu.pk, omar.arif@seecs.edu.pk

***Abstract*—This paper proposes a novel non-parametric method to robustly embed conditional and posterior distributions to reproducing Kernel Hilbert space (RKHS). Robust embedding is obtained by the eigenvalue decomposition in the RKHS. By retaining only the leading eigenvectors, the noise in data is methodically disregarded. The non-parametric conditional and posterior distribution embedding obtained by our method can be applied to a wide range of Bayesian inference problems. In this paper, we apply it to heterogeneous face recognition and zero-shot object recognition problems. Experimental validation shows that our method produces better results than the comparative algorithms.**

I. Introduction

Many problems in machine learning involve finding interesting non-linear relationships in high dimensional data. Kernel methods in statistical learning theory provide powerful techniques for analyzing such data. The basic idea behind kernel methods is to map/embed the data non-linearly to a higher dimensional feature space (reproducing kernel Hilbert space (RKHS) $\mathcal{H}$), where linear algorithms are applied. Kernel Principal Component Analysis (KPCA) is one such algorithm, where standard PCA is applied in higher dimensional space. The inner product between the embedded points in the RKHS is not explicitly computed but is given by a positive definite kernel function $\boldsymbol{k}$ in the input space. This property, also called "*kernel trick*", allows algorithms to effectively use the higher dimensional space without explicitly dealing with Hilbert space representation.

Smola et al., [**?**] introduced the idea of *kernel mean*, which is the generalization of mapping individual data points to mapping probability distributions to RKHS. The probability distribution is mapped by taking the expectation of the embedding points and is represented as a point in the RKHS. For characteristic kernels, such as Gaussian kernel, the mapping of probability distribution is one to one, i.e. the mapping uniquely identifies the distribution. Kernel mean mapping has found numerous applications in machine learning tasks. For example [**?**] uses kernel mean mapping to analyze and compare two distributions by taking the norm of the difference between the two distribution mappings. In [**?**], [**?**], kernel mean mapping is used in conjunction with Support Vector Machines (SVM).

The idea is further extended to embed conditional distributions (*conditional kernel mean*) in the Hilbert space [**?**]. Similarly [**?**] derives an expression for realizing Bayes' rule, where prior and conditional probabilities are represented in the RKHS using kernel mean mapping. A nonparametric Hidden Markov Model based on kernel mean representation is proposed in [**?**]. These methods allow a non-parametric way of representing conditional and posterior probabilities in a supervised learning settings.

In this paper, we derive and analyze the expression for the low rank kernel embeddings of conditional and posterior distributions. The motivation for this study is that the high dimensional data encountered in many problems lies intrinsically close to a low dimensional manifold. Learning this low dimensional manifold will not only help computationally by reducing the dimensions of the problem but by keeping only the most significant dimensions, the effect of noise and outliers present in the data is also reduced. We will call this low dimensional manifold *reduced RKHS*. The process is schematically explained in Figure 1, where a data point x is mapped to the RKHS, $\mathcal{H}$ via the kernel $\boldsymbol{k}$, and then mapped to the reduced space (R^q) via eigenvalue decomposition. Likewise, probability distributions can also be mapped to the reduced RKHS, the details of which will be discussed in the following sections. Low rank embedding of distributions has been studied before. In [**?**], [**?**], the mapped points in the Hilbert space are projected to a low dimensional subspace using Kernel principal component analysis and used to estimate the probability density, which leads to improved performance in density estimation. The framework is used in the context of visual object tracking. The same expression is reached in [**?**] using orthogonal series density estimation. As an example of improved performance, see Figure 2, which we will explain in more detail in Section III. Low rank embedding of probability distribution is also studied in [**?**].

