Robust Fusion of Uncertain Information

Haifeng Chen and Peter Meer

***Abstract*—A technique is presented to combine n data points, each available with point-dependent uncertainty, when only a subset of these points come from $N \ll n$ sources, where N is unknown. We detect the significant modes of the underlying multivariate probability distribution using a generalization of the nonparametric *mean shift* procedure. The number of detected modes automatically defines N, while the belonging of a point to the basin of attraction of a mode provides the fusion rule. The robust data fusion algorithm was successfully applied to two computer vision problems: estimating the multiple affine transformations, and range image segmentation.**

***Index Terms*—Computer vision, information fusion, mean shift, robust analysis.**

I. Introduction

Under its most general form, information fusion is a process in which the available data is combined to find representations of higher quality. See [31] for a discussion of the concepts involved in data fusion, and [30] for an extensive analysis of the recent literature.

In different research areas information fusion has different meanings. In engineering and applied sciences it is most often identified with sensor fusion, and the goal is to improve the representations provided by individual sensing modalities. See [27] and [28] for typical collections of papers. In pattern recognition, statistics and machine learning, information fusion is related to combining classifiers to increase classification accuracy, e.g., [18], [24], or handling ensembles of outputs obtained by resampling the input, e.g., [12], [20]. We approach the information fusion problem as a generic one and the proposed technique can be integrated into most of the methods described in the literature.

Let the p-dimensional data points be $\mathbf{x}_i \in R^p$, $i = 1, \ldots, n$, each associated with a covariance matrix $\mathbf{C}_i$ characterizes the uncertainty of the process through which $\mathbf{x}_i$ was obtained. The data can come from an *unknown* number of information sources, and will denote this number by N. To have a realistic setup, will also assume that only a subset of the data is actually related to the information sources, others (possibly the majority) are severely corrupted data points, i.e., outliers. The goal is to characterize the N sources.

Whenever a data point is related to one of the information sources, it is assumed to carry an *unbiased* representation of that source. Since we focus here on computational aspects of information fusion, no distinction has to be drawn based on the nature of the data. This distinction is important, however, when analyzing different classifier combination rules under the Bayesian paradigm [29], or in the context of machine learning [18].

There have been several studies in the computation of multiple source models in the computer vision tasks, such as multiple objects motions, multiple epipolar geometries, etc. However, those methods are either sensitive to the outliers [8], [10] or sensitive to the parameters of the algorithm [11]. Here we address those problems by robust data fusion. We show that our technique is superior to Hough transform and other clustering based methods in dealing with the multiple structured data. The fusion based algorithm was applied to two generic computer vision problems: estimation of the affine transformation parameters from moving objects, and range image segmentation. In both applications, the proposed algorithm gave satisfactory results.

In Section II kernel density estimation is reviewed. In Section III information fusion of data from a single or multiple sources is discussed. In Section IV the computational technique employed to characterize the information sources, the mean shift procedure is described. The main advantages of data fusion are presented in Section V. Two fusion examples are shown in Section VI, and the relation of the proposed method to other approaches is discussed in Section VII.

II. Kernel Density Estimation

Kernel density estimation, called the Parzen window technique in the pattern recognition literature [19, Sec.4.3], is the most popular density estimation method. See [32] for a thorough treatment of the subject in the statistical literature.

Given the data points $\mathbf{x}_i \in R^p$, the estimate of the underlying probability distribution at location $\mathbf{x}$ is computed as the average of *continuous* kernel functions centered on the data points

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}} (\mathbf{x} - \mathbf{x}_i) . \tag{1}$$

We will consider only radially symmetric kernels with bounded support, i.e., $K(\mathbf{x}) = 0$ for $\|\mathbf{x}\| > 1$. The $p \times p$ symmetric positive definite bandwidth matrix $\mathbf{H}$ scales the kernel support to the desired elliptical shape and size

$$K_{\mathbf{H}}(\mathbf{x}) = [\det[\mathbf{H}]]^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x}). \tag{2}$$

The case $\mathbf{H} = h^2 \mathbf{I}_p$ yields the often used multivariate kernel density estimator expression for (1), where the radius of the kernel is defined by the bandwidth h.

The bandwidth matrix $\mathbf{H}$ is the critical parameter of the kernel density estimator. For example, if the support of the kernel is too large, significant features of the distribution, like multimodality, can be missed by oversmoothing. A single bandwidth matrix may not be enough when the distribution has very different structure locally.

There are two ways to adapt the bandwidth to the local structure, in each case the adaptive behavior being achieved by first performing a pilot density estimation. The bandwidth matrix can be either associated with the location $\mathbf{x}$ in which the distribution is to be estimated, or each data point $\mathbf{x}_i$ can be taken into account in (1) with its own bandwidth matrix, i.e.,

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}_i} (\mathbf{x} - \mathbf{x}_i) . \tag{3}$$

It can be shown that (3), called the *sample point* density estimator, has superior statistical properties [22].

Our radially symmetric kernels satisfy

$$K(\mathbf{x}) = c_{k,p} k(\mathbf{x}^\top \mathbf{x}) > 0 \quad \|\mathbf{x}\| \leq 1 \tag{4}$$

where $k(x)$ is called the *profile* of the kernel, and the normalization constant $c_{k,p} > 0$ assures that $K(\mathbf{x})$ integrates to one. We have $k(x) > 0$ for $0 \leq x < 1$, and w.l.g. will be considered monotonically decreasing with x. The Epanechnikov kernel, having the profile

$$k(x) = \begin{cases} 1 - x, & 0 \leq x \leq 1 \\ 0, & x > 1 \end{cases} \tag{5}$$

Manuscript received 2003; revised 2004. This work was supported by the National Science Foundation under Grant IRI 99-87695. This paper was recommended by Guest Editor B. Bhanu.

H. Chen is with the NEC Laboratories America, Inc., Princeton, NJ 08540 USA (e-mail: haifeng@nec-labs.com)

P. Meer is with the Electrical and Computer Engineering Department, Rutgers University, Piscataway, NJ 08854 USA (e-mail: meer@caip.rutgers.edu)

Digital Object Identifier 10.1109/TSMCB.2005.846659

will be used in the examples. The Epanechnikov kernel is optimal in the sense of asymptotic mean square error of the estimated distribution [32, p. 104]. From (3), taking into account (4) and (2), we obtain

$$\hat{f}(\mathbf{x}) = \frac{c_{k,p}}{n} \sum_{i=1}^{n} [\det[\mathbf{H}_i]]^{-1/2} k\left((\mathbf{x}-\mathbf{x}_i)^\top \mathbf{H}_i^{-1}(\mathbf{x}-\mathbf{x}_i)\right). \quad (6)$$

III. Data Fusion

We discuss first the simplest case when *all* the available data is related to a single source of information. Next, will describe the more realistic situation in which not only several information sources should be handled simultaneously (without *a priori* knowledge of their number) but also the presence of spurious data should be tolerated.

