Robust Fitting of Multiple Structures: The Statistical Learning Approach

Tat-Jun Chin, Hanzi Wang and David Suter
School of Computer Science
The University of Adelaide, South Australia
{tjchin,hwang,dsuter}@cs.adelaide.edu.au

Abstract

We propose an unconventional but highly effective approach to robust fitting of multiple structures by using statistical learning concepts. We design a novel Mercer kernel for the robust estimation problem which elicits the potential of two points to have emerged from the same underlying structure. The Mercer kernel permits the application of well-grounded statistical learning methods, among which nonlinear dimensionality reduction, principal component analysis and spectral clustering are applied for robust fitting. Our method can remove gross outliers and in parallel discover the multiple structures present. It functions well under severe outliers (more than 90% *of the data) and considerable inlier noise without requiring elaborate manual tuning or unrealistic prior information. Experiments on synthetic and real problems illustrate the superiority of the proposed idea over previous methods.*

1. Introduction

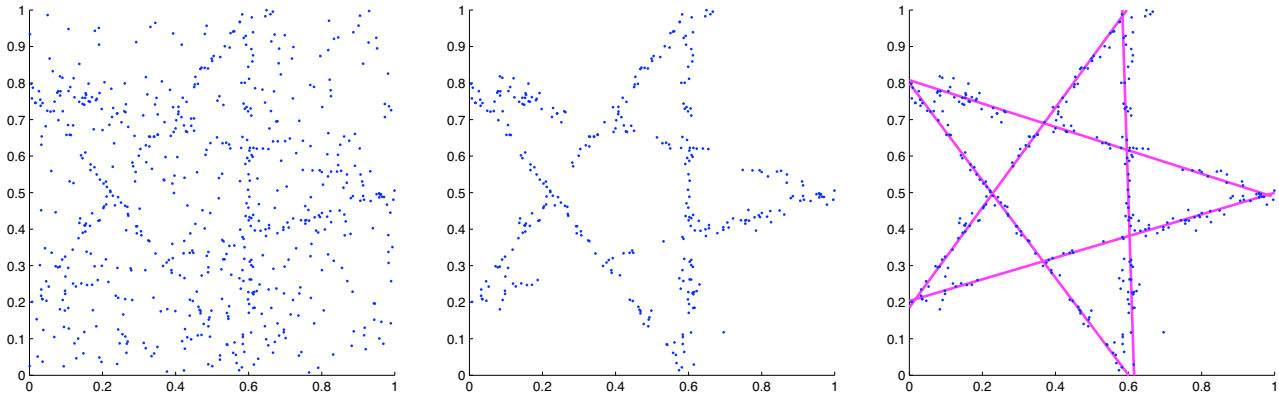
Outliers in data almost unavoidably arise in practical computer vision problems due to the imperfect processes in the feature extraction pipeline. To mitigate the debilitating influence of severe outliers on model fitting, robust statistical approaches have been applied extensively in computer vision. While many robust statistical approaches such as LMedS and M-estimators originated in the statistics community [7], the widespread usage of robust statistics in vision also motivated the invention of other methods such as RANSAC [4] and the Hough Transform [3].

In the context of practical vision applications, a robust fitting method should possess several desirable characteristics. Since outlier rates of more than 50% are very prevalent in vision, a method must be capable of tolerating a large number of outliers to ensure basic applicability. A competent method should also be able to handle significant inlier variability, and if possible provide an accurate estimate of the scale of inlier noise. It is also very common for the data to contain *multiple* instances of a model where the points belonging to each structure act as pseudo-outliers to the others, thus the method must also unearth all of the structures present without a priori knowing how many exist.

Generally speaking robust fitting techniques have followed either one of the two following paradigms: (1) Generate putative model hypotheses based on random subsets of the input data and find the hypothesis which maximizes some fitting criterion, e.g. [7, 4, 6, 13, 1, 17]. To fit multiple structures one can apply a particular method sequentially by removing the inliers of a structure at each iteration. (2) Detect clusters directly in the parameter space of the model, where each cluster is indicative of an instance of the model in the data, e.g. [3, 18, 11]. For computational feasibility, these methods often sample the parameter space by generating random hypotheses from subsets of the data.

The two categories differ in how well they satisfy the properties desirable of robust fitting methods. Techniques in the first group are generally very robust towards outliers, where *empirical* breakdown points of more than 80% have been reported [17]. However they are generally suboptimal in discovering multiple structures, since disastrous outcomes can be obtained if the initial fits are not accurate and the wrong inliers are removed (or even if the initial fits *are* accurate but the estimated inlier scale is wrong). Secondly, devising a stopping criterion to reflect the true number of structures is non-trivial. On the other hand methods in the second group are not affected by the perils of sequential fitting. However, besides suffering from poor computational efficiency and a generally lower tolerance to gross outliers, it is not easy to deduce the number of true clusters.

In this paper we propose a novel solution to robust statistics by using statistical learning concepts. Our method is fundamentally different since it does not follow either of the categories above. Instead of sampling and scoring random hypotheses or clustering in the parameter space, we examine relations between data points. Central to our approach is to craft a *Mercer kernel* between two points which elicits their potential of arising from a common structure. The Mercer kernel induces a Reproducing Kernel Hilbert Space (RKHS) which permits us to draw from the vast body of lit-



(a) Input data: 92% outliers, inlier $\sigma = 0.01$. (b) Gross outlier removal (c) Multiple structure discovery

Figure 1. Summary of the proposed method using *actual results* on 2D line fitting of 5 lines. (a) Input data of 650 points, with 50 inliers per line and 400 gross outliers (i.e. outlier rate of 92.31%). The inliers are perturbed with Gausian noise of $\sigma = 0.01$ which is high relative to the range of values of the data (i.e. $[0\ 1\ 0\ 1]$). (b) Step 1: Gross outlier removal. (c) Step 2: Multiple structure discovery.

erature on statistical learning theory [15, 9]. In the RKHS we perform dimensionality reduction, principal component analysis and spectral clustering *on the data points* for robust fitting. The proposed method can effectively remove gross outliers in the data and in parallel discover the multiple structures present. It exhibits considerable tolerance to inlier noise and high resistance to severe outliers encompassing more than 90% of the data. It also does not require excessive manual tuning or unrealistic prior information of the data typical of previous methods like RANSAC [4] (manually set inlier threshold) or multiRANSAC [21] (requires prior knowledge of the number of structures). Fig. 1 summarizes the proposed method.