Our contribution in this paper is to derive expressions for mapping conditional and posterior probability distributions to reduced RKHS using eigenvalue decomposition of the kernel matrix. The embedding is *robust* because the most significant eigenvectors are retained and eigenvectors corresponding to lower eigenvalues are rejected. To the best of our knowledge, robust embedding of conditional and posterior probability distributions to low rank/reduced RKHS has not been studied before. The resultant algorithm has applications in a wide variety of Bayesian inference problems. In this paper, we will apply it to heterogeneous face recognition and zero shot

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.153

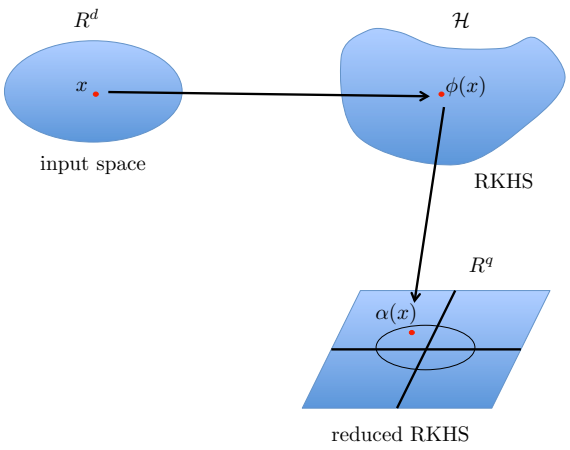

Fig. 1. RKHS

learning. To robustly embed the conditional distributions, we first perform principal component analysis using eigenvalue decomposition in the RKHS and retain the leading q eigenvectors. It is assumed that the eigenvectors corresponding to lower eigenvalues represent noise present in the data. The mapped points are then used to represent the mean mapping. All the computations in the Hilbert space are computed using the kernel function and no explicitly mapping is carried out.

The organization of this paper is as follows. Section II briefly goes over the basics of Kernel method, followed by Section III, which details our method. The algorithm is applied to two problems in computer vision in Section IV.

II. Kernel Method Basics

Let $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$ be a set of n training points drawn from the distribution P_x. A Mercer kernel is a function $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which implicitly defines a mapping $\phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space and the following equation holds:

$$\boldsymbol{k}(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle. \tag{1}$$

This means that the inner product in the RKHS, $\mathcal{H}$ is computed using the kernel $\boldsymbol{k}$ in the input space. If the finite sample of points $\{x_i\}_{i=1}^n$ are drawn i.i.d. from the distribution P_x, then the unbiased numerical estimate of the mean mapping, called the *kernel mean*, is given by

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = \frac{1}{n}\Phi 1_n, \tag{2}$$

where $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$ and 1_n is a $n \times 1$ vector of 1. The kernel mean can be used to compute the probability at a test point x by taking the inner product between the mapped test point $\phi(x)$ and the mean map [**?**]:

$$p(x) = \langle \phi(x), \mu_X \rangle = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{k}(x, x_i) = \frac{1}{n} K(x)^T 1_n, \tag{3}$$

where $K(x) = [\boldsymbol{k}(x_1, x), \ldots, \boldsymbol{k}(x_n, x)]$. This expression provides an alternative view of the kernel density estimation technique. For universal kernel functions, such as the Gaussian kernel, there is a one to one mapping between the distribution, P_x and the kernel mean mapping μ_X. This means that the distribution can be uniquely identified via its kernel mean map. In this paper, we will use the Gaussian kernel:

$$\boldsymbol{k}(x_i, x_j) = \exp\left(-\frac{1}{2}(x_i - x_j)^T \Sigma^{-1}(x_i - x_j)\right), \tag{4}$$

where Σ is a $d \times d$ dimensional covariance matrix.

A. Embedding via Eigenvalue Decomposition

One possible embedding that satisfies the inner product property of Equation (1) is obtained by the eigenvalue decomposition of kernel matrix K, with $K_{ij} = \boldsymbol{k}(x_i, x_j)$. Let E_x be a $n \times q$ matrix whose columns are the eigenvectors of the kernel matrix corresponding to the q largest eigenvalues and Λ_x be a $q \times q$ diagonal matrix of eigenvalues, then the embedding satisfying (1) is given by

$$K = E_x \Lambda E_x^T = E_x \Lambda_x^{\frac{1}{2}} (E_x \Lambda_x^{\frac{1}{2}})^T = \alpha^T \alpha, \tag{5}$$

where $\alpha = (E_x \Lambda_x^{\frac{1}{2}})^T$ is the robust embedding matrix. This embedding satisfies Equation (1) as $\alpha^T \alpha = K$.

