3D SHAPE ESTIMATION FROM SILHOUETTES USING MEAN-SHIFT

Donghoon Kim, Jonathan Ruttle & Rozenn Dahyot

School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland

ABSTRACT

In this article, a novel method to accurately estimate 3D surface of objects of interest is proposed. Each ray projected from 2D image plane to 3D space is modelled with the Gaussian kernel function. Then a mean shift algorithm with an annealing scheme is used to find maximums of the probability density function and recovers the 3D surface. Experimental results show that our method is more accurate to estimate 3D surface than the Radon transform-based approach.

Index Terms— 3D shape estimation, 3D shape recovery, 3D shape reconstruction, shape from silhouettes, mean shift

1. INTRODUCTION

3D shape estimation from 2D images has been widely studied in the field of computer vision. One of the most popular methods to compute 3D shape of interest objects is silhouette volume intersection [1]. Each 2D silhouette of the object is segmented from an image, and creates a cone to the 3D world. The intersection of these cones from all camera views gives an estimation of the 3D object volume and shape. This approach is called Visual Hull [2] or Shape from Silhouettes.

According to [3], most methods regarding Shape from Silhouettes can be categorized into a volume-based and surface-based approaches. The volume-based approach focuses on the volume of the visual hull which is discretized as voxels. This approach generally suffers from a heavy computation and memory requirement. The surface-based approach focuses on surface representation of the visual hull. The surface vertices and faces are estimated by intersecting the generalized cones from the occluding contours of the silhouettes. This method requires less computation and memory than the volume-based approach. However, the intersection in the 3D space is sensitive to numerical instabilities, especially in complicated objects.

Similarly, tomographic reconstruction is an important and active research topic in the field of medical image processing, for example Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). The context of tomographic reconstruction is similar to the volume-based approach in terms of a back-projection technique. However, 3D medical image is represented in three dimensions as a stack of two-dimensional images reconstructed from tomographic projections. Each slice in the stack is calculated by the Radon transform, which was first introduced by J. Radon in 1917 and referred to as the x-ray transform or the projection transform. The Radon transform is now a mathematical basis of medical image processing. An explicit and computationally efficient inversion algorithm exists for 2D Radon transforms called filtered back-projection [4, 5]. Tomographic reconstruction reconstructs 3D volume from density data, and silhouette images are a rough approximation of it. Consequently, similarly to the visual hull, using the inverse Radon transform on silhouette images of an object taken from different points of view, allows to reconstruct an approximation of the 3D shape [6].

The visual hull approach can be understood as estimating a 3D histogram describing the probability of a point in space to be part of the object. Instead of a histogram formalism, we propose here a kernel density estimate to reconstruct 3D shape of objects more accurately than the Radon transform based approach [6] (see section 2.2). This new modelling reconstructs each slice of the object with a smooth probability density function (p.d.f.) defined over the spatial domain and the surface of the object is then estimated using the gradient ascent Mean-shift algorithm to find the maxima of the p.d.f [7, 8, 9]. This new algorithm is presented in paragraph 2.3. Advantages of our method are that it does not need any camera calibration parameters since orthographic projection is assumed, it has a light memory requirement and it is numerically stable. We compare the accuracy of our new method to the Radon Transform (see section 3).

2. MEAN-SHIFT FOR 3D SHAPE INFERENCE

2.1. Hypotheses and Notations

The 3D shape is recovered by first reconstructing the 2D slices of the object from each lines of the silhouette images. For simplification, the camera matrix is chosen as an orthogonal projection: each foreground pixel on the 2D silhouette images is projected from the 2D image plane to 3D space as a ray using an orthographic projection. We use a polar coordinate system to define the ray created by the foreground pixel

i:

$$\rho_i = x\ \cos\theta_i + y\ \sin\theta_i \tag{1}$$

All parameters $\{(\rho_i, \theta_i)\}_{i=1,\cdots,n}$ for all foreground pixels in all image silhouettes from all camera views are known.

2.2. Kernel density estimator

Having only one ray (ρ_i, θ_i), we propose to model the probability density function of the random variable $\mathbf{x} = (x, y)$, representing the spatial position of the object in the 2D slice. If we wanted to have all possible positions to be exactly on the ray generated by (ρ_i, θ_i), then we could select the Dirac kernel as follow:

$$\hat{p}(\mathbf{x}|(\rho_i, \theta_i)) \quad \propto \quad \delta\left((\rho_i - x\ \cos\theta_i - y\ \sin\theta_i)\right) \tag{2}$$

Instead, we propose to use the gaussian kernel:

$$\hat{p}(\mathbf{x}|(\rho_i, \theta_i)) \propto \frac{1}{\sqrt{2\pi}h} \exp\left(\frac{-(\rho_i - x\ \cos\theta_i - y\ \sin\theta_i)^2}{2h^2}\right) \tag{3}$$

This latter choice allows to consider the positions close to the ray as potential positions for the object with a probability non-zero. Each foreground pixel now creates a fuzzy cylinder of possible positions of the object instead of a strict line (ray).