A. Single Source

When *all* the data points $\mathbf{x}_i \in R^p$, $i = 1, \ldots, n$ are related to the same source, the underlying probability distribution is unimodal. In the sequel the data (measurements) are assumed to be independent random variables, an assumption which may not necessarily be accurate in reality, and it is not crucial for the proposed method. The uncertainty about the data point $\mathbf{x}_i$ is described by the covariance matrix $\mathbf{C}_i$. The data points are unbiased estimates of $\hat{\mathbf{x}}$, the value characterizing the information source. Therefore, the source should be taken as the center of the cluster of these points, i.e.,

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum_{i=1}^{n} (\mathbf{x}-\mathbf{x}_i)^\top \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{x}_i) \quad (7)$$

where the covariances are assumed to have full rank. The rank deficient case can be easily handled by using pseudoinverses in (7), though the presence of null spaces may require some additional care. It is easy to show that

$$\hat{\mathbf{x}} = \left(\sum_{i=1}^{n} \mathbf{C}_i^{-1}\right)^{-1} \sum_{i=1}^{n} \mathbf{C}_i^{-1}\mathbf{x}_i \quad (8)$$

i.e., the source is characterized by the covariance weighted average of the data. The more uncertain is a data point (the inverse of its covariance has a smaller norm), the less it contributes to the result of the fusion. This is a desirable behavior, though will not defend against an outlier with small uncertainty far away from the good data. Thus, the fusion rule (8) is not robust.

The uncertainty of $\hat{\mathbf{x}}$ is also of interest. Without loss of generality we can assume that the *true* $\hat{\mathbf{x}} = \mathbf{0}$. The covariance of $\hat{\mathbf{x}}$, obtained after some simple manipulations, is

$$\hat{\mathbf{C}} = E[\hat{\mathbf{x}}\hat{\mathbf{x}}^\top] = \left(\sum_{i=1}^{n} \mathbf{C}_i^{-1}\right)^{-1}. \quad (9)$$

The covariance of the fusion output is the harmonic mean of the data covariances. Note that the estimator (8) is consistent. Indeed, by taking all $\mathbf{C}_i = \mathbf{C}$, the covariance estimate becomes $\hat{\mathbf{C}} = n^{-1}\mathbf{C}$ and vanishes when the number of measurements becomes very large.

B. Multiple Sources

We will relax now the single source condition imposed in the previous section. A data point may come from *any* of the $N \ll n$ information sources, or can be an outlier not related to these sources. It is assumed that the outliers are not structured, i.e., they do not yield "phantom" sources. The data points related to an information source being unbiased representations of that source, we should seek the centers of the N densest regions in R^p, i.e., the modes of the underlying probability distribution. These modes will be localized by using the sample point kernel density estimator (3).

The robustness of our approach will be assured by defining the bandwidth matrices as

$$\mathbf{H}_i = \chi^2_{\gamma,p}\mathbf{C}_i \quad i = 1, \ldots, n \quad (10)$$

where $\chi^2_{\gamma,p}$ is the chi-square value for p degrees of freedom and level of confidence γ. In our implementation $\gamma = 0.995$.

As a consequence of our choice for the bandwidth matrices the support of the kernel $K_{\mathbf{H}_i}(\mathbf{x}-\mathbf{x}_i)$ is the confidence region of $\mathbf{x}_i$ at coverage probability γ. Indeed, from (4) we have

$$K_{\mathbf{H}_i}(\mathbf{x}-\mathbf{x}_i) > 0 \text{ for } (\mathbf{x}-\mathbf{x}_i)^\top \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{x}_i) < \chi^2_{\gamma,p}. \quad (11)$$

When the data $\mathbf{x}_i$ is related to a source characterized by $\hat{\mathbf{x}}_l$, the ellipsoidal region in R^p centered on $\mathbf{x}_i$ and delineated by the condition (11) contains $\hat{\mathbf{x}}_l$ with probability γ. Note that the χ^2 values are used to scale the confidence region, which implicitly assumes that the difference $\mathbf{x}-\mathbf{x}_i$ is a zero mean Gaussian variable with covariance $\mathbf{C}_i$. In fact, the result of the fusion is only weakly dependent on the accurate size of the confidence region, as long as all these regions are scaled the same way.

Taking into account (10) the density estimate (6) becomes

$$\hat{f}(\mathbf{x}) = \frac{c_{k,p}}{n\left[\chi^2_{\gamma,p}\right]^{p/2}} \sum_{i=1}^{n} [\det[\mathbf{C}_i]]^{-1/2} \times k\left(\frac{1}{\chi^2_{\gamma,p}}(\mathbf{x}-\mathbf{x}_i)^\top \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{x}_i)\right) \quad (12)$$

with the modes being the significant *local* maxima

$$\hat{\mathbf{x}}_l = \arg\max_{\mathbf{x}} \hat{f}(\mathbf{x}) \quad l = 1, \ldots, N. \quad (13)$$

Note that by identifying the modes the value of N is automatically determined. They are defined by the zeros of the gradient of the density estimate. Stationary points also correspond to such zeros, but they can be avoided in our approach as will be discussed in Section IV. The gradient of the density estimate is computed from (13)

$$\nabla \hat{f}(\mathbf{x}) = \frac{c_{k,p}}{n\left[\chi^2_{\gamma,p}\right]^{(p/2+1)}} \times \sum_{i=1}^{n} [\det[\mathbf{C}_i]]^{-1/2}\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{x}_i) \times k'\left(\frac{1}{\chi^2_{\gamma,p}}(\mathbf{x}-\mathbf{x}_i)^\top \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{x}_i)\right). \quad (14)$$

