A New Distance for Scale-Invariant 3D Shape Recognition and Registration

Minh-Tri Pham[1] Oliver J. Woodford[1] Frank Perbet[1] Atsuto Maki[1] Björn Stenger[1] Roberto Cipolla[2]

[1]Cambridge Research Laboratory, Toshiba Research Europe Ltd, Cambridge, UK

[2]Department of Engineering, University of Cambridge, Cambridge, UK

{minhtri.pham,oliver.woodford,frank.perbet,atsuto.maki,bjorn.stenger}@crl.toshiba.co.uk cipolla@eng.cam.ac.uk

Abstract

This paper presents a method for vote-based 3D shape recognition and registration, in particular using mean shift on 3D pose votes in the space of direct similarity transforms for the first time. We introduce a new distance between poses in this space—the SRT distance. It is left-invariant, unlike Euclidean distance, and has a unique, closed-form mean, in contrast to Riemannian distance, so is fast to compute. We demonstrate improved performance over the state of the art in both recognition and registration on a real and challenging dataset, by comparing our distance with others in a mean shift framework, as well as with the commonly used Hough voting approach.

1. Introduction

This paper concerns itself with vote-based pose estimation techniques. These arise in many vision tasks including 2D object detection [13, 17, 22], motion segmentation [24, 25], and 3D shape registration and recognition [9, 12, 26]. These methods all share a common two stage framework: First they generate an empirical distribution of pose through the collation of a set of possible poses, or *votes*. The votes are often computed by matching local features from a test object to those in a library with known pose [9, 12, 13, 17, 22, 24, 25, 26]. The second step is then to find one or more "best" poses in the distribution (the maxima, in the case of ML/MAP estimation). This curation of data prior to inference makes such vote-based approaches more efficient and robust than competing techniques, *e.g.* global or appearance-based methods [16]. Here we focus on the latter, inference step, and assume the votes are given.

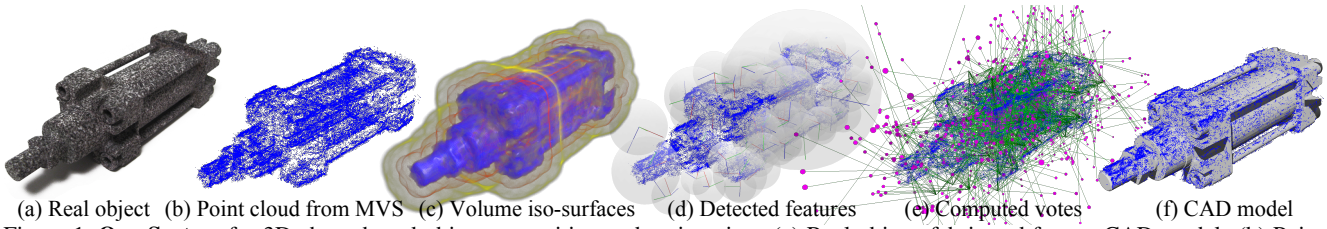
Two methods for finding the maxima are Hough voting and mean shift. In Hough voting the probability is computed on a regular grid over the pose parameter space. This discretization leads to loss of accuracy, as well as a complexity exponential in the pose dimensionality, but ensures coverage of the entire space. Mean shift [7] iteratively finds local maxima of probability, resulting in initialization issues but also high accuracy. The complexity of an iteration is usually linear in the pose dimensionality. The two methods are therefore somewhat complementary; indeed they are often used together [13, 17].

While Hough voting can easily be applied to any space (in our case that of all poses), this is not straightforward for mean shift; each iteration requires the computation of a weighted average of input votes, formulated as a least squares minimization of distances from input votes to the mean. In Euclidean space this minimization yields a unique, closed-form solution—the arithmetic mean. When poses lie on a non-linear manifold this mean is typically outside the manifold, requiring a projection onto it. A more direct approach is to minimize the geodesic arclengths over the manifold, known as the Riemannian distance.

In this paper we focus on 3D shape recognition and registration, as part of a system (see figure 1) for recognizing industrial parts. However, unlike existing approaches, where objects of interest are of either fixed (or omitted) scale [25] or rotation [13, 17, 22], here we recognize and register objects in the direct similarity group: the group of isotropic similarity transformations parameterized by translation, rotation *and* scale. Scale is necessary when the input data's scale is unknown, or when there is high intra-class scale variation. Rotation is necessary for full registration, leading to more accurate recognition. The resulting 7D pose space is currently too large to apply Hough voting to in practice [11]. Here we use mean shift, for which scale and rotation also introduce problems using existing distances: Euclidean distance is scale variant, and the induced mean of poses has a bias in scale. The mean of poses using Riemannian distance has no closed-form solution even when the poses are rotations [15], and is slow to compute [23].

The contribution of this work is to introduce a new distance on the direct similarity group. The distance provides scale, rotation and translation-invariance concomitantly. The weighted mean of this distance is unique, closed-form, and fast to compute, as well as having several key properties discussed in §2.2.3. We demonstrate this distance's performance in mean shift, in the context of our 3D shape registration and recognition system, comparing it with other distances on the same space, as well as a Hough voting method.

The paper is laid out as follows: The next section reviews



(a) Real object (b) Point cloud from MVS (c) Volume iso-surfaces (d) Detected features (e) Computed votes (f) CAD model

Figure 1. **Our System** for 3D-shape-based object recognition and registration. (a) Real object, fabricated from a CAD model. (b) Point cloud extracted using a multi-view stereo (MVS) system. (c) Iso-surfaces of the scalar volume computed from the points. (d) Features (with full scale, rotation and translation pose) detected in the volume. (e) Votes for the object centre, based on detected features matched with a library of learnt features. (f) The registered CAD model.

	$\mathbf{t}$	$\mathbf{t}$, s	$\mathbf{t}$, $\mathbf{R}$	$\mathbf{t}$, $\mathbf{R}$, s
Hough	[26]	[12]	[10]	[11]
Mean shift	–	[13, 17, 22]*	[25]	[24]*, **This work**

Table 1. **Methods of pose estimation** over different transformations. $\mathbf{t}$: translation; $\mathbf{R}$: rotation; s: scale. *Indicates 2D space.

the literature relevant to 3D shape recognition and registration inference, as well as means in Lie groups. In the following section we introduce our new distance on the direct similarity group, and its associated mean. In the final two sections we present our experiments, before concluding.

2. Background

We now briefly review the inference techniques used for vote-based pose estimation, and take a closer look at mean shift applied to this task.