Equation (5) provides embedding for the training points. For a new point x, the embedding is obtained using the Nystrom method [**?**]:

$$\alpha(x) = (E_x \Lambda_x^{-\frac{1}{2}})^T K(x), \tag{6}$$

where $K(x) = [\boldsymbol{k}(x_1, x), \ldots, \boldsymbol{k}(x_n, x)]$. Note that Equation (6) can also be used to compute the embedding for the training samples. Kernel embedding of data points is explained in Figure 1, where a sample point x is mapped to the RKHS via the kernel, $\boldsymbol{k}$. Furthermore, using eigenvalue decomposition, the embedding is obtained in the reduced RKHS (Equation (6)).

Before moving forward, lets summarize the notation used in the paper. Let (X, Y) be a multivariate random variable on $\mathcal{X} \times \mathcal{Y}$ with probability P and $\{x_i, y_i\}_{i=1}^n$ be n i.i.d samples. The kernel matrix for the samples of random variable X is given by K, with $K_{ij} = \boldsymbol{k}(x_i, x_j)$. The robust Hilbert space embedding is given by Equation (6). Similar notations apply to random variable Y. The notations are summarized in Table I.

III. Robust Hilbert Space Embedding of Distributions

In this section, we present the main contribution of the paper, which is the robust kernel embedding of the conditional (Section III-A) and posterior distributions (Section III-B). Recall that the kernel mean is represented as expectation in the RKHS (Equation 2). Robust kernel mean is obtained by taking the mean of the mapped points in the reduced RKHS:

$$\mu_X^r = \frac{1}{n}\sum_{i=1}^{n} \alpha(x) = a_x 1_n. \tag{7}$$

Random Variable	X	Y
Observation	x	y
Kernel matrix	K	G
Hilbert space	$\mathcal{H}_x$	$\mathcal{H}_y$
Eigenvector, Eigenvalues	E_x, Λ_x	E_y, Λ_y
Kernel embedding matrix	Φ	Ψ
Robust kernel embedding matrix	$\alpha = (E_x \Lambda_x^{\frac{1}{2}})^T$	$\beta = (E_y \Lambda_y^{\frac{1}{2}})^T$
Robust embedding of a new point	$\alpha(x) = (E_x \Lambda_x^{-\frac{1}{2}})^T K(x)$	$\beta(y) = (E_y \Lambda_y^{-\frac{1}{2}})^T G(y)$
Robust kernel mean	$\mu_X^r = \frac{1}{n}\alpha 1_n$	$\mu_Y^r = \frac{1}{n}\beta 1_n$

TABLE I
NOTATIONS USED IN THE PAPER

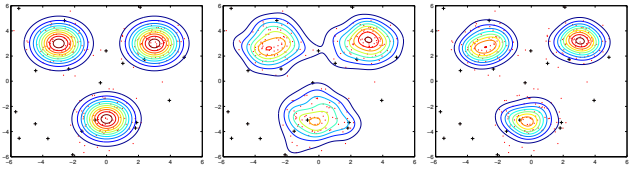

Fig. 2. Density estimation comparison. [Left]: data points in red with outliers in black. The contours depict the true density without the outliers. [Middle]: Density estimation using Equation (3). [Right]: Density estimation using Equation (8). The effect of outliers is less pronounced in the robust density estimation using eigenvalue decomposition.

The superscript $\cdot^r$ indicates that the mean is defined in the reduced RKHS, where only the leading q eigenvectors are retained.

The probability at a test point x can be computed by taking the inner product between the robust kernel mean and the mapped point:

$$p(x) = \langle \alpha(x), \mu_X^r \rangle = K(x)^T E_x E_x^T 1_n. \tag{8}$$

Compare Equation (8) with (3). Equation 3 computes the probability in the RKHS, while Equation 8 computes it in reduced RKHS. If all eigenvectors are retained in the eigenvalue decomposition, then $E_x E_x^T$ is equal to the identity matrix and Equation (8)) reduces to kernel density estimation Equation (3). Keeping only the significant eigenvectors has a smoothing effect on the density estimaion. This is also shown in Figure 2 using a toy example. 150 points are generated from a two dimensional multimodal distribution. Outliers are added from uniform distribution. The points are shown in red, while the outliers in black. Figure 2a shows the true density without the outliers. Figure 2b and c show the density estimation using Equation (3) and Equation (8). Density estimation using robust kernel mean ignores the effect of outliers by keeping only the significant eigenvectors and therefore follows the true density much more closely.