We will define the function $g(x) = -k'(x)$ which is also a profile since $k(x)$ is monotonically decreasing with x. After scaling the covariance matrices

$$\mathbf{W}_i = [\det[\mathbf{C}_i]]^{1/2}\mathbf{C}_i \quad (15)$$

the expression of the gradient estimate can be rewritten as

$$\nabla \hat{f}(\mathbf{x}) = \frac{c_{k,p}}{n\left[\chi^2_{\gamma,p}\right]^{(p/2+1)}} \left(\sum_{i=1}^{n} I_\gamma(\mathbf{x}-\mathbf{x}_i)\mathbf{W}_i^{-1}\right) \times \left[\left(\sum_{i=1}^{m} I_\gamma(\mathbf{x}-\mathbf{x}_i)\mathbf{W}_i^{-1}\right)^{-1} \times \left(\sum_{i=1}^{n} I_\gamma(\mathbf{x}-\mathbf{x}_i)\mathbf{W}_i^{-1}\mathbf{x}_i\right) - \mathbf{x}\right] \quad (16)$$

where

$$I_\gamma(\mathbf{u}) = g\left(\frac{\mathbf{u}^\top \mathbf{C}_{\mathbf{u}}^{-1}\mathbf{u}}{\chi^2_{\gamma,p}}\right) = \begin{cases} 1, & \mathbf{u}^\top \mathbf{C}_{\mathbf{u}}^{-1}\mathbf{u} \le \chi^2_{\gamma,p} \\ 0, & \mathbf{u}^\top \mathbf{C}_{\mathbf{u}}^{-1}\mathbf{u} > \chi^2_{\gamma,p} \end{cases} \quad (17)$$

with the second definition being the specific value for the case of the Epanechnikov kernel (5) used in the paper.

The zeros of the density gradient estimate (16) are then the solutions of the equation

$$\mathbf{x} = \left(\sum_{i=1}^{n} I_\gamma(\mathbf{x}-\mathbf{x}_i)\mathbf{W}_i^{-1}\right)^{-1}\left(\sum_{i=1}^{n} I_\gamma(\mathbf{x}-\mathbf{x}_i)\mathbf{W}_i^{-1}\mathbf{x}_i\right). \quad (18)$$

By comparing (8) with (18) we see that both expressions are covariance weighted averages of the data, but in (18) the scalar function $I_\gamma(\mathbf{x}-\mathbf{x}_i)$ is also present in the weights. This function has nonzero values only over the confidence region of $\mathbf{x}_i$, see (17). Thus, the computation of the kernel density estimates is restricted to local regions and the detection of each mode is handled separately, yielding a robust behavior. To solve (18) an iterative numerical procedure will be employed.

IV. Mean Shift Procedure

The mean shift property was first described in 1975 by Fukunaga and Hostetler [21] in the context of pattern recognition. It provides an efficient way to locally estimate the gradient of a density. Recently mean shift became a popular tool in computer vision. See [15]–[17] for issues related to the topic of this paper. The mean shift property is less known in the statistical literature. While the book [26, Sec.6.2.2] discusses [21], the usefulness of mean shift type approaches to density estimation was only recently recognized [14].

To describe the mean shift property, will assume for the moment that all the covariance matrices are the same and proportional to the identity matrix

$$\mathbf{C}_i = \frac{h^2}{\chi^2_{\gamma,p}}\mathbf{I}_p \quad i = 1,\ldots,n \quad (19)$$

where h is the radius (bandwidth) of the kernel. Then (17) becomes

$$I_h(\mathbf{x}-\mathbf{x}_i) = \begin{cases} 1, & \|\mathbf{x}-\mathbf{x}_i\| \le h \\ 0, & \|\mathbf{x}-\mathbf{x}_i\| > h. \end{cases} \quad (20)$$

Since now all $\mathbf{W}_i = c\mathbf{I}_p$ (15), we can define from (18) the *mean shift* vector

$$\mathbf{m}_h(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbf{x}_i I_h(\mathbf{x}-\mathbf{x}_i)}{\sum_{i=1}^{n} I_h(\mathbf{x}-\mathbf{x}_i)} - \mathbf{x} \quad (21)$$

i.e., the difference between the average of the data points in a window of radius h, and its center $\mathbf{x}$. The mean is biased toward the region containing the majority of the points, i.e., $\mathbf{m}_h(\mathbf{x})$ is oriented from the center toward the region of highest density in the window. Thus, moving the window into the location specified by the mean shift vector, will place it over a region of higher density than the current position. It can be shown that this iterative procedure is guaranteed to converge to the nearest stationary point where the density gradient estimate vanishes. See [16] for a more rigorous description of the mean shift property.

In most cases the point of convergence is a local maximum of the density, i.e., a mode. The modes have the property that after a small random perturbation the new mean shift process will return to the original point of convergence. For nonmaximum stationary points, like saddle point, this is much less probable and thus they can be easily eliminated.

The following procedure finds the modes of the distribution underlying the measurement.

Starting from $\mathbf{x} = \mathbf{x}_i$, $i = 1,\ldots,n$:

1) compute the mean shift vector $\mathbf{m}_h(\mathbf{x})$
2) translate the window by $\mathbf{m}_h(\mathbf{x})$
3) if $\mathbf{m}_h(\mathbf{x})$ larger than tolerance return to 1.
4) store point of convergence.

Due to numerical issues the n mean shift procedures may converge to n different points. However, the convergence points around each of the N modes can be grouped together using the bandwidth parameter h as a measure of proximity. This parameter sets the resolution of the density estimation and therefore the resolution of the mode detection process as well.

Given a detected mode, the data points which converged to it define its *basin of attraction*. Only the N modes have significant basins of attraction since the outliers in the background are in regions of low density. It is important to emphasize that a basin of attraction can have an arbitrary shape since each data point $\mathbf{x}_i$ was processed separately. This is the main advantage of the mean shift based clustering technique over the traditional approaches in which quadratic proximity measures are used and thus an elliptical shape is imposed for the clusters. See [15] for a detailed discussion of clustering with the mean shift procedure. The low density boundary regions of a basin of attraction can be removed by postprocessing.

The two-dimensional data in Fig. 1(a) is used as a simple example. In Fig. 1(b) the probability distribution underlying the data is shown. Note that presence of three significant modes, the complex shapes of two of the three clusters, and the large contamination of the background. Nevertheless, the mean shift procedure successfully recovers the three modes, and their basins of attraction are close to the original shape of the clusters [Fig. 1(c)].

A. Application to Data Fusion

We return now to the general case when each data point is associated with a different covariance matrix. The mean shift iterations (18) become

$$\mathbf{x}^{(m+1)} = \left(\sum_{i=1}^{n} I_\gamma\left(\mathbf{x}^{(m)}-\mathbf{x}_i\right)\mathbf{W}_i^{-1}\right)^{-1} \times \left(\sum_{i=1}^{n} I_\gamma\left(\mathbf{x}^{(m)}-\mathbf{x}_i\right)\mathbf{W}_i^{-1}\mathbf{x}_i\right) \quad (22)$$

where m is the iteration index. Taking into account (15) and (20) and we can recognize at every iteration of (22) an expression similar to (8) is computed. However, now the computations are restricted to a local region due to the presence of the indicator function I_γ. After the density gradient was estimated in $\mathbf{x}$, to compute the next location of the window of analysis, i.e., the mean shift vector, *only* those $\mathbf{x}_i$ are taken into consideration whose region of confidence contains $\mathbf{x}$. This is the property which assures the robustness of the multiple source data fusion.