2.1. Vote-based pose estimation

The inference step of vote-based pose estimation involves the computation of maxima in the empirical distribution of poses defined by a set of input votes. Two main techniques for this are Hough voting (an extension of the Generalized Hough Transform [4]) and mean shift [7]. Table 1 lists a representative subset of works that use each of these methods, and under which transformations.

Using Hough voting, Khoshelham [11] quantizes the 7D space of 3D translation, rotation and scale for object registration. This creates a trade-off between pose accuracy and computational requirements, the latter proving to be costly. Other methods seek to reduce this complexity by shrinking the pose space and marginalizing over some parameters. Fisher *et al*. [10] quantize translations and rotations in two separate 3D arrays; peak entries in both arrays indicate the pose of the object, but multiple objects create ambiguities. Knopp *et al*. [12] show effective object recognition using Hough voting over 3D translation and scale. Tombari & Di Stefano [26] first compute Hough votes over translation, assuming known scale in their 3D object recognition and registration application, then determine rotation by averaging the rotations at each mode. Geometric hashing [9, 14] is a similar technique to Hough voting which reparameterizes pose in a lower dimensional space before clustering. However, all these dimensionality reduction techniques lead to an increased chance of false positive detections.

Mean shift avoids the trade-off suffered by Hough voting methods, being both accurate and having lower (usually[1] linear) complexity in the pose dimensionality, making it suitable for inference in the full 7D pose space of the direct similarity group in 3D. To date it has been used in 2D applications: object detection over translation and scale [13, 17, 22], and motion segmentation over affine transformations [25], as well as in 3D for motion segmentation over translation and rotation [24]. This is the first paper we know of to apply mean shift to a 3D application using translation, rotation and scale simultaneously. A reason for this could be the problems associated with computing means using Euclidean and Riemannian distances in this space, which we discuss below.

2.2. Mean shift

The mean shift algorithm [7] (Algorithm 1) is a nonparametric kernel density estimator that finds the local modes of a density function by coordinate ascent. Given a distance function $d(\cdot,\cdot)$ on the input space, the kernel density estimate is defined as

$$\hat{f}_K(\mathbf{X}) = \sum_{i=1}^{N} \frac{1}{\zeta} \lambda_i K(d^2(\mathbf{X}, \mathbf{X}_i)), \tag{1}$$

where $\mathbf{X}$ is the random variable, $\mathcal{X} = \{\mathbf{X}_i, \lambda_i\}_{i=1}^N$ is a set of input points with weights $\lambda_i \geq 0$, $K(\cdot) \geq 0$ is a kernel function, and ζ is a volume density function which normalizes $K(d^2(\cdot, \mathbf{X}_i))$. The most common (and our) choice for $K(\cdot)$ is the Gaussian kernel, $\exp\left(-\frac{\cdot}{2\sigma^2}\right)$, where σ is the bandwidth of the kernel. On Euclidean spaces a natural choice for $d()$ is the Euclidean distance, $d_{\mathrm{E}}()$; *e.g*. if $\mathbf{X}$ and $\mathbf{Y}$ are matrices, $d_{\mathrm{E}}(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}$ where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Under the Euclidean distance, the solution of step 5 in Algorithm 1,

$$\mu(\mathcal{X}) = \operatorname*{argmin}_{\mathbf{X}} \sum_i w_i d^2(\mathbf{X}, \mathbf{X}_i), \tag{2}$$

where the w_i are weights, is the arithmetic mean, *i.e*. $\mu(\mathcal{X}) = \frac{\sum_i w_i \mathbf{X}_i}{\sum_i w_i}$.

In pose estimation, votes are represented by linear transformations which form a matrix Lie group. This paper is concerned with the direct similarity group $S^+(n) \subset$

[1] Certain distance computations are not linear, *e.g*. that of §2.2.2.

Algorithm 1 Mean shift [7] (for notation see text)

Require: $\mathcal{X} = \{\mathbf{X}_i, \lambda_i\}_{i=1}^N$, distance function $d(\cdot,\cdot)$

1: Initialize $\mathbf{X}$
2: **repeat**
3: $\quad \mathbf{X}_{\text{old}} := \mathbf{X}$
4: $\quad w_i := \lambda_i K(d^2(\mathbf{X}, \mathbf{X}_i)) \qquad \forall i = 1, .., N$
5: $\quad \mathbf{X} := \text{argmin}_{\mathbf{X}} \sum_i w_i d^2(\mathbf{X}, \mathbf{X}_i)$
6: **until** $d(\mathbf{X}_{\text{old}}, \mathbf{X}) < \epsilon$
7: **return** $\mathbf{X}$

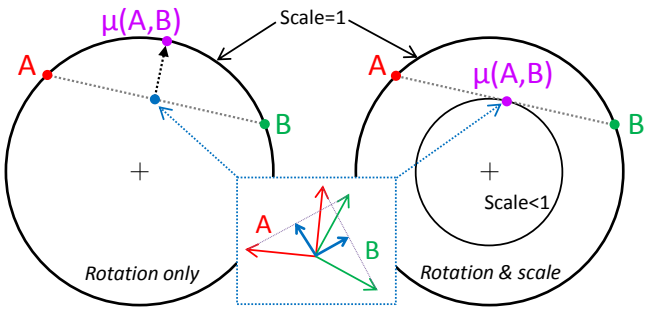


Figure 2. **Scale bias of the extrinsic mean.** Let us consider $S^+(2)$ (without translation): on a plane, a rotation can be represented as a point on a circle, the radius being the scale. *Left*: with rotation only, the arithmetic mean of $\mathbf{A}$ and $\mathbf{B}$ leads to a smaller scale but the reprojection onto the manifold (*i.e.* the unit circle) gives a reasonable result. *Right*: with rotation and scale, the mean is already on the manifold, but with a smaller scale.

$GL(n+1, \mathbb{R})$, which is the set of all affine transformation matrices acting on $\mathbb{R}^n$ preserving angles and orientations [21]. Using homogeneous coordinates, every matrix $\mathbf{X} \in S^+(n)$ is represented as follows:

$$\mathbf{X} = \begin{bmatrix} s(\mathbf{X})\mathbf{R}(\mathbf{X}) & \mathbf{t}(\mathbf{X}) \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix}, \tag{3}$$

where $s(\mathbf{X}) \in \mathbb{R}^+$, $\mathbf{R}(\mathbf{X}) \in SO(n, \mathbb{R})$, and $\mathbf{t}(\mathbf{X}) \in \mathbb{R}^n$ are the scale, rotation, and translation components of $\mathbf{X}$.