A. Robust Conditional Distribution Embedding

In this section, we derive a robust measure of conditional distribution $p(y|x)$. Following the discussion of the previous section, the robust conditional distribution $p(y|x)$ can be computed by taking the inner product of the embedded point, $\beta(y)$ with the robust conditional kernel mean, $\mu_{Y|x}^r$:

$$p(y|x) = \left\langle \beta(y), \mu_{Y|x}^r \right\rangle.$$

We will define the robust conditional kernel mean, $\mu_{Y|x}^r$, using the approach used in [**?**], where it is defined in terms of cross-covariance operators: $\mu_{Y|x} = C_{YX} C_{XX}^{-1} \phi(x)$. The covariance operators are estimated using n i.i.d. samples: $C_{XX} = \frac{1}{n}\Phi\Phi^T$, and $C_{YX} = \frac{1}{n}\Psi\Phi^T$.

However, unlike [**?**], we will compute the expectations in the reduced RKHS. This gives us the following definition of robust conditional kernel mean:

$$\mu_{Y|x}^r = {C_{YX}^r}^{-1} C_{XX}^r \alpha(x),$$

where the covariance operators are estimated using n i.i.d. samples: $C_{XX}^r = \frac{1}{n}\alpha\alpha^T$ and $C_{YX}^r = \frac{1}{n}\beta\alpha^T$.

Proposition 1. The robust kernel conditional mean is estimated by

$$\mu_{Y|x}^r = \beta K^{-1} K(x).$$

Proof: Let $h = {C_{XX}^r}^{-1}\alpha(x)$, then $C_{XX}^r h = \alpha(x)$ and $n^{-1}\alpha\alpha^T h = \alpha(x)$. Multiplying both sides by α^T, we get $n^{-1}K\alpha^T h = \alpha^T\alpha(x) \rightarrow \alpha^T h = nK^{-1}\alpha^T\alpha(x)$. Therefore $\mu_{Y|x}^r = \beta K^{-1}\alpha^T\alpha(x) = \beta K^{-1}K(x)$.

The conditional distribution $p(y|x)$ is then computed by taking the inner product between the mapped point $\beta(y)$ and the robust conditional kernel mean $\mu_{Y|x}^r$:

$$p(y|x) = \beta(y)^T \beta K^{-1} K(x) = G(y)^T E_y E_y^T K^{-1} K(x). \tag{9}$$

Consider the toy example of Figure 2 again, where 150 points are sampled from a mixture of three Gaussian distributions. Assume that each data point is also accompanied by the cluster center, $\{x_i, y_i\}_{i=1}^{150}$, where x_i is the cluster center and the y_i is the data point. Figure 3 shows the conditional probability distribution $p(y|x)$, when x is the top right cluster. The first figure uses the conditional kernel mean formulation [**?**], while the second figure uses the formulation derived in this section (Equation 9). The robust conditional embedding has a smoothing effect on the density estimation. When $E_y E_y^T = I$, Equation (9) reduces to the conditional kernel

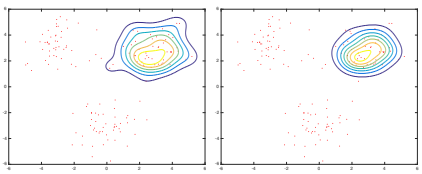

Fig. 3. Conditional density estimation comparison. 150 data points are sampled from a mixture of three Gaussian distributions [Left]: Conditional density estimation using conditional kernel mean [**?**] [Right]: Robust conditional density estimation using Equation (9). Density estimation is much smoother and closer to the true density.

mean expression. In Section IV, the robust conditional kernel mean embedding is applied to heterogeneous face recognition.