We emphasize that the Mercer kernel is primarily used in statistical learning, and we do not perform kernel-based mean-shift clustering [2, 11] or density estimation [17, 16].

Our method follows recent developments in robust statistics. Zhang and Kosecká [20, 19] advanced a different view of the problem of analyzing the distribution of the residuals of random hypotheses to a point. In such an arrangement, the multiple structures are revealed as multiple modes in the distribution, and it is proposed [20] that these can be discovered via nonparametric mode seeking. Unfortunately severe outliers and incorrect bandwidth estimates for density estimation can easily produce many false peaks and valleys which obscure the genuine modes. In a later work [19] simple statistics like skewness and kurtosis of the distribution are used to separate inliers and outliers, but this is confined to data with a single structure only. Building upon [20], Toldo and Fusiello [12] proposed a "conceptual representation" for robust fitting, essentially a reduction of the parameter space to a one-dimensional discrete space of hypothesis indices. Robust fitting proceeds by agglomerative clustering of the conceptual representation of the data points. This however has serious drawbacks. Firstly to build the representation a manually determined inlier threshold must be supplied, and secondly their agglomerative clustering method requires a pre-defined cut-off threshold related to the prior knowledge of how many points each underlying model instance possesses. Their approach is thus mired in a RANSAC-like dependence on manual parameter input.

Our major contribution is a novel Mercer kernel for the robust estimation problem. In Sec. 2, we describe the Mercer kernel and show how it can be used in conjunction with statistical learning concepts for effective gross outlier removal. Sec. 3 explains how, based on the Mercer kernel, nonlinear principal component analysis and spectral clustering are performed on the data for multiple structure discovery. Sec. 4 presents results on synthetic and real data, and in Sec. 5 we draw conclusions and state future work.

2. Gross Outlier Removal

This section describes how gross outliers can be effectively removed with kernel methods. Let the model to be fitted be determined by p parameters. Given input data $\{x_i\}_{i=1,\dots,N}$ of N points our approach begins by randomly sampling a set of M model hypotheses $\{\theta_j\}_{j=1,\dots,M}$, where each hypothesis θ_j is fitted from a minimal subset of p points. Various sampling strategies [13, 20, 12] can be applied to ensure that at least K hypotheses, where $K < M$, are generated from pure inliers only. We emphasize that we do not score and rank the random hypotheses.

2.1. A Mercer Kernel for Robust Fitting

For each data point x_i compute its absolute residual set $\mathbf{r}_i = \{r_1^i, \dots, r_M^i\}$ as measured to the M hypotheses. We sort the elements in $\mathbf{r}_i$ to obtain the sorted residual set $\tilde{\mathbf{r}}_i = \{r_{\lambda_1^i}^i, \dots, r_{\lambda_M^i}^i\}$, where the permutation $\{\lambda_1^i, \dots, \lambda_M^i\}$ is obtained such that $r_{\lambda_1^i}^i \leq \cdots \leq r_{\lambda_M^i}^i$. Define the sorted

hypothesis set of point x_i as

$$\tilde{\boldsymbol{\theta}}_i := \{\lambda_1^i, \ldots, \lambda_M^i\}, \tag{1}$$

i.e. $\tilde{\boldsymbol{\theta}}_i$ depicts the order in which x_i becomes the inlier of the M hypotheses as a fictitious inlier threshold is increased from 0 to ∞. We define the Ordered Residual Kernel (ORK) between two data points as

$$k_{\tilde{r}}(x_{i_1}, x_{i_2}) := \frac{1}{Z} \sum_{t=1}^{M/h} z_t \cdot k_{\cap}^t(\tilde{\boldsymbol{\theta}}_{i_1}, \tilde{\boldsymbol{\theta}}_{i_2}), \tag{2}$$

where $z_t = \frac{1}{t}$ are the harmonic series and $Z = \sum_{t=1}^{M/h} z_t$ is the (M/h)-th harmonic number. Without lost of generality assume that M is wholly divisible by h. Step size h is used to obtain the Difference of Intersection Kernel (DOIK)

$$k_{\cap}^t(\tilde{\boldsymbol{\theta}}_{i_1}, \tilde{\boldsymbol{\theta}}_{i_2}) := \frac{1}{h}(|\tilde{\boldsymbol{\theta}}_{i_1}^{1:\alpha_t} \cap \tilde{\boldsymbol{\theta}}_{i_2}^{1:\alpha_t}| - |\tilde{\boldsymbol{\theta}}_{i_1}^{1:\alpha_{t-1}} \cap \tilde{\boldsymbol{\theta}}_{i_2}^{1:\alpha_{t-1}}|) \tag{3}$$

where $\alpha_t = th$ and $\alpha_{t-1} = (t-1)h$. Symbol $\tilde{\boldsymbol{\theta}}_i^{a:b}$ indicates the set formed by the a-th to the b-th elements of $\tilde{\boldsymbol{\theta}}_i$. Since the contents of the sorted hypotheses set are merely permutations of $\{1 \ldots M\}$, i.e. there are no repeating elements,

$$0 \leq k_{\tilde{r}}(x_{i_1}, x_{i_2}) \leq 1. \tag{4}$$

Note that $k_{\tilde{r}}$ is independent of the type of model to be fitted.