When applying mean shift on a matrix Lie group, the choice of $d()$ is crucial since it affects both the computation of weights and the mean (steps 4 & 5 of Algorithm 1). Two well-known distances arise in the literature: Euclidean and Riemannian. We now review how existing methods utilize these distances in mean shift on matrix Lie groups.

2.2.1 Euclidean distance

Given a matrix Lie group $\mathcal{G} \subset GL(n, \mathbb{R})$, since $GL(n, \mathbb{R}) \subset \mathbb{R}^{n^2}$ (up to an isomorphism), the most straightforward way to apply mean shift on $\mathcal{G}$ is to embed $\mathcal{G}$ in the Euclidean space $\mathbb{R}^{n^2}$ and run mean shift on this instead. However, at each iteration the arithmetic mean may not lie in $\mathcal{G}$. It is therefore projected back to $\mathcal{G}$ via the mapping:

$$\pi : \mathbb{R}^{n^2} \to \mathcal{G} : \pi(\mathbf{X}) = \underset{\mathbf{Y} \in \mathcal{G}}{\text{argmin}} \, \|\mathbf{Y} - \mathbf{X}\|_{\text{F}}^2 . \tag{4}$$

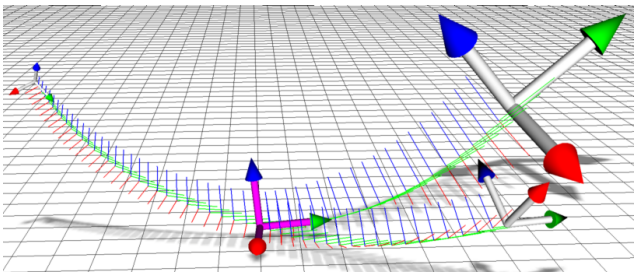


Figure 3. **The intrinsic mean.** Three poses in $S^+(3)$ (with different scales, rotations and translations) and their intrinsic mean (pink). The geodesics between the mean and input poses are also drawn. Note that the shortest distance between two transformations is not necessarily a straight line in terms of translation.

The projected arithmetic mean, $\mu(\mathcal{X}) = \pi\left(\frac{\sum_i w_i \mathbf{X}_i}{\sum_i w_i}\right)$, is referred to in the literature as the *extrinsic* mean [15, 23].

Mean shift using Euclidean distance (extrinsic mean shift) has shown good results on Stiefel and Grassmann manifolds [6]. However, there are two drawbacks with extrinsic mean shift applied to $S^+(n)$. First, $d_{\text{E}}()$ is invariant to rotation and translation but not to scaling, making the weights, w_i, computed by mean shift scale variant. Thus, although the extrinsic mean is scale-covariant[2], extrinsic mean shift is *not*. Second, the extrinsic mean of rotation and scale transformations causes a bias towards smaller scales, as illustrated in figure 2.

2.2.2 Riemannian distance

An alternative choice for $d()$ is the Riemannian distance. $d_{\text{R}}()$. Given $\mathbf{X}, \mathbf{Y} \in \mathcal{G}$, $d_{\text{R}}(\mathbf{X}, \mathbf{Y})$ is defined as the arclength of the geodesic between $\mathbf{X}$ and $\mathbf{Y}$ (see figure 3). Since $d_{\text{R}}()$ depends only on the *intrinsic* geometry of $\mathcal{G}$, the mean defined as the solution of equation (2) using $d_{\text{R}}()$ is called the intrinsic mean [15, 19]. Efficient formulations of $d_{\text{R}}()$ exist for some $\mathcal{G}$, notably $SE(3)$ [2], which can be adapted to $S^+(3)$. However, in $S^+(n)$ for $n > 3$, $d_{\text{R}}()$ generally has no closed-form formulation, taking $O(n^4)$ time to compute [8].

Intrinsic mean shift methods have been proposed [6, 25]. The intrinsic mean itself has multiple non-closed-form solutions [15]; in our experiments we compute an approximation using a single step[3] of the iterative method of [25].

2.2.3 Properties of a good distance in $S^+(n)$

In the context of mean shift, and subsequent to our overview of Euclidean and Riemannian distances, we propose the following list of desirable properties for a distance in $S^+(n)$ and its associated mean:

[2] Scale-covariant means a scale transformation of input data produces the same transformation on the output.

[3] This is equivalent to computing a mean using the *log-Euclidean* distance [3], $d(\mathbf{X}, \mathbf{Y}) = \|\text{logm}(\mathbf{X}) - \text{logm}(\mathbf{Y})\|_{\text{F}}$.

1. *Unique:* The mean should have a unique solution.

2. *Closed-form:* For efficient computation, the mean should have a closed-form solution.

3. *Scale-compatible:* If all rotations and translations are equal, the mean should behave as an average of the scales. Mathematically, if $\forall \mathbf{X}_i \in \mathcal{X} : \mathbf{R}(\mathbf{X}_i) = \mathbf{R}', \mathbf{t}(\mathbf{X}_i) = \mathbf{t}'$ for some $\mathbf{R}'$ and $\mathbf{t}'$, then we would like $\mathbf{R}(\mu(\mathcal{X})) = \mathbf{R}'$, $\mathbf{t}(\mu(\mathcal{X})) = \mathbf{t}'$, and $s(\mu(\mathcal{X}))$ to be an average of $s(\mathbf{X}_i)$'s. In this case, we say that μ is scale-compatible.

4. *Rotation-compatible:* If $\forall \mathbf{X}_i \in \mathcal{X} : s(\mathbf{X}_i) = s'$, $\mathbf{t}(\mathbf{X}_i) = \mathbf{t}'$, then $s(\mu(\mathcal{X})) = s'$, $\mathbf{t}(\mu(\mathcal{X})) = \mathbf{t}'$ and $\mathbf{R}(\mu(\mathcal{X}))$ is an average of $\mathbf{R}(\mathbf{X}_i)$'s.

5. *Translation-compatible:* If $\forall \mathbf{X}_i \in \mathcal{X} : s(\mathbf{X}_i) = s'$, $\mathbf{R}(\mathbf{X}_i) = \mathbf{R}'$, then $s(\mu(\mathcal{X})) = s'$, $\mathbf{R}(\mu(\mathcal{X})) = \mathbf{R}'$ and $\mathbf{t}(\mu(\mathcal{X}))$ is an average of $\mathbf{t}(\mathbf{X}_i)$'s.