The basin of attraction of each point of convergence defines the subset of the data to be used when computing its characteristics, $\hat{\mathbf{x}}$ (8) and $\hat{\mathbf{C}}$ (9). To group the points of convergence around a significant mode their mutual relation is verified. Let $\hat{\mathbf{x}}_a$, $\hat{\mathbf{C}}_a$ and $\hat{\mathbf{x}}_b$, $\hat{\mathbf{C}}_b$ be associated with the points of convergence labeled a and b respectively. They will be grouped together, iff

$$(\hat{\mathbf{x}}_a - \hat{\mathbf{x}}_b)^\top \hat{\mathbf{C}}_j^{-1}(\hat{\mathbf{x}}_a - \hat{\mathbf{x}}_b) \le \chi^2_{\gamma,p} \quad j = a, b \quad (23)$$

i.e., each point of convergence has to be inside the region of confidence of the other one. At the end of the grouping process, the basins of attraction associated with the modes are defined and the final $\hat{\mathbf{x}}$ and $\hat{\mathbf{C}}$ are computed. The condition (23) can also be used before the mean shift procedure to eliminate isolated outliers and thus reduce the amount of computations.

V. Why Robust Data Fusion

Clustering based feature space analysis is a common technique to approach the computer vision problems and has numerous advantages [15], [16]. The fusion strategy we proposed in this paper actually is a clustering algorithm. However, our method addresses the drawbacks of traditional clustering methods, such as the k-means.

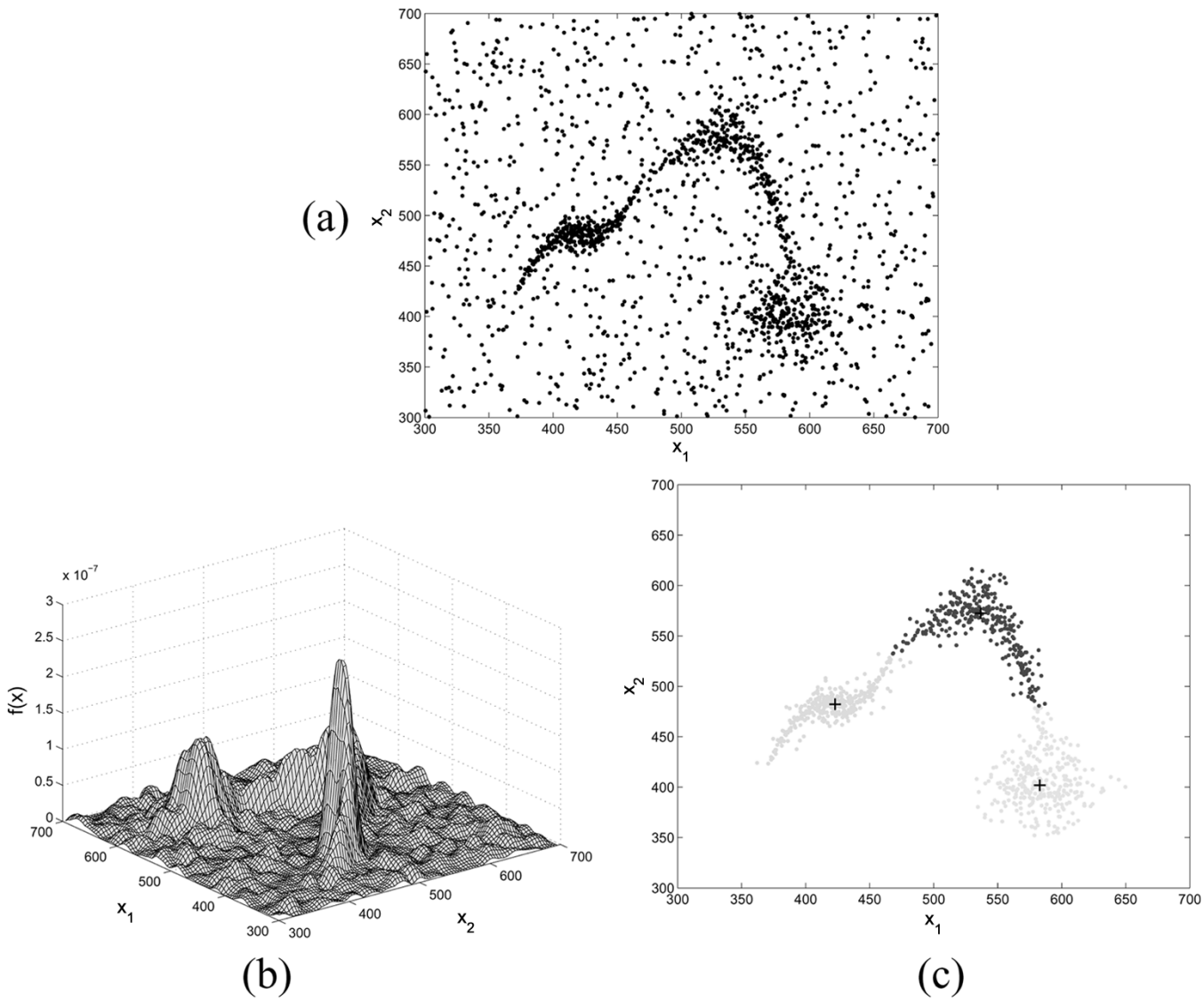

Fig. 1. Example of clustering using the mean shift procedure. (a) Two-dimensional input. (b) Kernel density estimate of the underlying distribution. (c) Basins of attraction of the three significant modes (marked "+").

One major drawback of the traditional clustering methods is their sensitivity to the choice of the number of clusters in the feature space. This sensitivity is an important limitation in computer vision, where only rarely is the number of significant feature properties known *a prior*. In our robust fusion approach, the number of detected modes automatically defines the number of information sources, which is equivalent to the number of clusters k. Another problem of the traditional clustering methods is their sensitivity to the outliers. In the k-means clustering, adding an outlier located very far away from the clusters may change the results a lot. To solve this problem, a lot of robust clustering methods were proposed [5], [16]. However, those methods are still sensitive to the parameters used by the algorithm. For instance, the efficacy of mean shift [16] involves the specification of a bandwidth parameter, which is critical to the performance of the algorithm. Our approach adapts the bandwidth to the local structure of the data distribution. The fusion process exploits the uncertainty of each measurement, and incorporates such uncertainties into the bandwidth determination. By doing so, we could avoid the bandwidth selection problem, and achieve much reliable results.