When considering the set of all rays, $\{(\rho_i, \theta_i)\}_{i=1,\cdots,n}$ and assuming them all equiprobable, then we can propose the following kernel estimate for $\mathbf{x}$:

$$\hat{p}(\mathbf{x}) \propto \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h} \exp\left(\frac{-(\rho_i - x\ \cos\theta_i - y\ \sin\theta_i)^2}{2h^2}\right) \tag{4}$$

As an illustration, Figure 1 shows the probability density generated by equation (4) for a slice of a spherical object (soccer ball) reconstructed using 36 equally spaced camera views, and compared with the one estimated by the Radon transform. As can be noticed, the Radon transform reconstructs a surface (or empty object), whereas the $\hat{p}(\mathbf{x})$ models a volume object (or full object). Our purpose is now to recover the positions $\mathbf{x}$ corresponding to maxima of $\hat{p}(\mathbf{x})$. This will be performed using Mean-shift.

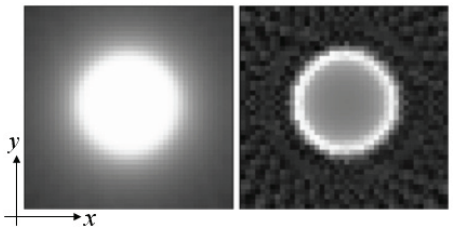

Fig. 1. Probability density $\hat{p}(\mathbf{x})$ (left), Radon transform (right).

2.3. Mean shift Algorithm

We note $d_i(\mathbf{x}) = (\rho_i - x\cos\theta_i - y\sin\theta_i)$. The gradient of $\hat{p}(\mathbf{x})$ is:

$$\nabla\hat{p}(\mathbf{x}) = \left[\begin{array}{l} \dfrac{\partial\hat{p}(\mathbf{x})}{\partial x} \propto \frac{1}{\sqrt{2\pi}nh^3}\displaystyle\sum_{i=1}^{n} d_i(\mathbf{x})\ \cos\theta_i\ \exp\left(\frac{-d_i^2(\mathbf{x})}{2h^2}\right) \\ \\ \dfrac{\partial\hat{p}(\mathbf{x})}{\partial y} \propto \frac{1}{\sqrt{2\pi}nh^3}\displaystyle\sum_{i=1}^{n} d_i(\mathbf{x})\ \sin\theta_i\ \exp\left(\frac{-d_i^2(\mathbf{x})}{2h^2}\right) \end{array}\right] \tag{5}$$

Using (5), starting from an initial position $\mathbf{x}^{(0)}$, the mean-shift iteration to converge towards the nearest local maximum, is then:

$$\mathbf{x}^{(m+1)} = \left(L(\mathbf{x}^{(m)})\right)^{-1} \cdot M(\mathbf{x}^{(m)}). \tag{6}$$

where $L(\mathbf{x})$ is a 2×2 matrix:

$$L(\mathbf{x}) = \sum_{i=1}^{n} \exp\left(\frac{-d_i^2(\mathbf{x})}{2h^2}\right) \times \left[\begin{array}{ll} \cos^2\theta_i & \sin\theta_i\ \cos\theta_i \\ \cos\theta_i\ \sin\theta_i & \sin^2\theta_i \end{array}\right] \tag{7}$$

and $M(\mathbf{x})$ is the 2×1 vector:

$$M(\mathbf{x}) = \sum_{i=1}^{n} \exp\left(\frac{-d_i^2(\mathbf{x})}{2h^2}\right) \times \left[\begin{array}{l} \rho_i\cos\theta_i \\ \rho_i\sin\theta_i \end{array}\right]. \tag{8}$$

From the starting position $\mathbf{x}^{(0)}$, the iteration (6) is repeated until convergence.

2.4. Bandwidths

If the number of cameras is small, the probability density might be noisy, and have many spurious modes. In that case, the ordinary mean shift algorithm might get trapped in meaningless local maxima. To overcome this problem, we propose to use the mean shift algorithm with a simulated annealing scheme [10, 11].