6. *Left-invariant:* A left-invariant distance is one that is unchanged by any post-transformation, *i.e.* $d(\mathbf{ZX}, \mathbf{ZY}) = d(\mathbf{X}, \mathbf{Y})$ $\forall \mathbf{X}, \mathbf{Y}, \mathbf{Z} \in S^+(n)$. This property is crucial for two reasons: (a) it leads to a left-covariant mean: $\mu(\mathbf{Z}\mathcal{X}) = \mathbf{Z}\mu(\mathcal{X})$[4] *i.e.* if all poses $\mathbf{X}_i$ are transformed by $\mathbf{Z}$, the mean is transformed by $\mathbf{Z}$ as well, and (b) it ensures that the weights w_i computed in mean shift are invariant to any post-transformation $\mathbf{Z}$, leading to left-covariant mean shift.

A symmetric distance, *s.t.* $d(\mathbf{X}, \mathbf{Y}) = d(\mathbf{Y}, \mathbf{X})$ $\forall \mathbf{X}, \mathbf{Y} \in S^+(n)$, intuitively seems desirable, but its absence does not prevent a distance from being used in mean shift and furthermore, given the properties listed, it is not necessary. Right-invariance might also be considered a desirable property, but in the context of 3D recognition this occurrence does not relate to any meaningful behaviour.

3. The SRT distance and its mean

In this section, we describe our new distance on $S^+(n)$, which fulfills all the desirable properties defined in §2.2.3. We call it the SRT distance, with corresponding mean μ_{SRT}.

3.1. Distance definition

We first define the following component-wise distances:

$$d_s(\mathbf{X}, \mathbf{Y}) = \left| \log \left(\frac{s(\mathbf{X})}{s(\mathbf{Y})} \right) \right|, \quad (5)$$

$$d_r(\mathbf{X}, \mathbf{Y}) = \|\mathbf{R}(\mathbf{X}) - \mathbf{R}(\mathbf{Y})\|_{\text{F}}, \quad (6)$$

$$d_t(\mathbf{X}, \mathbf{Y}) = \frac{\|\mathbf{t}(\mathbf{X}) - \mathbf{t}(\mathbf{Y})\|}{s(\mathbf{Y})}, \quad (7)$$

[4] $\mathbf{Z}\mathcal{X} = \{\mathbf{ZX} : \mathbf{X} \in \mathcal{X}\}$ is a left coset of $\mathcal{X}$. Proof in [20, App. A.4].

Properties	Extrinsic	Intrinsic	SRT
Distance:			
Symmetric	✓	✓	✗
Left-invariant	✗	✓	✓
Mean:			
Unique	✓	✗†	✓
Closed-form	✓	✗	✓
Scale-compatible	✓	✓	✓
Rotation-compatible	✗	✓	✓
Translation-compatible	✓	✗†	✓

Table 2. **Properties** of distances and associated means in $S^+(n)$. †The approximation of [25] is, however, unique and translation compatible.

in which $d_s()$, $d_r()$ and $d_t()$ measure scale, rotation and translation distances respectively, with $\mathbf{X}$ and $\mathbf{Y}$ in $S^+(n)$. Given some bandwidth coefficients $\sigma_s, \sigma_r, \sigma_t > 0$, the SRT distance is defined as:

$$d_{\text{SRT}}(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{d_s^2(\mathbf{X}, \mathbf{Y})}{\sigma_s^2} + \frac{d_r^2(\mathbf{X}, \mathbf{Y})}{\sigma_r^2} + \frac{d_t^2(\mathbf{X}, \mathbf{Y})}{\sigma_t^2}}. \quad (8)$$

By controlling $\sigma_s, \sigma_r, \sigma_t$, it is possible to create an SRT distance that is more sensitive to one type of transformations among scale, rotation, and translation than the others. In this sense, the SRT distance is more flexible than the Euclidean and Riemannian distances.

Theorem 1. *$d_{SRT}()$ is left-invariant.*

Proof. $d_{\text{SRT}}()$ is related to a pseudo-seminorm on $S^+(n)$, *i.e.* $d_{\text{SRT}}(\mathbf{X}, \mathbf{Y}) = \left\|\mathbf{Y}^{-1}\mathbf{X}\right\|_{\text{SRT}}$, where

$$\|\cdot\|_{\text{SRT}} = \sqrt{\frac{\log^2(s(\cdot))}{\sigma_s^2} + \frac{\|\mathbf{R}(\cdot) - \mathbf{I}\|_{\text{F}}^2}{\sigma_r^2} + \frac{\|\mathbf{t}(\cdot)\|^2}{\sigma_t^2}}. \quad (9)$$

It follows that $d_{\text{SRT}}()$ is left invariant: $d_{\text{SRT}}(\mathbf{X}, \mathbf{Y}) = \left\|\mathbf{Y}^{-1}\mathbf{X}\right\|_{\text{SRT}} = \left\|\left(\mathbf{X}_i^{-1}\mathbf{Z}^{-1}\right)(\mathbf{ZX})\right\|_{\text{SRT}} = d_{\text{SRT}}(\mathbf{ZX}, \mathbf{ZY})$. □

Note that, unlike $d_{\text{E}}()$ and $d_{\text{R}}()$, $d_{\text{SRT}}()$ is not symmetric; it could be made symmetric by a slight modification of the translation component, but at the expense of the translation-compatibility of the corresponding mean.

3.2. Mean computation

Having defined $d_{\text{SRT}}()$, we now derive the mean μ_{SRT} induced by $d_{\text{SRT}}()$ using equation (2), which is:

$$\mu_{\text{SRT}}(\mathcal{X}) = \underset{\mathbf{X} \in S^+(n)}{\operatorname{argmin}} \sum_i w_i d_{\text{SRT}}^2(\mathbf{X}, \mathbf{X}_i). \quad (10)$$

and show that it is closed-form and generally unique.