Let τ be a *fictitious* inlier threshold. The kernel $k_{\tilde{r}}$ captures the intuition that, if τ is low, two points arising from the same structure will have high normalized intersection since they share many common hypotheses. If τ is high, implausible hypotheses fitted on outliers start to dominate and decrease the normalized intersection. Step size h allows us to quantify the rate of change of intersection if τ is increased from 0 to ∞, and since z_t is decreasing, $k_{\tilde{r}}$ will evaluate to a high value for two points from the same structure. In contrast, $k_{\tilde{r}}$ is always low for points not from the same structure or that are outliers. Fig. 2 demonstrates this effect. Note that τ is merely an abstract construction— $k_{\tilde{r}}$ **does not require a user input inlier threshold**. Also parameter h depends on M, a value determined based on the size of the minimal subset and the number of data [20, 12] and **is not contingent on knowledge of the true inlier noise scale** σ. Fig. 2 depicts the independence of h with respect to σ. This is further substantiated by experiments in Sec. 4.

Proof of satisfying Mercer's condition. Let D be a fixed domain, and $\mathcal{P}(D)$ be the power set of D, i.e. the set of all subsets of D. Let $S \subseteq \mathcal{P}(D)$, and $p, q \in S$. If μ is a measure on the domain D, then

$$k_{\cap}(p, q) = \mu(p \cap q), \tag{5}$$

called the intersection kernel, is provably a valid Mercer kernel [9]. The DOIK can be rewritten as

$$k_{\cap}^t(\tilde{\boldsymbol{\theta}}_{i_1}, \tilde{\boldsymbol{\theta}}_{i_2}) = \frac{1}{h}(|\tilde{\boldsymbol{\theta}}_{i_1}^{(\alpha_{t-1}+1):\alpha_t} \cap \tilde{\boldsymbol{\theta}}_{i_2}^{(\alpha_{t-1}+1):\alpha_t}|$$

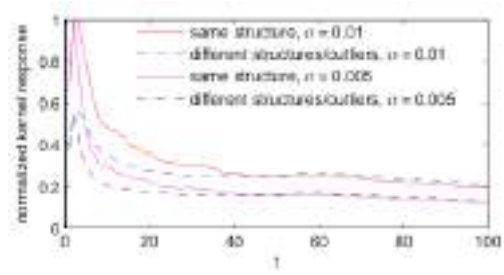


Figure 2. Normalized value of each DOIK component for $k_{\tilde{r}}$ evaluated between two points from the same structure and two points not from the same structure or that are gross outliers. The result here is averaged from the 650 points in Fig. 1(a) with inlier noise $\sigma = 0.01$ and 0.005. M and h are respectively fixed at 5000 and 50. The kernel evaluates to high and low values accordingly without having to tune h with respect to the inlier noise scale.

$$+|\tilde{\boldsymbol{\theta}}_{i_1}^{1:(\alpha_{t-1})} \cap \tilde{\boldsymbol{\theta}}_{i_2}^{(\alpha_{t-1}+1):\alpha_t}| + |\tilde{\boldsymbol{\theta}}_{i_1}^{(\alpha_{t-1}+1):\alpha_t} \cap \tilde{\boldsymbol{\theta}}_{i_2}^{1:(\alpha_{t-1})}|). \tag{6}$$

If we let $D = \{1 \ldots M\}$ be the set of all possible hypothesis indices and μ be uniform on D, each term in Eq. (6) is simply an intersection kernel multiplied by $|D|/h$. Since multiplying a kernel with a positive constant and adding two kernels respectively produce valid Mercer kernels [9], the DOIK and ORK are also valid Mercer kernels.•

A Mercer kernel $k(\cdot, \cdot)$ induces a mapping ϕ from the input space X to a possibly infinite dimensional feature space

$$\phi : x \in X \longmapsto \phi(x) = k(x, \cdot) \in F_k, \tag{7}$$

where $\phi(x)$ belongs to a function space F_k that has the structure of a so-called Reproducing Kernel Hilbert Space (RKHS) [9]. The RKHS is endowed with an inner product, and Mercer's theorem implies that

$$\langle \phi(x_{i_1}), \phi(x_{i_2}) \rangle = k(x_{i_1}, x_{i_2}). \tag{8}$$

As a valid Mercer kernel, the ORK also induces a RKHS, and with $k_{\tilde{r}}$ we are able to perform dot products in $F_{k_{\tilde{r}}}$ without explicitly characterizing or evaluating ϕ.

Encapsulating a robust fitting solution in a Mercer kernel also allows us to apply model- or domain-specific information in a theoretically consistent manner by manipulating the kernel function to produce a new kernel function, e.g.

$$k_{new}(\cdot, \cdot) = \beta_1 k_{\tilde{r}}(\cdot, \cdot) + \beta_2 k_2(\cdot, \cdot) + \beta_3 k_3(\cdot, \cdot) + \ldots \tag{9}$$

where β_1, β_2, β_3, $\ldots$ are positive constants and $k_2(\cdot, \cdot)$, $k_3(\cdot, \cdot)$, $\ldots$ are Mercer kernels pertaining to other information. For example, in line or plane fitting we can exploit the Gaussian kernel [9]

$$k(x_{i_1}, x_{i_2}) = \exp(-\|x_{i_1} - x_{i_2}\|^2 / 2\sigma^2) \tag{10}$$

to enforce the knowledge that two points arising from the same line/plane should be relatively close in space.

2.2. Kernel SVD Gross Outlier Removal

Denoting by $\mathbf{A} = [\phi(x_1) \dots \phi(x_N)]$ the matrix of the input data *after* it is mapped to RKHS $F_{k_{\tilde{r}}}$, the *kernel matrix* $\mathbf{K} = \mathbf{A}^T\mathbf{A}$ is computed using the kernel function $k_{\tilde{r}}$ as

$$\mathbf{K}_{p,q} = \langle \phi(x_p), \phi(x_q) \rangle = k_{\tilde{r}}(x_p, x_q), \quad p, q \in \{1 \dots N\}. \tag{11}$$