B. Robust Kernel Bayes' Rule

This section derives the expression for robust kernel Bayes' Rule, which will be described in terms of robust kernel means. The goal of the Bayesian inference is to find the posterior probability according to the Bayes' theorem:

$$q(y|x) = \frac{p(x|y)\pi(y)}{q(x)}$$

where $q(y|x)$ is the posterior probability of y given x, $p(x|y)$ and $\pi(y)$ are the likelihood and prior distribution. We will follow the derivation of [**?**] and refer the reader to the said paper and the references therein. The novelty of this section is to derive an expression for Bayes' rule in reduced RKHS instead of RKHS. We call the derived expression robust kernel Bayes's rule.

Let $\{u_i\}_{i=1}^{l}$ be samples from the prior distribution and μ_π^r and C_π^r be the estimation of prior robust kernel mean and kernel covariance. The estimated robust kernel mean of the posterior is given by

$$\mu_{Y|x}^r = MW^{-1}\alpha(x),$$

where $M = (C_{XY}^r C_{YY}^{r\,-1} C_\pi)^T$ and $W = C_{(XX)Y}^r C_{YY}^{r\,-1}\mu_\pi$.

Proposition 2: The covariance matrices M and W are estimated by

$$W = \alpha D \alpha^T, \text{ where } D = diag(G^{-1}G_{yu}1_l),$$
$$M = \Psi^\pi A^T \alpha^T, \text{ where } A = G^{-1}G_{yu}diag(1_l),$$

where G_{yu} is a matrix of dimension $n \times l$ containing $G_{yu_{ij}} = \boldsymbol{k}(y_i, u_j)$.

Proposition 3: The robust posterior kernel mean is estimated by

$$\mu_{Y|x}^r = \beta_\pi A^T (DK)^{-1} K(x),$$

where β_π is the robust kernel embedding matrix of the prior data points. We omit the proof of proposition 2 and 3 as they are similar to the proof of proposition 1. The posterior distribution $q(y|x)$ can be computed by taking the inner product between the mapped point $\beta(y)$ and the robust posterior kernel mean $\mu_{Y|x}^r$ as

$$q(y|x) = G_{yu}A^T(DK)^{-1}K(x). \tag{10}$$

C. Implementation

Equation (9) and (10), which are required to compute robust conditional probability and posterior probabilities respectively, involve matrix inversion. To ensure that the problem remains well posed, we use Tikhonov regularization. The resultant equations are

$$p(y|x) = G(y)^T E_y E_y^T (K + \epsilon I)^{-1} K(x), \tag{11}$$
$$q(y|x) = G_{yu}A^T((KD)^2 + \epsilon I)^{-1} KDK(x), \tag{12}$$

where I is the identity matrix and ϵ is the coefficient of Tikhonov regularization.

IV. Experiments

A. Heterogeneous Face Recognition

In the first example, we use the robust conditional embedding of distributions (Section III-A) for heterogeneous face recognition. In heterogeneous face recognition, face images from different modalities are matched. We consider face photo retrieval using sketch image. Given a set of n sketch-face pair $\{x, y\}_{i=1}^n$, we want to retrieve the most likely photo from the database, given a sketch image. We use CUHK student data base [**?**] for this purpose. Some sample face-sketch pairs are shown in Figure 4. The dataset consists of 188 face photos. For each face, there is a sketch drawn by an artist based on a photo taken in a frontal pose, under normal lighting condition, and with a neutral expression. In the dataset, each face photo and sketch is manually annotated with 35 landmark points such as eyes, nose etc. We use these fiducial points to align the face photos and sketches to mean photo and sketch using affine transformation. The images are cropped to a height of 250 pixels and width of 200 pixels. The photo images are converted to grayscale. We perform 5-fold cross validation on the dataset, training on four folds and testing on the remaining single fold. The experiment is repeated 10 times. The training data is used to compute the robust conditional kernel mean map (Equation (III-A)). Given a test sketch image x, the robust conditional probability (Equation (9)) is computed for each face image in the test set. The face with the highest probability is considered to be the corresponding face for the test sketch:

$$y = \underset{y}{\text{argmax}}\ p(y|x).$$

The σ for the Gaussian kernel and the number of eigenvectors retained are found by another level of 5 fold cross validation. We compare the method against conditional kernel density estimation and RKHS embedding [**?**]. The performance is evaluated using flat hit@(1-3) accuracy[1]. The results are shown in Table II. Robust RKHS embedding using eigenvalue decomposition greatly improves the recognition accuracy.