Theorem 2. *The solution of equation (10) is given as:*

$$s\left(\mu_{SRT}(\mathcal{X})\right) = \exp\left(\frac{\sum_i w_i \log s(\mathbf{X}_i)}{\sum_i w_i}\right), \quad (11)$$

$$\mathbf{R}\left(\mu_{SRT}(\mathcal{X})\right) = \operatorname{sop}\left(\frac{\sum_i w_i \mathbf{R}(\mathbf{X}_i)}{\sum_i w_i}\right), \quad (12)$$

$$\mathbf{t}\left(\mu_{SRT}(\mathcal{X})\right) = \sum_i \frac{w_i \mathbf{t}(\mathbf{X}_i)}{s^2(\mathbf{X}_i)} \Big/ \sum_i \frac{w_i}{s^2(\mathbf{X}_i)} \quad (13)$$

where $\operatorname{sop}(\mathbf{X}) = \operatorname{argmin}_{\mathbf{Y}\in SO(n,\mathbb{R})} \|\mathbf{Y}-\mathbf{X}\|_{\mathrm{F}}$ *is the orthogonal projection of matrix* $\mathbf{X}$ *onto* $SO(n,\mathbb{R})$.

Proof. The sum in equation (10) can be rewritten as

$$\sum_i w_i d^2_{\mathrm{SRT}}(\mathbf{X},\mathbf{X}_i) = \frac{F_s(\mathbf{X})}{\sigma_s^2} + \frac{F_r(\mathbf{X})}{\sigma_r^2} + \frac{F_t(\mathbf{X})}{\sigma_t^2}, \quad (14)$$

where[5] $F_\star(\mathbf{X}) = \sum_{i=1}^N w_i d^2_\star(\mathbf{X},\mathbf{X}_i)$. Since $s(\mathbf{X})$ only appears in $F_s(\mathbf{X})$, we can reformulate

$$s\left(\mu_{\mathrm{SRT}}(\mathcal{X})\right) = \operatorname*{argmin}_{s\in\mathbb{R}^+} \sum_i w_i \log^2\left(\frac{s(\mathbf{X})}{s(\mathbf{X}_i)}\right), \quad (15)$$

yielding the solution (11). Similarly, since $\mathbf{t}(\mathbf{X})$ only appears in $F_r(\mathbf{X})$, after rewriting

$$\mathbf{t}\left(\mu_{\mathrm{SRT}}(\mathcal{X})\right) = \operatorname*{argmin}_{\mathbf{t}\in\mathbb{R}^n} \sum_i w_i \frac{\|\mathbf{t}(\mathbf{X})-\mathbf{t}(\mathbf{X}_i)\|^2}{s^2(\mathbf{X}_i)}, \quad (16)$$

we get equation (13). Finally, since $\mathbf{R}(\mathbf{X})$ only appears in $F_r(\mathbf{X})$, we rewrite

$$\mathbf{R}\left(\mu_{\mathrm{SRT}}(\mathcal{X})\right) = \operatorname*{argmin}_{\mathbf{R}\in SO(n,\mathbb{R})} \sum_i w_i \|\mathbf{R}(\mathbf{X})-\mathbf{R}(\mathbf{X}_i)\|^2_{\mathrm{F}}. \quad (17)$$

This is precisely the definition of the Euclidean (extrinsic) mean of rotation matrices in [15], except that $SO(3,\mathbb{R})$ is generalized to $SO(n,\mathbb{R})$, and with the inclusion of weights w_i. The uniqueness and closed-form solution of equation (17) in the case of $n=3$ is given in [15, §3.1], in which it is shown that $\mathbf{R}\left(\mu_{\mathrm{SRT}}(\mathcal{X})\right) = \operatorname{sop}\left(\frac{1}{N}\sum_i \mathbf{R}(\mathbf{X}_i)\right)$. □

It can be further verified that $\mu_{\mathrm{SRT}}(\mathcal{X})$ is:

Scale-compatible: If $\forall \mathbf{X}_i \in \mathcal{X} : \mathbf{R}(\mathbf{X}_i) = \mathbf{R}'$, $\mathbf{t}(\mathbf{X}_i) = \mathbf{t}'$ then $\mathbf{R}(\mu_{\mathrm{SRT}}(\mathcal{X})) = \mathbf{R}'$, $\mathbf{t}(\mu_{\mathrm{SRT}}(\mathcal{X})) = \mathbf{t}'$ and $s(\mu_{\mathrm{SRT}}(\mathcal{X}))$ is the geometric mean of $s(\mathbf{X}_i)$'s.

Rotation-compatible: If $\forall \mathbf{X}_i \in \mathcal{X} : s(\mathbf{X}_i) = s'$, $\mathbf{t}(\mathbf{X}_i) = \mathbf{t}'$ then $s(\mu_{\mathrm{SRT}}(\mathcal{X})) = s'$, $\mathbf{t}(\mu_{\mathrm{SRT}}(\mathcal{X})) = \mathbf{t}'$ and $\mathbf{R}(\mu_{\mathrm{SRT}}(\mathcal{X}))$ is the extrinsic mean of $\mathbf{R}(\mathbf{X}_i)$'s.

Translation-compatible: If $\forall \mathbf{X}_i \in \mathcal{X} : s(\mathbf{X}_i) = s'$, $\mathbf{R}(\mathbf{X}_i) = \mathbf{R}'$ then $s(\mu_{\mathrm{SRT}}(\mathcal{X})) = s'$, $\mathbf{R}(\mu_{\mathrm{SRT}}(\mathcal{X})) = \mathbf{R}'$, and $\mathbf{t}(\mu_{\mathrm{SRT}}(\mathcal{X}))$ is the arithmetic mean of $\mathbf{t}(\mathbf{X}_i)$'s.

Table 2 summarizes the desirable properties of the SRT distance and mean, and contrasts them with those of the Euclidean and Riemannian distances.

[5] $\star$ should be replaced with s, r or t.

3.3. SRT mean shift

We form our mean shift algorithm on $S^+(n)$ using $d_{\mathrm{SRT}}()$ and $\mu_{\mathrm{SRT}}(\mathcal{X})$ in steps 4 & 5 of Algorithm 1 respectively. It follows from the left-invariance of d_{SRT} that SRT mean shift is left-covariant.

The coefficients $\sigma_s, \sigma_t, \sigma_r$ act in place of the kernel bandwidth σ in equation (1). Also note that, while the coefficient ζ is constant in Euclidean space, it is *not* constant in a non-Euclidean space, in which case $\zeta = \zeta(\mathbf{X}_i)$ [18, 25] cannot be factored out of the kernel density estimate. Since $\zeta(\mathbf{X}_i)$ can be costly to compute (sometimes non-closed-form), existing mean shift algorithms on Lie groups [6, 25] replace $\zeta(\mathbf{X}_i)$ with a constant. However, in the case of $d_{\mathrm{SRT}}()$, indeed any left-invariant distance, it can be shown that $\zeta(\mathbf{X}_i)$ is constant:

Lemma 3. *Using* d_{SRT}*, the volume densities are constant:* $\forall \mathbf{X},\mathbf{Y} \in S^+(n) : \zeta(\mathbf{X}) = \zeta(\mathbf{Y})$.