To illustrate the potential of the proposed data fusion method we discuss a simple example. We return to the data in Fig. 1. Twenty points were drawn randomly from each of the three clusters (information sources), and forty points from the outliers. The 100 measurements are shown in Fig. 2(a) where the true values of the information sources, the centers of the three clusters, are marked "+." Each data point was associated with a randomly generated region of confidence [Fig. 2(b)]. For the 60 points chosen from the clusters, the related information source was inside the region of confidence. The robust data fusion process accurately extracts the three information sources, and associates small regions of confidence with them [Fig. 2(c)]. However, the k-means based clustering is unable to recover the correct information sources, because it could not deal with outliers in the data. Also for the fixed bandwidth mean shift, its performance is dependent on the correct bandwidth selection since it did not account the uncertain information of each point. On the other hand, our fusion based approach comes up with satisfactory results. Note that for visualization purposes, we only deal with low dimensional data in this example. However, much higher dimensional and more complex data can be handled the same way.

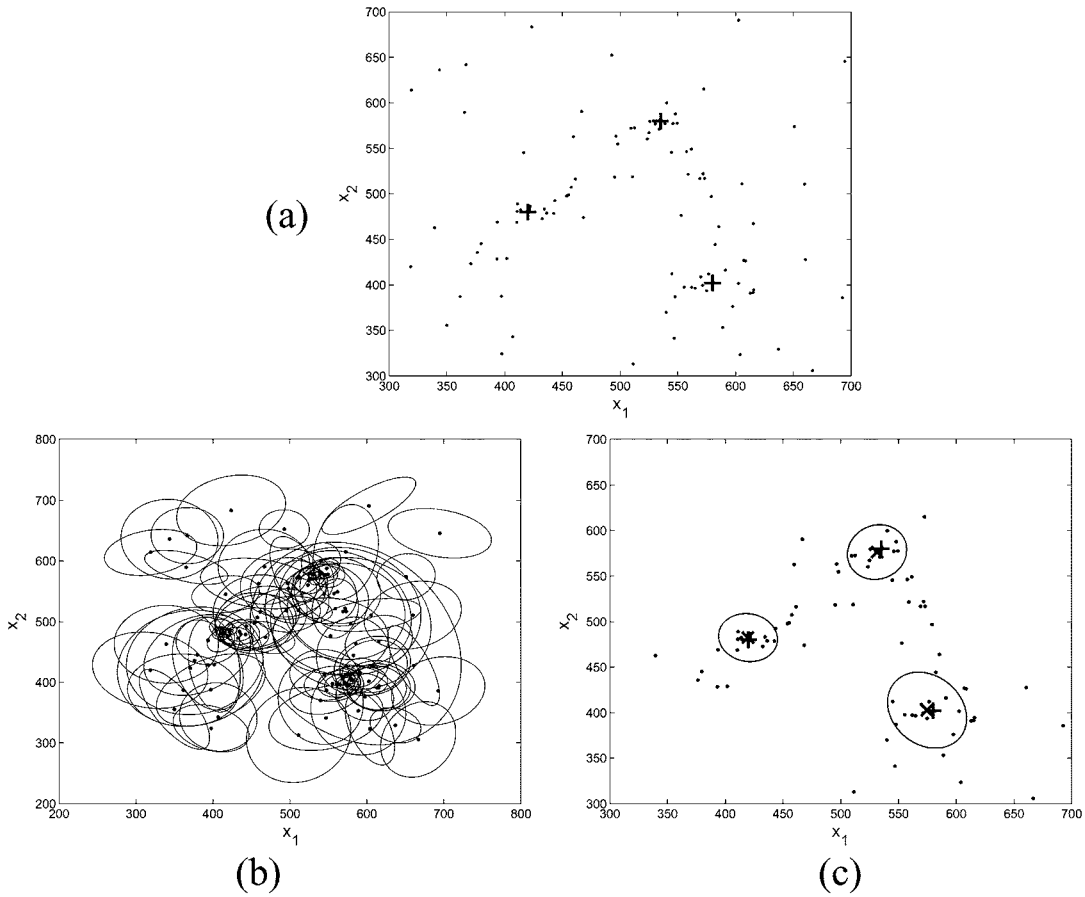

Fig. 2. Robust fusion of measurements drawn from the data in Fig. 1(a). (a) Selected data points. The information sources are marked "+." (b) Associated confidence regions. Note the slightly different scale of the graph. (c) Regions of confidence of the three detected modes (marked "×"). The true values are marked "+."

VI. Applications

Data containing multiple structures is characterized by the presence of several instances of the same model, each defined with a different set of parameters. As two computer vision examples, these structures may be different surfaces in depth measurements, or multiple moving objects in motion estimation. There have been a lot of studies of dealing with multiple structures [8], [10], [11]. But these methods are either sensitive to the outliers [8], [10], or sensitive to the parameters used in the algorithm [11]. For instance, Hough transform is a traditional technique for multiple structures extraction. But care has to be taken when one quantizes the feature space. When the bins of the feature space is too fine, the two candidates for the same structure may be in different bins. If the bin size is not fine enough, on the other hand, two similar structures (like two parallel lines which are close to each other) will lie in the same bin. To real applications, however, there is not enough *a priori* knowledge to reliably define the bin size. Furthermore, our research shows that when the noise level corrupting the data is high, there may not exist suitable bin size for the Hough Transform [9]. In this paper, we apply robust data fusion technique in the Hough space to detect the multiple structures. Our fusion process is a clustering based analysis. However, as we discussed in the previous section, it has superior advantages than traditional clustering methods. We have conducted two real experiments: estimation of the affine transformation parameters from moving objects, and range image segmentation. Our results have shown that the fusion based approach indeed has the desired abilities and great advantages.

A. Multiple Affine Transformations

In this example, a pair of images of three independently moving objects was considered [Fig. 3(a) and (b)]. The motion of each object is modeled by an affine transformation between the (unknown) true coordinates of salient points in correspondence $\mathbf{y}_{ko} = [y_{k1o} y_{k2o}]^\top$, $k = 1, 2$,

$$\begin{bmatrix} y_{21o} \\ y_{22o} \end{bmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{bmatrix} y_{11o} \\ y_{12o} \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \tag{24}$$

which can be decoupled into two three-dimensional problems [23]. Here will consider the one in variables (y_{11}, y_{12}, y_{21}) and parameters a_{11}, a_{12}, t_1.

To obtain the data, first a large number of point correspondences were established and the matches on the static background, identified by having zero displacement, were removed. The estimation process used the remaining 282 point correspondences. When this data is mapped into the space of the variables, the correspondences associated with the three moving objects yield points near to three planes [Fig. 3(c)]. Note also the points in the background due to erroneous matches.