[1]Image is correctly classified if it is within top (1-3) predicted labels.

Method	1	2	3
Condition KDE	$.25 \pm .03$	$.38 \pm .02$	$.47 \pm .02$
RKHS embedding [**?**]	$.71 \pm .01$	$.80 \pm .02$	$.85 \pm .02$
Our Method	$\mathbf{.81} \pm .03$	$\mathbf{.86} \pm .02$	$\mathbf{.90} \pm .01$

TABLE II
HETEROGENEOUS FACE RECOGNITION ACCURACY ON CUHK STUDENT DATABASE [**?**]. ROBUST RKHS EMBEDDING USING EIGENVALUE DECOMPOSITION IMPROVES THE RECOGNITION ACCURACY.

Fig. 4. Sample face-sketch pairs

B. Object Recognition by Zero-Shot Learning

Object recognition by zero-shot learning aims to recognize objects for which we have no training examples (unseen object classes), only the description of the object classes is available. This is achieved by first learning to recognize objects in seen classes and then transferring knowledge from seen to unseen object classes. Over the years many zero-shot object recognition algorithms have been developed such as [**?**], [**?**], [**?**]. In this section, zero-shot learning is framed as a Bayesian inference problem and solved using robust kernel Bayes' theorem derived in the previous section.

We use Animal With Attributes (AwA) dataset [**?**]. AwA provides 50 classes of animals with 30,475 images altogether. Each class is associated with 85 attributes (see Table III and Figure 5). The dataset also provides the training/testing split with 10 classes held out for testing with 6180 images. Let $\{(x_i, y_i)\}_{i=1}^{n}$ be a set of all (animal, attribute) pair, including both training and testing images. The objective is to compute the probability $q(y|x)$ for unseen/testing images. At the time of training we have access to the attributes (y_i) of the test images but not the images itself. Using Bayes' theorem, the probability is given by $q(y|x) = p(x|y)\pi(y)/q(x)$. The restriction is that the likelihood distribution ($p(x|y)$) can only be learned using the training sample pairs. The prior distribution $\pi(y)$ is modeled on the class attributes of the unseen classes as they are available at the time of training. We used convolutional neural network (CNN) OverFeat[2] image features obtained from [**?**]. The overfeat feature for each image is a 4096 dimensional vector. Linear kernel is used to compute the embedding. Number of eigenvector retained is 100. We got 72.1% accuracy in recognizing the test class images (6180 images). The accuracy of other methods on the same dataset is listed in Table IV. All methods use CNN features. Our method performs considerably better than other methods. Figure 6 shows the class-wise accuracy of the proposed method and ESZSL [**?**]. Our proposed method outperforms ESZSL in 7 out of 10 classes. Figure 7 shows the effect of choosing the number of eigenvectors on the accuracy of the algorithm. The figure indicates that 4096 dimensional overfeat feature vectors contain redundant information, which can be effectively reduced to about 50 dimensions using our algorithm without any loss of accuracy.

black	blue	brown	patches	spots	stripes	furry
big	flippers	hands	hooves	paws	longneck	tail
horns	tusks	smelly	hops	swims	walks	fast
hairless	claws	hibernate	bush	tree	domestic	water

TABLE III
ANIMAL WITH ATTRIBUTES. EACH CLASS OF ANIMALS IS ASSOCIATED WITH 85 ATTRIBUTES. SOME OF THE ATTRIBUTES ARE LISTED ABOVE.

Fig. 5. Class-attribute matrix. White color indicates presence of attribute while black indicates absence.