Proof. Let $\mathbf{Z} = \mathbf{X}\mathbf{Y}^{-1}$. By definition, we have $\zeta(\mathbf{Y}) = \int_{S^+(n)} K\left(d^2_{\mathrm{SRT}}(\mathbf{U},\mathbf{Y})\right) d\nu(\mathbf{U})$ where $\nu(\mathbf{U})$ is a (left-)Haar measure on $S^+(n)$. If we assume that $K(\cdot)$ is integrable then $\zeta(\mathbf{Y})$ is a Haar integral [20]. Using the substitution $\mathbf{V} = \mathbf{Z}\mathbf{U}$, we have $\zeta(\mathbf{Y}) = \int K\left(d^2_{\mathrm{SRT}}(\mathbf{Z}^{-1}\mathbf{V},\mathbf{Y})\right) d\nu(\mathbf{V}) = \int K\left(d^2_{\mathrm{SRT}}(\mathbf{Z}\mathbf{Z}^{-1}\mathbf{V},\mathbf{Z}\mathbf{Y})\right) d\nu(\mathbf{V}) = \zeta(\mathbf{X})$. □

4. Experiments

4.1. Experimental setup

Our experimental data consists of 12 shape classes, for which we have both a physical object and matching CAD model. We captured the geometry of each object[6] in the form of point clouds (figure 1(b)), 20 times from a variety of angles. Along with the class label, every shape instance has an associated ground truth pose, computed by first approximately registering the relevant CAD model to the point cloud manually, then using the Iterative Closest Point algorithm [5] to refine the registration.

4.1.1 Pose vote computation

Given a test point cloud and set of training point clouds (with known class and pose), the computation of input pose votes $\mathcal{X}$ is a two stage process[7] similar to [12, 26]. In the first stage, local shape features, consisting of a descriptor and a scale, translation and rotation relative to the object, are computed on all the point clouds (figure 1(c)). This is done by first converting a point cloud to a 128^3 voxel volume (figure 1(d)) using a Gaussian on the distance of each voxel centre to the nearest point. Then interest points are

[6] We used an implementation of [28], kindly provided by the authors.

[7] We keep the description of this process short as it is not central to the evaluation of relative performance, since all inference methods use the same set of input pose votes.

localized in the volume across 3D location and scale using the Difference of Gaussians operator, and a canonical orientation for each interest point computed [27], to generate a local feature pose. Finally a basic, 31-dimensional descriptor is computed by simply sampling the volume (at the correct scale) at 31 regularly distributed locations around the interest point.

In the second stage each test feature is matched to the 20 nearest training features, in terms of Euclidean distance between descriptors. Each of these matches generates a vote (figure 1(e)), $\mathbf{X}_i = \mathbf{A}\mathbf{B}^{-1}\mathbf{C}$, for the test object's pose, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ being the test feature, training feature and training object's ground truth pose respectively. In addition each vote has a weight, λ_i, computed as $(N_{\mathrm{C}} N_{\mathrm{I}})^{-1}$, N_{C} being the number of training instances in the class and N_{I} the number of features found in the feature's particular instance.

4.2. Inference

Mean shift Mean shift finds a local mode, and its weight, in the output pose distribution for a given object class. Since there may be many such modes we start mean shift from 100 random input poses for each class. Each mode, duplicates excepted, is then added to a list of candidate poses across all classes.

In $S^+(3)$ it is possible to use the quaternion representation of rotation, $\mathbf{q}(\mathbf{X})$, which we do. We therefore alternately define the rotation component of $d_{\mathrm{SRT}}()$ as

$$d_r(\mathbf{X}, \mathbf{Y}) = \sqrt{1 - |\mathbf{q}(\mathbf{X})^{\mathsf{T}} \mathbf{q}(\mathbf{Y})|}, \qquad (18)$$

where $|\cdot|$ is needed to account for the fact that $\mathbf{q}(\mathbf{X})$ and $-\mathbf{q}(\mathbf{X})$ represent the same rotation. This formulation confers a small computational advantage over other, non-component-wise distances in this space.

Hough voting We implemented a published Hough voting scheme [12] to compare with the mean shift inference approaches. This computes sums of weights of the pose votes which fall into each bin of a 4D histogram over translation and scale, effectively marginalizing over rotation. The bin widths are set to be 0.16 times the width (or scale) of the average shape in each of the 4 dimensions. The highest bin sum for each class defines a pose mode. Note that we used our own pose votes and weights, and not those computed using the method described in [12].

4.3. Evaluation

We use cross validation on our training data for evaluation—a training set is created from 19 of the 20 shape instances in each class, and the remaining instance in each class becomes a test shape. Each test shape undergoes 5 random transformations (over translation, rotation and scale in the range 0.5–2), and this process is repeated with each training shape being the test shape, creating 100 test instances per class. We use 10 classes in our evaluation (shown in figure 4), so 1000 tests in all. The remaining 2 classes are used to learn the optimal kernel bandwidth, σ, for each inference method. We have made the data used in this evaluation publicly available [1].

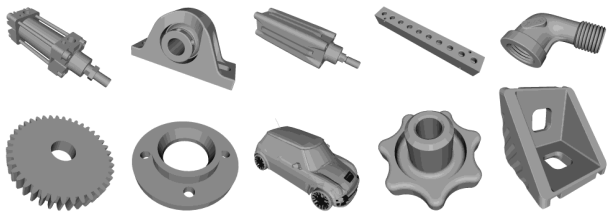


Figure 4. **Test objects.** CAD models of the 10 real objects used for evaluation. *Top*: piston2, bearing, piston1, block, and pipe. *Bottom*: cog, flange, car, knob, and bracket.

We evaluate each inference method on two criteria: Recognition rate and registration rate.

Recognition rate As described above, each inference method generates a list of modes across pose and class for a given test instance, each with an associated weight. The output class is that of the mode of highest weight. A confusion matrix logs the output class versus ground truth class across all tests. The recognition rate is given by the trace of this matrix, *i.e.* the number of correct classifications.