To identify the points belonging to the individual objects we solve the multiple regression problem with a technique which can be regarded as a generalization of the Hough transform. One hundred and fifty samples, each containing around 30 points, were chosen by guided random sampling with preference toward denser regions in the data. From each sample the parameters of a plane were robustly estimated. The plane is parameterized by (x_1, x_2, x_3), the coordinates of its point closest to the origin. See [13] for a detailed description of all the issues about sampling and parameter estimation.

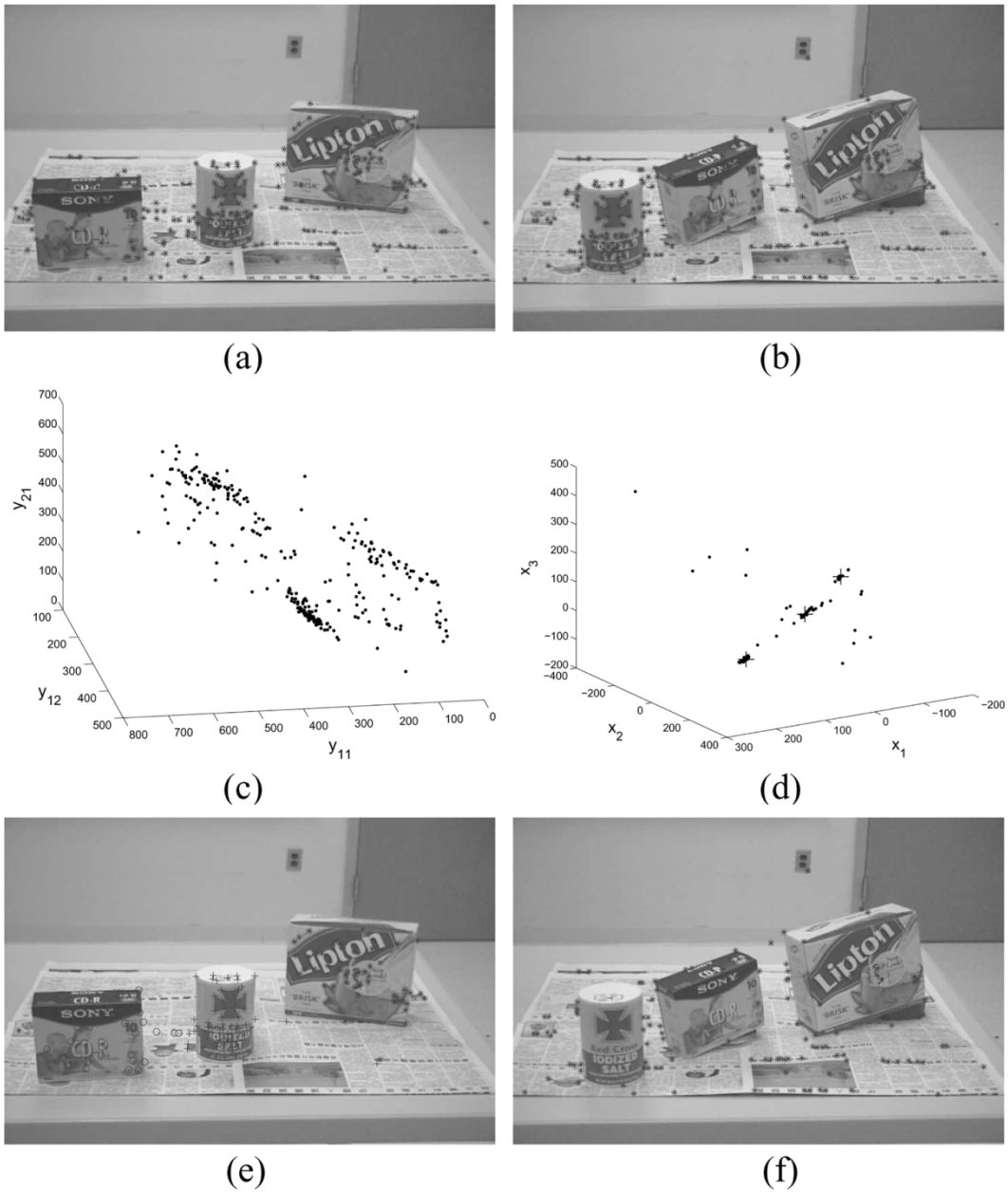

Fig. 3. Robust fusion in a multiple regression problem, detection of three moving objects through estimation of three affine transformations. (a)–(b) Two image frames with marked salient points. (c) Distribution of point correspondences in the subspace (y_{11}, y_{12}, y_{21}). (d) Parameter space used for data fusion. The three detected modes are marked "+." (e) Points associated with the moving objects, as seen in the first frame. Each object is labeled with a different symbol. (f) Points not associated with any object, as seen in the second frame.

The parameter estimation process also returns the covariance of the estimates. As the associated confidence regions are difficult to visualize, only the parameter estimates are shown in Fig. 3(d). The robust fusion technique described in this paper was then used to detect three modes, marked "+." The associated confidence ellipsoids are too small to display. By deriving the number of modes from the data we also determined the number of significant independently moving objects in the scene. The basins of attraction of the modes delineate the points related to each of the moving objects. The final delineation of the objects is based on the joint results of the two estimation processes.

The estimated affine transformation parameters for each object are presented in Table I. Among the 282 points, we detected 169 object points and 113 background points. As can be seen in Fig. 3(e), the detected object points are correctly assigned to its related object (marked by "$*$," "$\circ$," and "x," respectively). Only a few background points are wrongly taken as object points, which is probably because those background points happen to satisfy the affine motion constraints. From Fig. 3(f), it is shown that the detected background points are either on the background or on the boundaries of objects. Those boundary points

TABLE I
ESTIMATED TRANSFORMATION PARAMETERS

	a_{11}	a_{12}	a_{21}	a_{22}	t_1	t_2
tea box	0.737	4.455	0.794	2.825	19.79	115.8
cd box	0.627	4.611	0.760	2.837	156.2	12.09
salt container	2.324	1.518	0.818	3.112	151.2	15.11