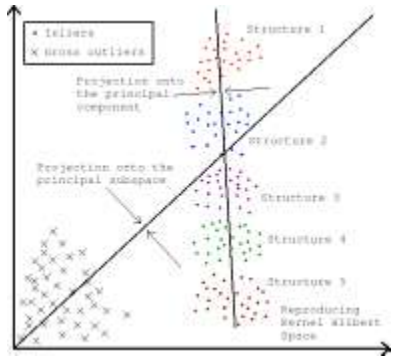
Since $k_{\tilde{r}}$ is a valid Mercer kernel, $\mathbf{K}$ is guaranteed to be positive semi-definite [9]. Let $\mathbf{K} = \mathbf{Q}\mathbf{\Delta}\mathbf{Q}^T$ be the eigenvalue decomposition (EVD) of $\mathbf{K}$. Then the rank-n Kernel Singular Value Decomposition (SVD) [9] of $\mathbf{A}$ is

$$\mathbf{A}^n = [\mathbf{A}\mathbf{Q}^n(\mathbf{\Delta}^n)^{-\frac{1}{2}}][(\mathbf{\Delta}^n)^{\frac{1}{2}}][(\mathbf{Q}^n)^T] \equiv \mathbf{U}^n\mathbf{\Sigma}^n(\mathbf{V}^n)^T. \tag{12}$$

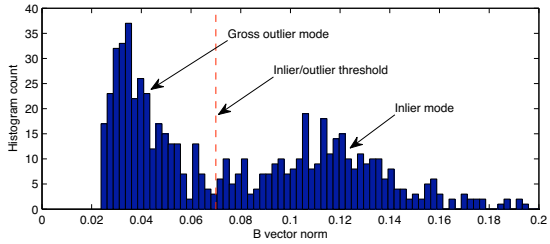
Via the Matlab notation, $\mathbf{Q}^n = \mathbf{Q}_{:,1:n}$ and $\mathbf{\Delta}^n = \mathbf{\Delta}_{1:n,1:n}$. The left singular vectors $\mathbf{U}^n$ is an orthonormal basis for the n-dimensional principal subspace of $\mathbf{A}$ in $F_{k_{\tilde{r}}}$. Projecting the data onto the principal subspace yields

$$\mathbf{B} = [\mathbf{A}\mathbf{Q}^n(\mathbf{\Delta}^n)^{-\frac{1}{2}}]^T\mathbf{A} = (\mathbf{\Delta}^n)^{\frac{1}{2}}(\mathbf{Q}^n)^T, \tag{13}$$

where $\mathbf{B} = [b_1 \dots b_N] \in \mathbb{R}^{n \times N}$ is the reduced dimension version of $\mathbf{A}$. Directions of the principal subspace are dominated by inlier points, since $k_{\tilde{r}}$ evaluates to a high value generally for them, but always to a low value for gross outliers. Thus the vectors in $\mathbf{B}$ have high norms if they correspond to inlier points and vice versa. Fig. 3(a) illustrates.



(a) Input data mapped to the RKHS $F_{k_{\tilde{r}}}$



(b) Histogram of vector norms in the kernel principal subspace

Figure 3. (a) Gross outlier removal with Kernel SVD and structure discovery with Kernel PCA. (b) The histogram is obtained from the actual input data in Fig. 1(a).

This observation is exploited for gross outlier removal. Fig. 3(b) shows the actual histogram of $\mathbf{B}$ vector norms of the data in Fig. 1(a) for $n = 6$, a value allowing $span(\mathbf{U}^n)$ to encompass 90% of the singular values in $\mathbf{\Sigma}^n$. The existence of two distinct modes, corresponding respectively to inliers and gross outliers, is evident. We can thus safely discard data with low norms as gross outliers. The cut-off threshold ψ can be set by analyzing the distribution of the norms. For instance we can fit a 1D Gaussian Mixture Model (GMM) with two components

$$f(b) = \sum_{c=1,2} \pi_c \mathcal{N}(b|\mu_c, \sigma_c) \tag{14}$$

on the $\mathbf{B}$ vector norms, where $\mathcal{N}$ is a Gaussian with mean μ_c and standard deviation σ_c, and π_c is the mixing coefficient. The threshold can be obtained as the point of equal Mahalanobis distance as in Fig. 3(b), i.e.

$$\sigma_2(\psi - \mu_1)^2 = \sigma_1(\psi - \mu_2)^2, \tag{15}$$

or as the average between the two means, i.e. $\psi = 0.5(\mu_1 + \mu_2)$. A threshold which is less dependent on the shape of the distribution is the following

$$\psi = \rho \max_{i=1,\dots,N}(\|b_i\|^2), \tag{16}$$

where $\rho = 0.3$ is empirically justified to be effective. Eq. (16) is suitable for both clean and noisy data, i.e. there exists either one or two modes in the $\mathbf{B}$ vector norm distribution. Fig. 1(b) shows an actual result of the method.

Our outlier removal scheme is considerably more tractable than the mode seeking-based method of [20]. There, the number of modes in the residual distribution equals the unknown number of structures, thus the problem is non-trivial (see Sec. 1). Contrast this to our case where it is known beforehand that there are at most two modes in the norm distribution, thus the problem is greatly simplified.

Our subspace operation also vastly differs from the pbM-estimator's [1], where putative subspaces in the *input space* are generated, each equivalent to a model hypothesis. The pbM method then seeks the subspace (equivalently, model hypothesis) which maximizes the mode of the projection [1]. Being a method in Group 1 (see Sec. 1), pbM faces difficulty in determining the number of structures. In contrast, our method performs subspace projection in the *RKHS deterministically* and can automatically deduce the number of structures of generic models (Sec. 3 elaborates).

3. Discovering Multiple Structures

We fit multiple model instances based on the idea that points from the same structure concentrate at a location in RKHS $F_{k_{\tilde{r}}}$; see Fig. 3(a). This is because the kernel $k_{\tilde{r}}$ (which is equivalent to a dot product in $F_{k_{\tilde{r}}}$) evaluates to a high value for points from the same structure and vice versa, and our task is to cluster the *data* in $F_{k_{\tilde{r}}}$. This differs from the Hough Transform [3] or mean shift-based methods [11] which cluster the *hypotheses* in the *parameter* space.