Registration rate The output pose for a given test instance is given by that of the weightiest mode whose class matches the ground truth class. We choose to consider a pose $\mathbf{X}$ to be correct if its scale is within 5%, orientation is within 15° and translation is within 10% (of the object size) of the ground truth's. Explicitly, the criteria to be met are

$$\left| \log \left(\frac{s(\mathbf{X})}{s(\mathbf{Y})} \right) \right| < 0.05, \qquad (19)$$

$$\operatorname{acos} \left(\frac{\operatorname{trace}(\mathbf{R}(\mathbf{X})^{-1} \mathbf{R}(\mathbf{Y})) - 1}{2} \right) < \pi/12, \qquad (20)$$

$$\frac{\|\mathbf{t}(\mathbf{X}) - \mathbf{t}(\mathbf{Y})\|}{\sqrt{s(\mathbf{X}) s(\mathbf{Y})}} < 0.1, \qquad (21)$$

with $\mathbf{Y}$ being the ground truth pose. In the case of an object having symmetries there are multiple $\mathbf{Y}$'s, and distance to the closest is used.

4.3.1 Learning σ

We learn the mean shift kernel bandwidth, σ (or in the case of SRT, σ_s, σ_r and σ_t), used for each mean shift algorithm by maximizing the registration rate from cross-validation on two training classes (which are not used in the final evaluation). Registration rate is maximized using local search: an initial bandwidth is chosen, then the registration rate computed for this value and the values 1.2 and $1/1.2$ times this value. That value with the highest score is chosen, and the

process is repeated until convergence. With 3 parameters to learn, the local search is computed over a 3D grid.

4.4. Results

Table 3 summarizes the quantitative results for the four inference methods tested. It shows that SRT mean shift performs best at both recognition and registration. The third row gives registration rate taking into account scale and translation only (as the Hough method only provides these), indicating that mean shift performs considerably better than Hough voting at registration. Also given (row 5) is the mean of output scales (each as a ratio of the output scale over the ground truth scale) of the registration result, which shows a marked bias towards a smaller scale when using extrinsic mean shift. Whilst better than extrinsic mean shift at registration, intrinsic mean shift is the slowest[8] method by an order of magnitude.

The per-class registration rates of the mean shift methods are given in table 4, showing that SRT out-performs extrinsic mean shift in 9 out of 10 classes, and intrinsic mean shift in 7 out of 10. The scale-invariance of registration rate, and hence, by implication, recognition rate, using SRT and intrinsic mean shift, and the contrasting scale-variance of extrinsic mean shift (as discussed in §2.2.1), is shown empirically in figure 5.

The confusion matrices for the four inference methods are shown in figure 6. Hough voting performs very poorly on bracket, car and pipe, getting a recognition rate of just 1.3% on average for these classes, which all have low rotational symmetry; in particular it prefers cog and flange (which both have high rotational symmetry), no doubt due to the marginalization this method performs over rotation. Intrinsic mean shift shows a tendency to confuse block, and cog and piston1 to a lesser degree, for other classes, whilst extrinsic and SRT mean shift confuse cog, and block and piston1 to a lesser degree for other classes.

Finally, figure 7(a) demonstrates that SRT mean shift applied to a real scene containing multiple objects. Given a threshold weight above which modes are accepted, mean shift on the votes can produce many false positive detections, as shown by the low precision at high recall rates in figure 7(b). Our system can additionally (though not used here) filter the list of output poses using physical constraints such as the position of the ground plane and collision detection, which we found removed the majority of false positive results, including those shown in the figure.

5. Conclusion

We have introduced the SRT distance for use in mean shift on poses in the space of direct similarity transformations, $S^+(n)$. We have proven the distance to be left-invariant, and have a unique, closed-form mean with the desirable properties of scale, rotation and translation compatibilities. We have demonstrated the use of this distance for registration and recognition tasks on a challenging and realistic 3D dataset which combines real-world objects, with and without rotational symmetries, together with a vision-based geometry capture system and basic features.

Our results show that SRT mean shift has better recognition and registration rates than both intrinsic and extrinsic mean shift, as well as Hough voting. We also show that extrinsic mean shift is not only scale-variant but also biases output scale, and that intrinsic mean shift is slower to compute. In addition to the performance increase over Hough voting, especially in the presence of rotationally symmetric objects, we demonstrate for the first time that mean shift on the full 7D pose space of $S^+(3)$ is not only possible, but that it also provides accurate 7D registration, including rotation. This is not practical using Hough-based approaches, due to their exponential memory requirements.

We address the issue of poor precision at high recall rates in other work [29].

	SRT	Extrinsic	Intrinsic	Hough
Recognition	**64.9%**	49.6%	45.5%	56.1%
Registration	**68.3%**	52.0%	62.0%	–
Reg. (t,s)	**79.8%**	62.0%	75.7%	57.3%[9]
Proc. time	1.6s	9.7s	127s	**0.043s**
Mean scale	**0.995**	0.959	0.987	–

Table 3. **Quantitative results** for the four inference methods tested. The SRT mean shift method is best in all respects except speed, for which it is better than the other mean shift methods.

	bearing	block	bracket	car	cog	flange	knob	pipe	piston1	piston2
SRT	**77**	13	**95**	75	**100**	**41**	**88**	**86**	**44**	63
Extr.	36	12	90	50	80	32	53	63	37	**67**
Intr.	54	**19**	83	**90**	90	36	65	82	34	**67**

Table 4. **Registration rate** per class (%). SRT mean shift performs best on 7/10 classes.

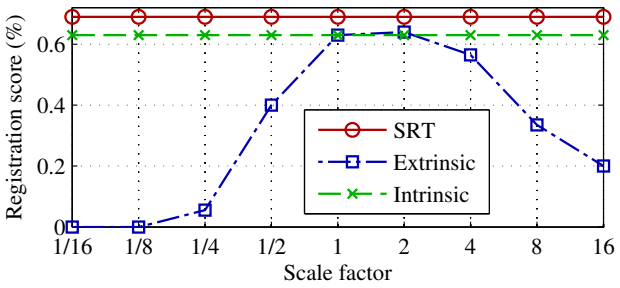
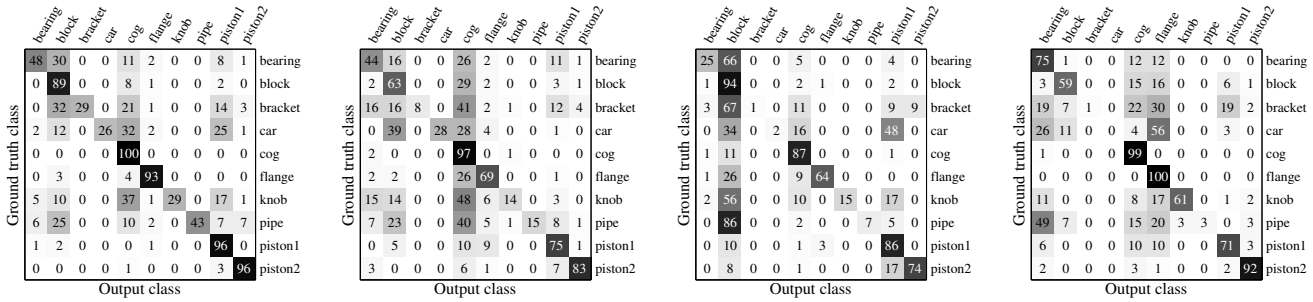