Method	Features	Accuracy
IAP[**?**]	DECAF	44.5
DAP[**?**]	DECAF	53.2
DAP[**?**]	VGG19	60.8
AMP [**?**]	OverFeat	66.0
ESZSL [**?**]	OverFeat	66.7
Our Method	OverFeat	**72.1**

TABLE IV
RESULT OF CLASS RECOGNITION ACCURACY (%) ON AWA. RESULTS ARE COMPARED AGAINST METHODS THAT USE CNN FEATURES. DECAF: HTTP://CAFFE.BERKELEYVISION.ORG/MODEL_ZOO.HTML VGG19:HTTP://WWW.ROBOTS.OX.AC.UK/~VGG/RESEARCH/

V. CONCLUSION

We have proposed a non-parametric method to robustly estimate the conditional and posterior distributions (Bayes' rule). Robust estimation is based on the subspace embedding of kernel mean using eigenvalue decomposition of the kernel matrix. The method is applied to Heterogenous face recognition and zero-shot object recognition problems. In both problems, we show better accuracy than the compared methods.

[2]http://cilvr.nyu.edu/doku.php?id=code:start

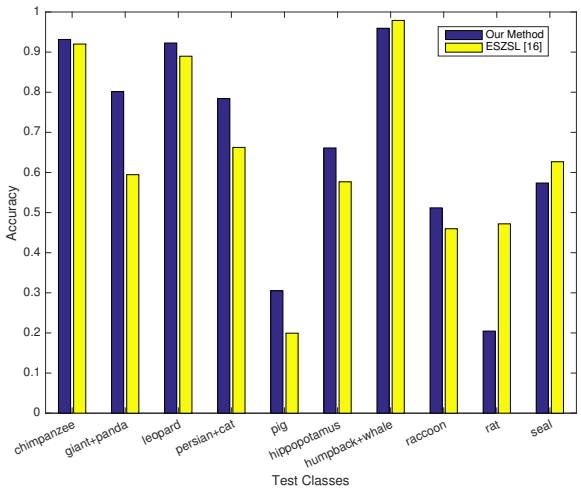

Fig. 6. Class-wise accuracy of the proposed method on 10 test classes.

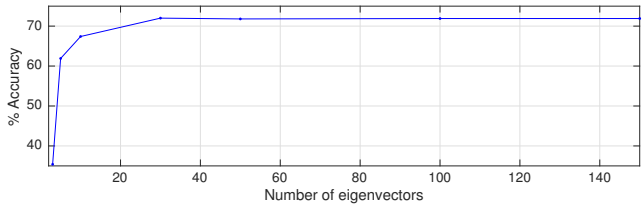

Fig. 7. Effect of choosing the number of leading eigenvectors on the recognition accuracy of AwA dataset.

REFERENCES

[1] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf, "A hilbert space embedding for distributions," in *Algorithmic Learning Theory*. Springer, 2007, pp. 13–31.

[2] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[3] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf, "Learning from distributions via support measure machines," in *Advances in neural information processing systems*, 2012, pp. 10–18.

[4] Yuya Yoshikawa, Tomoharu Iwata, and Hiroshi Sawada, "Latent support measure machines for bag-of-words data classification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1961–1969.

[5] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 961–968.

[6] Kenji Fukumizu, Le Song, and Arthur Gretton, "Kernel bayes' rule.," in *NIPS*, 2011, pp. 1737–1745.

[7] Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey J Gordon, and Alex J Smola, "Hilbert space embeddings of hidden markov models," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 991–998.

[8] Omar Arif and Patricio A Vela, *Robust Density Comparison Using Eigenvalue Decomposition*, INTECH Open Access Publisher, 2012.

[9] Omar Arif and Patricio A Vela, "Robust density comparison for visual tracking.," in *BMVC*, 2009, pp. 1–10.

[10] Mark Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, no. 3, pp. 669–688, 2002.

[11] Le Song and Bo Dai, "Robust low rank kernel embeddings of multivariate distributions," in *Advances in Neural Information Processing Systems*, 2013, pp. 3228–3236.

[12] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[13] Kenji Fukumizu, Francis R Bach, and Michael I Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces," *The Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.

[14] Xiaogang Wang and Xiaoou Tang, "Face photo-sketch synthesis and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 11, pp. 1955–1967, 2009.

[15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 951–958.

[16] Bernardino Romera-Paredes and PHS Torr, "An embarrassingly simple approach to zero-shot learning," in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 2152–2161.

[17] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong, "Zero-shot object recognition by semantic manifold distance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2635–2644.

[18] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Attribute-based classification for zero-shot visual object categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 453–465, 2014.