3.1. Kernel PCA and Spectral Clustering

Using Kernel PCA [9], we first seek a parsimonious representation of the data which maximizes its spread in $F_{k_{\tilde{r}}}$. Let $\{y_i\}_{i=1,\ldots,N'}$ be the N'-point subset of the input data that remains after outlier removal, where $N' < N$. Denote by $\mathbf{C} = [\phi(y_1) \ldots \phi(y_{N'})]$ the data matrix after mapping the data to $F_{k_{\tilde{r}}}$, and by symbol $\tilde{\mathbf{C}}$ the result of adjusting $\mathbf{C}$ with the empirical mean of $\{\phi(y_1), \ldots, \phi(y_{N'})\}$. The *centered* kernel matrix $\tilde{\mathbf{K}}' = \tilde{\mathbf{C}}^T\tilde{\mathbf{C}}$ can be obtained as

$$\tilde{\mathbf{K}}' = \boldsymbol{\nu}^T\mathbf{K}'\boldsymbol{\nu}, \quad \boldsymbol{\nu} = [\mathbf{I}_{N'} - \frac{1}{N'}\mathbf{1}_{N',N'}], \tag{17}$$

where $\mathbf{K}' = \mathbf{C}^T\mathbf{C}$ is the *uncentered* kernel matrix, $\mathbf{I}_s$ and $\mathbf{1}_{s,s}$ are respectively the $s \times s$ identity matrix and matrix of ones. If $\tilde{\mathbf{K}}' = \mathbf{R}\boldsymbol{\Omega}\mathbf{R}^T$ is the EVD of $\tilde{\mathbf{K}}'$, then we obtain first-m kernel principal components $\mathbf{P}^m$ of $\mathbf{C}$ as the first-m left singular vectors of $\tilde{\mathbf{C}}$ [9] , i.e.

$$\mathbf{P}^m = \tilde{\mathbf{C}}\mathbf{R}^m(\boldsymbol{\Omega}^m)^{-\frac{1}{2}}, \tag{18}$$

where $\mathbf{R}^m = \mathbf{R}_{:,1:m}$ and $\boldsymbol{\Omega}_{1:m,1:m}$; see Eq. (12). Projecting the data on the kernel principal components yields

$$\mathbf{D} = [d_1 \ldots d_{N'}] = (\boldsymbol{\Omega}^m)^{\frac{1}{2}}(\mathbf{R}^m)^T, \tag{19}$$

where $\mathbf{D} \in \mathbb{R}^{m \times N'}$. The *affine* subspace $span(\mathbf{P}^m)$ maximizes the spread of the *centered* data in the RKHS, as Fig. 3(a) illustrates, and the projection $\mathbf{D}$ offers an effective representation of the data for clustering.

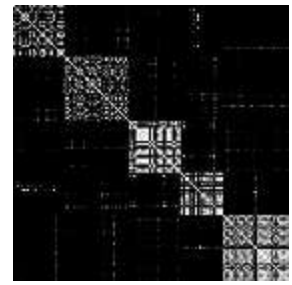
Various methods can be applied to cluster $\mathbf{D}$, and we achieve it using the Normalized Cut (Ncut) [10] method due to its effectiveness. A fully connected graph is first derived from the data, where its weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{N' \times N'}$ is obtained as

$$\mathbf{W}_{p,q} = \exp(-\|d_p - d_q\|^2/2\delta^2), \tag{20}$$

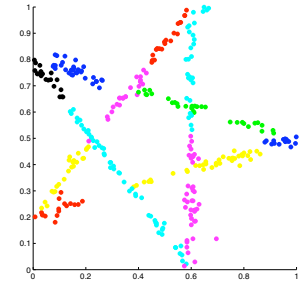
and δ is taken as the average nearest neighbour distance in the Euclidean sense among the vectors in $\mathbf{D}$. Fig. 4(a) shows $\mathbf{W}$ for the input data in Fig. 1(b) after gross outlier removal. It can be seen that strong affinity exists between points from the same structure. The degree $\mathbf{G}$ and Laplacian $\mathbf{L}$ matrices, both of size $N' \times N'$, are obtained as

$$\mathbf{G}_{p,p} = \sum_{q=1}^{N'} \mathbf{W}_{p,q} \quad \text{and} \quad \mathbf{L} = \mathbf{G} - \mathbf{W}, \tag{21}$$

where the off-diagonal elements of $\mathbf{G}$ are zero. Under Ncut, the number of clusters l embedded in the data is revealed as the number of eigenvalues of $\mathbf{L}$ which are zero [10]. Denoting by $\mathbf{E} = [e_1 \ldots e_l] \in \mathbb{R}^{N' \times l}$ the l eigenvectors of $\mathbf{L}$ with zero eigenvalues, a subsequent k-means step with $k = l$ is then performed on the rows of $\mathbf{E}$ to extract the clusters.



(a) Weighted adjacency matrix with points re-arranged based on true cluster membership.



(b) Ncut reveals 12 clusters. The figure is best viewed in colour. Note that the colours repeat.

Figure 4. (a) Weighted adjacency matrix for the data in Fig. 1(b). (b) Normalized Cut clustering results on for the data in Fig. 1(b).

In practice, due to the presence of noise and the limits of computational precision, it is unlikely that the eigenvalues are exactly zero. Finding a consistently accurate thresholding scheme is also non-trivial, if not impossible. Thus in our work we set a relatively high threshold of $1.0 \times e^{-3}$ to deliberately oversegment the data, as Fig. 4(b) shows. We then resolve the redundancies by merging the structures.

3.2. Structure Merging Scheme

Our structure merging scheme operates under the objective of fitting the data with the least number of structures possible. A model instance is first estimated from each point cluster with LMedS [7]. The algorithm then sequentially merges structures by testing, if a structure is merged with another structure, whether the data can still be "explained" satisfactorily by the remaining structures. This proceeds until the condition of satisfactory explanation is violated. Algorithm 1 lists the structure merging scheme.