Figure 5. **Scale-invariance.** Registration rate over scale, showing that only extrinsic mean shift varies with scale.

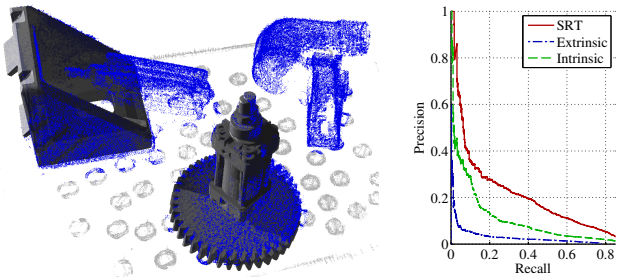
[8] We used optimized implementations for all methods.

[9] This score is the percentage of ground truth poses that were in the same bin as the output pose.



(a) SRT mean shift (b) Extrinsic mean shift (c) Intrinsic mean shift (d) Hough voting

Figure 6. **Confusion matrices** for the four inference methods tested. The Hough voting method performs poorly on objects with low rotational symmetry, while mean shift methods, and in particular SRT, perform better.



(a) SRT output with six objects (b) Precision–recall

Figure 7. **Performance with multiple objects**. (a) Given a point cloud with 6 objects, SRT mean shift finds 3 of them in the top 6 modes, the piston multiple times. (b) Precision-recall curves of the mean shift methods for correct registration and recognition jointly.

References

[1] Toshiba CAD model point clouds dataset. http://www.toshiba-europe.com/research/crl/cvg/projects/stereo_points.html. 6

[2] M. Agrawal. A Lie algebraic approach for consistent pose registration for general euclidean motion. In *IROS*, pages 1891–1897, 2006. 3

[3] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache. A Log-Euclidean polyaffine framework for locally rigid or affine registration. In *Biomedical Image Registration*, volume 4057, pages 120–127. 2006. 3

[4] D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981. 2

[5] P. Besl and N. McKay. A method for registration of 3D shapes. *TPAMI*, 14(2), 1992. 5

[6] H. Cetingul and R. Vidal. Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds. In *CVPR*, pages 1896–1902, 2009. 3, 5

[7] Y. Cheng. Mean shift, mode seeking, and clustering. *TPAMI*, 17:790–799, 1995. 1, 2, 3

[8] P. I. Davies and N. J. Higham. A Schur-Parlett algorithm for computing matrix functions. *SIAM J. Matrix Anal. Appl*, 25:464–485, 2003. 3

[9] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *CVPR*, pages 998–1005, 2010. 1, 2

[10] A. Fisher, R. B. Fisher, C. Robertson, and N. Werghi. Finding surface correspondence for object recognition and registration using pairwise geometric histograms. In *ECCV*, pages 674–686, 1998. 2

[11] K. Khoshelham. Extending generalized Hough transform to detect 3D objects in laser range data. *Workshop on Laser Scanning*, XXXVI:206–210, 2007. 1, 2

[12] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool. Hough transform and 3D SURF for robust three dimensional classification. In *ECCV*, pages 589–602, 2010. 1, 2, 5, 6

[13] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008. 1, 2

[14] A. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *TPAMI*, 28(10):1584–1601, 2006. 2

[15] M. Moakher. Means and averaging in the group of rotations. *SIAM J. Matrix Anal. Appl.*, 24:1–16, 2002. 1, 3, 5

[16] J. L. Mundy. Object recognition in the geometric era: A retrospective. In *Toward Category-Level Object Recognition*, pages 3–28. Springer, New York, 2006. 1

[17] A. Opelt, A. Pinz, and A. Zisserman. Learning an alphabet of shape and appearance for multi-class object detection. *IJCV*, 80(1), 2008. 1, 2

[18] B. Pelletier. Kernel density estimation on Riemannian manifolds. *Statistics Probability Letters*, 73(3):297–304, 2005. 5

[19] X. Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *JMIV*, 25(1):127–154, 2006. 3

[20] X. Pennec and N. Ayache. Uniform distribution, distance and expectation problems for geometric features processing. *J. Math. Imaging Vis.*, 9:49–67, 1998. 4, 5

[21] E. Schramm and P. Schreck. Solving geometric constraints invariant modulo the similarity group. In *ICCSA*, pages 356–365, 2003. 3

[22] J. Shotton, A. Blake, and R. Cipolla. Multiscale categorical object recognition using contour fragments. *TPAMI*, 30(7):1270–1281, 2008. 1, 2

[23] A. Srivastava and E. Klassen. Monte Carlo extrinsic estimators of manifold-valued parameters. *IEEE Trans. on Signal Processing*, 50(2):299–308, 2002. 1, 3

[24] R. Subbarao and P. Meer. Nonlinear mean shift for clustering over analytic manifolds. In *CVPR*, volume I, pages 1168–1175, 2006. 1, 2

[25] R. Subbarao and P. Meer. Nonlinear mean shift over Riemannian manifolds. *IJCV*, 84(1), 2009. 1, 2, 3, 4, 5

[26] F. Tombari and L. Di Stefano. Object recognition in 3D scenes with occlusions and clutter by Hough voting. In *PSIVT*, pages 349–355, 2010. 1, 2, 5

[27] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *ECCV*, 2010. 6

[28] G. Vogiatzis and C. Hernández. Video-based, real-time multi view stereo. *Image and Vision Computing*, 29(7):434–441, 2011. 5

[29] O. J. Woodford, M.-T. Pham, A. Maki, F. Perbet, and B. Stenger. Demisting the Hough transform for 3D shape recognition and registration. In *BMVC*, 2011. 7