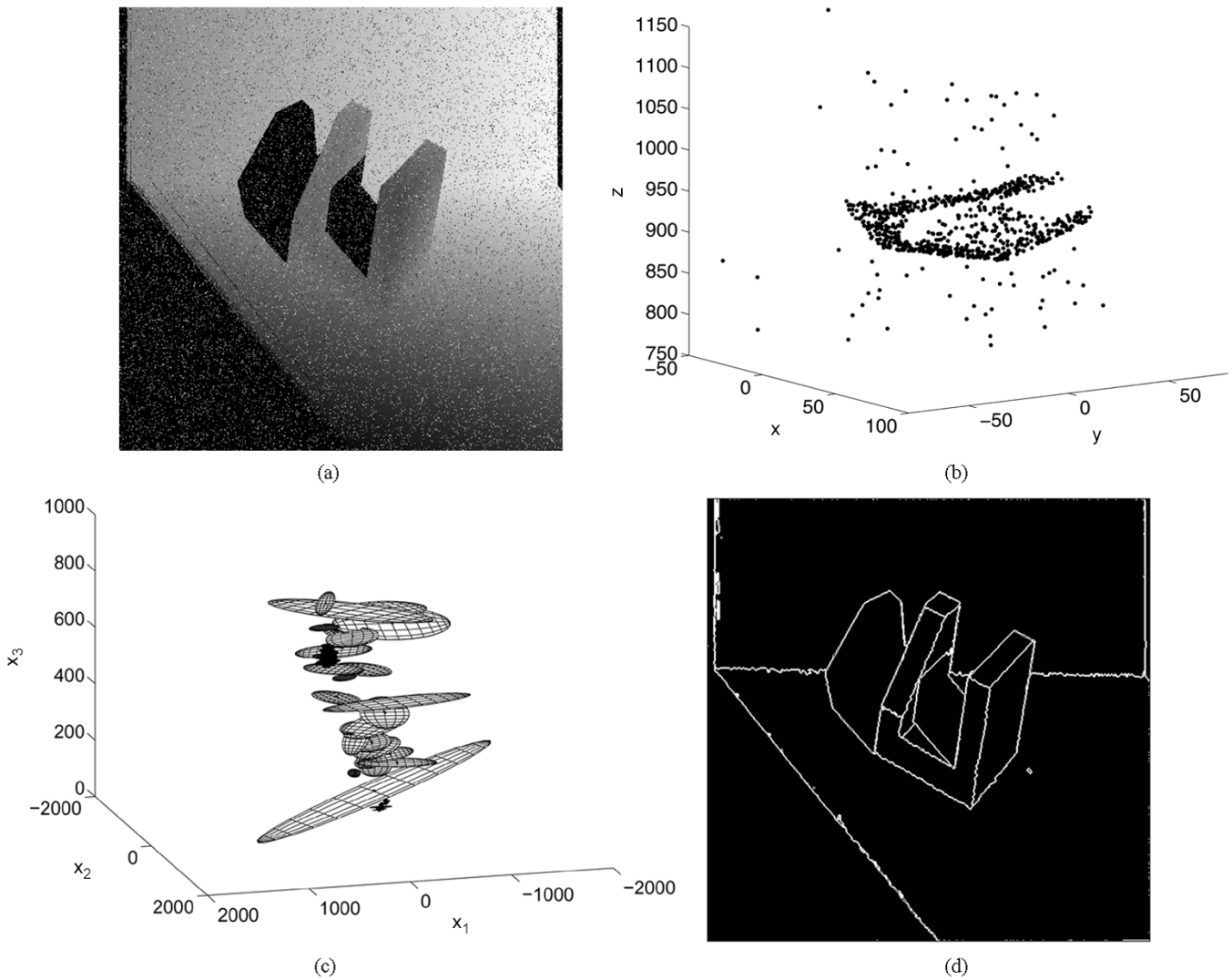

Fig. 4. Range image segmentation by robust data fusion. (a) Range image from USF database with 26 200 random outliers added. (b) Sampled points in the second iterative step. (c) Associated confidence regions related to the estimated regression parameters. (d) Segmentation results.

are treated as outliers since they are occluded in the other image. Due to the outliers and the unavailability of the knowledge about number of moving objects, traditional clustering methods, such as the k-means, is unable to segment three moving object correctly.

B. Range Image Segmentation

Range image provides three-dimensional geometric information about the objects in a scene. Segmentation is to break the image into some meaningful nonoverlapping homogeneous regions whose union is the entire image. An often adopted approach is to fit surface models to range data subsets and then calculate geometric features using the estimated model parameters [2], [5]. To make the algorithm tolerate the outliers, some robust estimators were employed in fitting the surface model [2], [3]. Since the robust estimators only extract one dominant structure in the data set, it has to be employed multiple times to iteratively extract all the surfaces from the range image. Research also shows that under certain circumstances, none of the current robust estimators could correctly extract the surface model from the data [9].

Here we propose a new segmentation algorithm that takes advantages of the robust data fusion technique. Given a range image, we randomly select n points from that image ($n = 1000$ in this experiment), and use the multiple regression method introduced in the previous example to extract all the surfaces that are related to the sampled points. Unlike the robust estimator based methods, our sampling and surface extraction processes are only run a small number of iterative steps (usually three or four) to obtain all the surface models. After the surfaces related to the sampled points are extracted in each iterative step, we remove all the points in the image that belong to those surfaces. When all the surfaces are extracted, some isolated outliers may remain. Also there may exist the problem of virtual intersection. This is because the true extent of the surface being extracted is not always known, and points of other surfaces, which are intercepted by the unbounded surface being considered, may be incorrectly assigned to that surface. At this stage, a post processing procedure, like the connected component algorithm [4], is made to deal with those problems.

Fig. 4 illustrates a fusion based range image segmentation example. In order to show the advantages of the proposed method, we add 26 200 random noise points to the range image taken from the USF ABW range image database (test 22), as shown in Fig. 4(a). The sampling and fusion are employed three times for this image to extract all the planar primitives. In the first iterative step, the two background planes are estimated. In the following iterative steps, all the six planes related to the

object in the scene are extracted. Fig. 4(b) shows the sampled data in the second iterative step. The associated confidence regions related to the estimated parameters in the multiple regression stage are shown in Fig. 4(c). The robust fusion is conducted on the parameter space, and the recovered information sources are related to the surface models. As we can see in Fig. 4(d), the algorithm extracted all the surfaces correctly. The accuracy of the surface parameters ensures that the location and boundary of the planes are aligned with those in the original image [Fig. 4(a)].

VII. Discussion

Describing the proposed fusion technique from a Bayesian viewpoint provides additional insight. To obtain the expression (8) for a single source, we must assume that the data points are independent, have conditional normal distributions $\mathbf{x}_i \sim NI(\hat{\mathbf{x}}, \mathbf{C}_i)$, and that no *a priori* information is available about $\hat{\mathbf{x}}$. In this case, both the maximum *a posteriori* (MAP) and the maximum likelihood estimates of $\hat{\mathbf{x}}$ yield (8).