Algorithm 1 Structure merging scheme after Ncut

input: Set of l° point clusters $\mathcal{C} = \{\mathcal{C}_l\}_{l=1,\ldots,l^\circ}$.
while $continue = true$ **do**
  $\forall\ \mathcal{C}_l$, estimate model $\mathcal{M}_l$ using LMedS.
  Get $r^i_{l,m}$ as residual of the i-th point in $\mathcal{C}_l$ to $\mathcal{M}_m$.
  Get φ_l as inlier threshold of $\mathcal{M}_l$ by Eq. (22).
  $continue = false$.
  for $l = 1, \ldots, |\mathcal{C}|$ **do**
    if $(\sum_{i,m \neq l} \delta(|r^i_{l,m}| <= \varphi_m)) \geq |\mathcal{C}_l|$ **then**
      For all m and i, if $|r^i_{l,m}| \leq \varphi_m$ move point i from $\mathcal{C}_l$ to $\mathcal{C}_m$ until $\mathcal{C}_l$ is empty.
      $\mathcal{C} \leftarrow \mathcal{C} - \mathcal{C}_l$, $continue = true$.
      break
    end if
  end for
end while
output: Model parameters for $|\mathcal{C}| \leq l^\circ$ structures.

The inlier threshold for each model $\mathcal{M}_l$ in Step 5 of the algorithm is calculated as

$$\varphi_l = 0.5(med(|r|_{inliers}) + med(|r|_{pseudo})), \quad (22)$$

$$\text{where} \quad |r|_{inliers} = \{|r^i_{l,l}|\}_{i=1,\ldots,|\mathcal{C}_l|} \quad (23)$$

$$\text{and} \quad |r|_{pseudo} = \cup_{\forall m \neq l}\{|r^i_{m,l}|\}_{i=1,\ldots,|\mathcal{C}_m|} \quad (24)$$

are respectively the set of absolute residuals of points in cluster $\mathcal{C}_l$ to model $\mathcal{M}_l$ and the set of absolute residuals of the pseudo-outliers of model $\mathcal{M}_l$. Function $\delta(\cdot)$ in Step 8 is the Kronecker delta. Fig. 1(c) shows the result of this algorithm on the clusters in Fig. 4(b). Note that the algorithm is applicable to generic model types, and that compared to other model selection based techniques, our task is considerably easier since the gross outliers have been removed.

4. Results

We evaluate the performance of the proposed method (henceforth known as *Kernel Fitting* or KF) in various applications with an emphasis on multiple structure discovery. The Mercer kernel is implemented efficiently with complexity $\mathcal{O}(M)$ using symbol tables [8] (less than 10 seconds in total for 500 data points and 5000 random hypotheses).

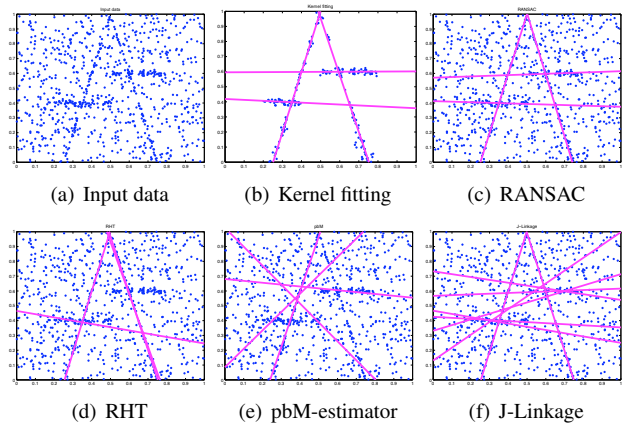
Multiple 2D line fitting. Eight methods are compared in this experiment. Table 1 depicts their dependence on manual parameter inputs. Sequential fitting methods require the true number of structures as a stopping criterion, while RHT and J-Linkage prune clusters based on the expected number of points per structure. Only RHA and KF derive the number of structures automatically from the data. A total of 5000 random hypotheses are generated and reused across all methods which require them. In RANSAC and J-Linkage, the required inlier threshold is set as twice the true inlier noise scale. In KF, we add to the ORK the Gaussian kernel by using the average nearest neighbour distance as its width, and h is fixed at 100. The codes of pbM and J-Linkage are obtained from the web[1] while we implemented the others.

	Methods							
Parameter	1	2	3	4	5	6	7	8
Inlier noise scale/threshold	*						*	
Number of structures	*	*	*	*				
No. of points per structure					*		*	

Table 1. Manual parameter inputs required for each method. 1-RANSAC [4], 2-LMedS [7], 3-ALKS [6], 4-pbM-estimator [1], 5-Randomized Hough Transform (RHT) [18], 6-Residual Histogram Analysis (RHA) [20], 7-J-Linkage [12] and 8-Kernel Fitting (KF).

The type of data used in this experiment is depicted in Fig. 5 along with a few sample results (more extensive results follow). The four lines in the data are arranged to produce a challenging configuration for line fitting. Each line contains 50 inliers contaminated with Gaussian noise of standard deviation σ. A total of L points of gross outliers are also randomly inserted while maintaining the range of the data in $[0\ 1\ 0\ 1]$. The total outlier rate is thus given by $100\% \times (L + 150)/(L + 200)$.

[1] Respectively from http://www.caip.rutgers.edu/riul/research/code.html and http://profs.sci.univr.it/~fusiello/demo/jlk/.



(a) Input data (b) Kernel fitting (c) RANSAC

(d) RHT (e) pbM-estimator (f) J-Linkage

Figure 5. Input data and sample results for the multiple 2D line fitting experiment. In this particular example there are 50 points per line and 700 gross outliers. The inlier noise scale is 0.01.

For a particular fitting result, let $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_N\}$ and $\hat{\boldsymbol{\omega}} = \{\hat{\omega}_1, \ldots, \hat{\omega}_{\hat{N}}\}$ respectively be the set of true and estimated line parameters of a particular method, where $\|\omega_p\| = 1$ and $\|\hat{\omega}_q\| = 1$. The error between a pair of parameters is obtained as $\|\omega_p - \hat{\omega}_q\|/\sqrt{2}$. We compute the *multi-structure fitting error* between $\boldsymbol{\omega}$ and $\hat{\boldsymbol{\omega}}$ as

$$\varepsilon = |N - \hat{N}| + \sum_{n=1}^{\min(N,\hat{N})} \min \epsilon_n. \quad (25)$$

The first term penalizes incorrect estimation of the number of structures. Symbol ϵ_n represents the set of all pairwise error between elements in $\boldsymbol{\omega}$ and $\hat{\boldsymbol{\omega}}$ at the n-th summation, where at each summation the pair with the lowest error in the *previous* summation are removed from $\boldsymbol{\omega}$ and $\hat{\boldsymbol{\omega}}$.

We test the performance of the methods under the influence of various outlier rates and inlier noise scales. For the former, we fix σ at 0.01 and vary L from 0 to 700 in steps of 50 (i.e. outliers rates from 75% to 94%), while for the latter we fix L at 200 and vary σ from 0.0025 to 0.025 in steps of 0.0025. For each L and σ, 100 repetitions of the data are created. We compute and average the fitting error of all methods across the repetitions. Fig. 6 shows the results. We stress that, as shown in Table 1, the methods differ in their level of dependence on manual parameter inputs, and KF is given none of the prior information available to the others.

The results reveal that the simplest method (RANSAC) can competently segment all the lines *if* the inlier noise scale and number of structures is known a priori. Our proposed method, however, is as accurate as "ideal" RANSAC

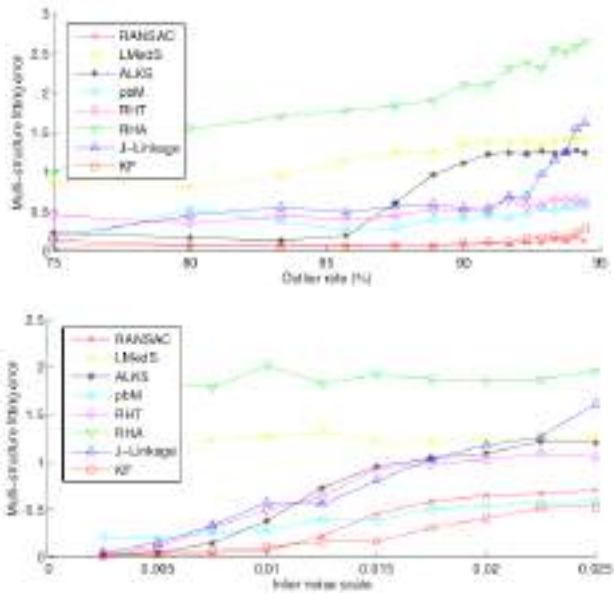


Figure 6. Performance comparison under various outlier rates (top) and inlier noise scale (bottom). For both experiments, 5000 random hypotheses are generated for all data repetitions, while parameter h for Kernel Fitting (KF) is fixed at 100.

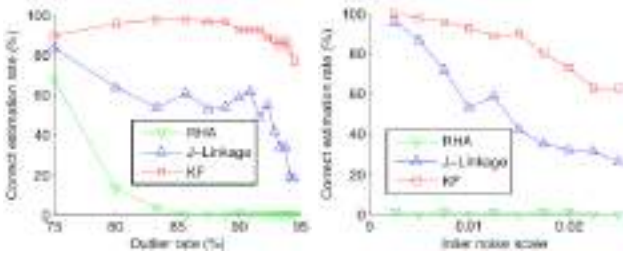


Figure 7. Performance of three methods in estimating the number of structures in the data as outlier rates and inlier noise scale vary.

without having to know these parameters in advance. The proposed Mercer kernel is also able to perform well under a large range of inlier noise scale, as Fig. 6 (bottom) shows, despite not being subjected to tuning (h was fixed at 100). Among the other methods, pbM and RHT returned the lowest error rates but they still differ by a large margin from KF and RANSAC. *Furthermore*, pbM and RHT require the unrealistic prior information of the number of structures or the number of points per structure. We also compare the ability of three "automatic" methods (RHA, J-Linkage, KF) in estimating the number of structures in the data. Fig. 7 shows the percentage of correct estimation across the repetitions. It can be seen that our method is able to estimate correctly at about 80% of the time, whereas RHA and J-Linkage succumb easily to gross outliers and inlier noise.

Fig. 8 shows more results of the proposed method on other 2D data, including on non-linear models.

Homography estimation. We test the ability of KF to

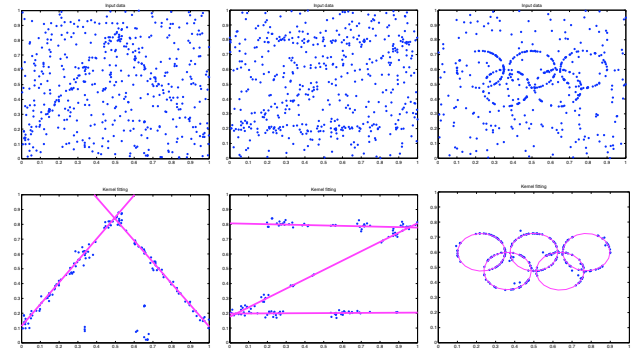


Figure 8. Results of KF on other 2D data. Left & centre: 90% outlier rate, $\sigma = 0.015$. Right: 93% outlier rate, $\sigma = 0.001$.

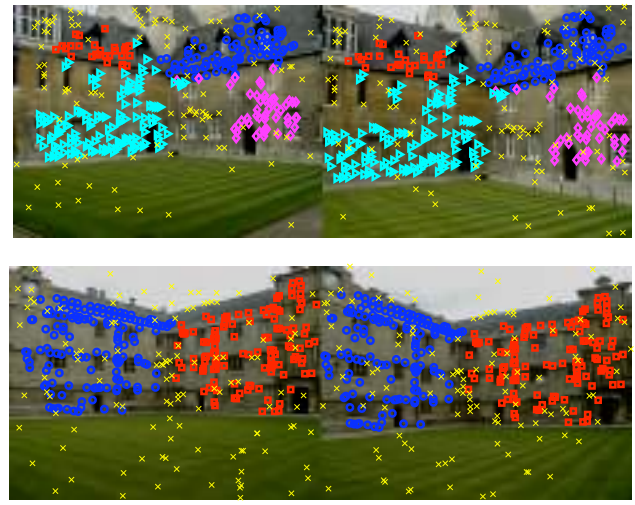


Figure 9. KF homography estimation results (in colour). The four (top pair) and two (bottom pair) planar structures were correctly detected. Yellow crosses are gross outliers as determined by KF.