For multiple sources, a Bayesian model is more complicated. We have to introduce the (unknown) probability of a data point being an inlier or an outlier. The representation of the sources, $\hat{\mathbf{x}}$, becomes a discrete random variable taking one of N unknown values, where the number N also has to be determined. There is no information about the prior probability of these sources.

The difficulties of such parametric model are avoided in our method by seeking the modes of the distribution underlying the data with the nonparametric mean shift procedure. At each iteration of (22) the mean shift vector provides the update of a MAP estimate computed only from those points whose region of confidence contains the current estimate. Any data distribution can be handled, as long as there are enough data points to support the mean shift procedure.

In the machine learning literature the variable-kernel similarity (VSM) metric method [25] is close in spirit to what was proposed in this paper. There, to classify a new measurement, the most probable class is determined by conditional density estimation. A Gaussian kernel is centered on each training data point in a neighborhood, and the probability of the test point belonging to a class is computed with a kernel density estimator. The kernel parameters are determined by an optimization procedure. In our method the basin of attraction of each mode provides a similar way to classify new data, but the kernel parameters are replaced by the information about the measurement uncertainty. This could yield a more accurate classification.

Integration of the proposed information fusion method with resampling based machine learning techniques, such as bagging [12], can yield a new class of robust bootstrap-type learning techniques. In this case, each data point is obtained from a bootstrap sample, i.e., a bootstrap estimate, and the aggregation process is now based on the detected modes and their basins of attraction. Since the covariances associated with the bootstrap samples provide a measure of confidence in the sample, our mode delineation automatically takes into account the relevance of the sample during fusion, with the outlying samples being discarded. Note that by using the modes, the two traditional learning strategies, averaging and voting, are combined together. Indeed, the mode is a local maximum of the *a posteriori* distribution, while its position is computed as a weighted average.

VIII. Conclusion

A new technique for the robust fusion of information was presented in the paper. Given the measurements with their uncertainty described by the covariance matrices, we find the characteristics of an *unknown* number of information sources by minimizing the sum of Mahalanobis distances from the measurements to those sources. Multivariate Epanechnikov kernels are used to compute the sample point density estimate of the measurements. Adaptive mean shift is employed to locate the modes of such multivariate density distribution, which are related to the information sources. Two computer vision applications, multiple affine transforms estimation and range image segmentation, showed the effectiveness of this technique.

References

[1] R. N. Dave and K. J. Patel, "Progressive fuzzy clustering algorithms for characteristic shape recognition," in *Proc. N. Amer. Fuzzy Information Processing Soc.*, 1990, pp. 121–124.

[2] K.-M. Lee, P. Meer, and R.-H. Park, "Robust adaptive segmentation of range images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 2, pp. 200–205, Feb. 1998.

[3] C. V. Stewart, "MINPRAN: a new robust estimator for computer vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 10, pp. 925–938, Oct. 1998.

[4] R. Lumia, L. Shapiro, and O. Zungia, "A new connected components algorithm for virtual memory computer," *Comput. Vis., Graph. Image Process.*, vol. 22, pp. 287–300, 1983.

[5] K. Koster and M. Spann, "MIR: An approach to robust clustering—Applications to range image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 5, pp. 430–444, May 2000.

[6] M. A. Wani and B. G. Batchelor, "Edge region based segmentation of range images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 3, pp. 314–319, Mar. 1994.

[7] R. L. Hoffman and A. K. Jain, "Segmentation and classification of range images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, no. 5, pp. 608–620, Sep. 1987.

[8] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. I, Madison, WI, Jun. 2003, pp. 621–628.

[9] H. Chen, P. Meer, and D. E. Tyler, "Robust regression for data with multiple structures," in *Proc. IEEE Con. Computer Vision and Pattern Recognition*, vol. I, Kauai, HI, Dec. 2001, pp. 1069–1075.

[10] J. Costeira and T. Kanade, "A multibody factorization method for motion analysis," in *Proc. 5th Int. Conf. Computer Vision*, vol. I, Vancouver, BC, Canada, Jun. 1995, pp. 1071–1078.

[11] T. Y. Tian and M. Shah, "Recovering 3-D motion of multiple objects using adaptive Hough transform," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 10, pp. 1178–1183, Oct. 1997.

[12] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[13] H. Chen and P. Meer, "Robust computer vision through kernel density estimation," in *Proc. 7th Eur. Conf. Computer Vision*, vol. I, Copenhagen, Denmark, May 2002, pp. 236–250.

[14] E. Choi and P. Hall, "Data sharpening as a prelude to density estimation," *Biometrika*, vol. 86, pp. 941–947, 1999.

[15] D. Comaniciu and P. Meer, "Distribution free decomposition of multivariate data," *Pattern Anal. Applicat.*, vol. 2, pp. 22–30, 1999.

[16] ——, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[17] D. Comaniciu, V. Ramesh, and P. Meer, "The variable bandwidth mean shift and data-driven scale selection," in *Proc. 8th Int. Conf. Computer Vision*, vol. I, Vancouver, BC, Canada, Jul. 2001, pp. 438–445.

[18] T. G. Dietterich, "Machine learning research: Four current directions," *AI Mag.*, vol. 18, no. 4, pp. 97–136, 1997.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.

[20] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, 1997.

[21] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inform. Theory*, vol. 21, no. 1, pp. 32–40, Jan. 1975.

[22] P. Hall, T. Hui, and J. Marron, "Improved variable window kernel estimates of probability densities," *Ann. Statist.*, vol. 23, pp. 1–10, 1995.

[23] E.-Y. Kang, I. Cohen, and G. Medioni, "Robust affine motion estimation in joint image space using tensor voting," in *Proc. 16th Int. Conf. Computer Vision and Pattern Recog.*, vol. IV, Quebec City, QC, Canada, Aug. 2002, pp. 256–259.

[24] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[25] D. G. Lowe, "Similarity metric learning for a variable-kernel classifier," *Neural Comput.*, vol. 7, pp. 72–85, 1995.

[26] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.

[27] *Proc. IEEE, Special Issue on Data Fusion*, vol. 85, no. 1, Jan. 1997.

[28] *Pattern Recognit., Special Issue on Data and Information Fusion in Image Processing and Computer Vision*, vol. 34, no. 8, Aug. 2001.

[29] D. M. J. Tax, M. van Breukelen, R. P. W. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?," *Pattern Recognit.*, vol. 33, pp. 1475–1486, 2000.

[30] L. Valet, G. Mauris, and P. Bolon, "A statistical overview of recent literature in information fusion," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 16, pp. 7–14, 2001.

[31] L. Wald, "Some terms of reference in data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1190–1193, May 1999.

[32] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London, U.K.: Chapman & Hall, 1995.