detect planar homographies. Images of buildings in multiple views were obtained from the web[2] along with their precomputed interest point correspondences. For each image pair, 100 spurious correspondences were randomly added as gross outliers. For an image pair, we sample p-subsets of 8 points which form 4 correspondences and estimate a homography using the Direct Linear Transformation (DLT) algorithm [5]. We generate 5000 hypotheses in this manner and set $h = 100$ for the Mercer Kernel. The residual is computed as the geometric distance [5] between homography transformations. We complement the ORK with the Gaussian kernel, since points from the same plane should be close in 2D space. The results in Fig. 9 show that KF is able to simultaneously recover and estimate the number of homographies. The gross outliers were also successfully detected and precluded from homography estimation.

Motion segmentation. We also apply KF to the task of

[2]From http://www.robots.ox.ac.uk/ vgg/data/data-mview.html.

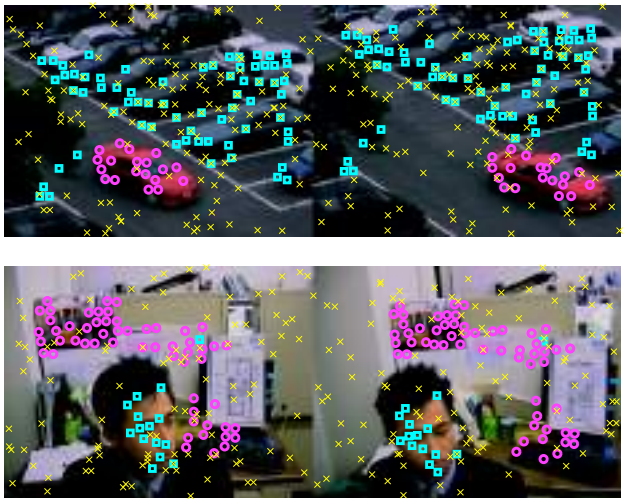


Figure 10. KF motion segmentation results (best viewed in colour). The two objects in each sequence were correctly segmented. Yellow crosses indicate gross outliers as determined by KF.

segmenting the motions of multiple rigidly moving objects under the affine camera model [14], where each motion occupies a subspace in the trajectory space. We obtain from the web[3] image sequences of multiple moving objects and the trajectories of feature points detected within. For each sequence, we randomly generate 100 spurious trajectories as gross outliers. As in [14] we generate 4D subspaces as hypotheses by invoking the SVD on p-subsets of size 4. We produce 5000 random hypotheses per sequence, and h is set to 100 for the Mercer kernel. The residual is computed as the orthogonal projection distance onto the subspace. The results[4] in Fig. 10 show that KF is able to separate the true trajectories from the false trajectories, discover the correct number of motions in the sequences and label the feature points according to the objects they belong to.

5. Conclusions and Future Work

We have presented a novel approach to robust fitting of multiple structures by using statistical learning techniques. Central to our idea is a Mercer kernel designed for the task of robust fitting. Our approach can identify and remove gross outliers, discover the true number of model instances and estimate model parameters for the individual structures. Our experiments show that the proposed method outperforms other methods in terms of fitting accuracy, and that it is also highly competent in practical vision tasks.

We plan to evaluate further the performance of the proposed method on publicly available benchmark datasets, e.g. for motion segmentation [14], so that we can obtain a comprehensive comparison against other methods. Since our method is a generic robust fitting approach, it would also be interesting to customize it for specific tasks.

[3] From http://www.suri.cs.okayama-u.ac.jp/e-program-separate.html.

[4] Extended results are available in the supplementary material.

References

[1] H. Chen and P. Meer. Robust regression with projection based M-estimators. In *ICCV*, 2003.

[2] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, 24(5):603–619, 2002.

[3] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Comm. of the ACM*, 15:11–15, 1972.

[4] M. A. Fischler and R. C. Bolles. Random sample concensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24:381–395, 1981.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, 2000.

[6] K.-M. Lee, P. Meer, and R.-H. Park. Robust adaptive segmentation of range images. *TPAMI*, 20(2):200–205, 1998.

[7] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley, 1987.

[8] R. Sedgewick. *Algorithms in C: Parts 1–4*. Addison-Wesley, 3rd edition, 1998.

[9] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[10] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 22(8):888–905, 2000.

[11] R. Subbarao and P. Meer. Nonliner mean shift for clustering over analytic manifolds. In *CVPR*, 2006.

[12] R. Toldo and A. Fusiello. Robust multiple structures estimation with J-Linkage. In *ECCV*, 2008.

[13] P. Torr and A. Zisserman. MLESAC: A new robust estimator with applications to estimating image geometry. *CVIU*, pages 138–156, 2000.

[14] R. Tron and R. Vidal. A benchmark for the comparison of 3-D motion segmentation algorithms. In *CVPR*, 2007.

[15] V. Vapnik. *The nature of statistical learning theory*. Berlin: Springer-Verlag, 1995.

[16] M. P. Wand and M. C. Jones. *Kernel smoothing*. Chapman & Hall, 1995.

[17] H. Wang and D. Suter. Robust adaptive-scale parametric model estimation for computer vision. *TPAMI*, 26(11):1459–1474, 2004.

[18] L. Xu, E. Oja, and P. Kultanen. A new curve detection method: randomized Hough transform (RHT). *Pattern Recognition Letters*, 11(5):331–338, 1990.

[19] W. Zhang and J. Kosecká. Ensemble method for robust motion estimation. In *25 years of RANSAC workshop, CVPR*, 2006.

[20] W. Zhang and J. Kosecká. Nonparametric estimation of multiple structures with outliers. In *Dynamical Vision, ICCV 2005 and ECCV 2006 Workshops*, 2006.

[21] M. Zuliani, C. S. Kenney, and B. S. Manjunath. The multi-RANSAC algorithm and its application to detect planar homographies. In *ICIP*, 2005.