The bandwidth starts large which results in a smoother probability density function with less local maxima. As the mean-shift point approaches the global maximum the bandwidth is decreased to achieve the greatest accuracy possible. This scheme allows our method to robustly and quickly converge. The rate at which the bandwidth decreases from h_{max} to $h_{min} = 1$ is based on a geometric rate [11]:

$$h_b = \alpha^b h_{max} \text{ until } h_b = h_{min} \text{ with } \alpha = 0.98, \qquad (9)$$

The minimum bandwidth reflects the uncertainty on the pixel resolution $h_{min} = 1$. $h_{max} = 10$ has been chosen experimentally. In the case when few cameras are used, the mean-shift iteration may be trapped in a local maximum. This local maxima can be avoided by re-increasing the bandwidth, since we know the value of the density on the object $\max \hat{p}(\mathbf{x})$ which can be calculated by the number of maximum intersected rays.

Several guess points are created in the spatial domain around the object, and these are moved using the simulated Mean-shift algorithm until convergence. The contour of the object in the slice is then inferred by connecting the closest points together.

3. EXPERIMENTAL RESULTS

To test our method, a dataset with a ground truth was created using Autodesk 3ds Max (see examples Fig. 4). 10 mesh objects of various sizes and proportions were selected and the silhouette of each was orthographically projected into 360 image planes equally spaced around each object (equivalent setting as a turning table [6]).

The 3D surface of each object in the dataset is estimated from the silhouette images using our Mean-shift method and the Radon transform-based approach (computed by the function *iradon* in Matlab and associated with the canny edge detector for finding the maxima) [6]. The original mesh object was used as the ground truth. Figure 2 presents the contours estimated by Mean-Shift (green) and the Radon Transform (blue) in a slice of the face (see Fig. 4) compared to the ground truth (red). Both estimates are convex approximation to the reference and the mean-shift reconstruction is closer to the ground truth.

An euclidian distance is computed between the reconstructed meshes and the reference one, and normalized to remove the effect of the scale of the object. Figure 3 shows these distances computed for several camera views (4 to 36 cameras) and averaged over all our 10 objects. Note that for each number of cameras (abscissa in Figure 3), the experiment is repeated on each object by selecting randomly the equally spaced camera views. This is done to remove any effect that some particular views maybe more informative than others. The standard error is also reported. As can be seen in Figure 3, the Mean-shift based reconstruction outperforms the Radon transform one (the distance to the reference is smaller) and

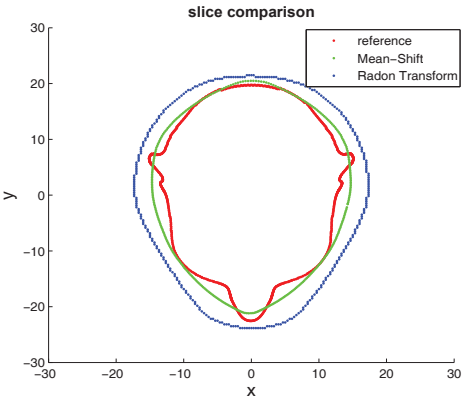

Fig. 2. Slice reconstruction of the object *head* (using 36 views).

as expected for both methods, the distance decreases up to a point as more camera views are available. The two graphs have a similar pattern as the number of cameras changes. However, our method is more accurate, and when the distance is close to the limit of 0.5, adding more cameras does not improve the accuracy.

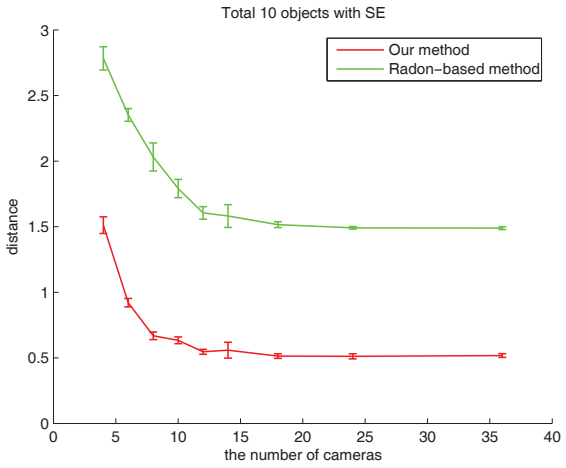

Fig. 3. Distance plot with standard error: Mean-shift (red) and Radon Transform (green).

Figure 4 presents a few 3D objects used with the Mean-shift reconstructions estimated from 36 camera views.

4. CONCLUSIONS AND FUTURE WORK

We have proposed a new approach to 3D shape estimation using 2D silhouette images recorded from several camera views. It is based on a new kernel density estimate of the density

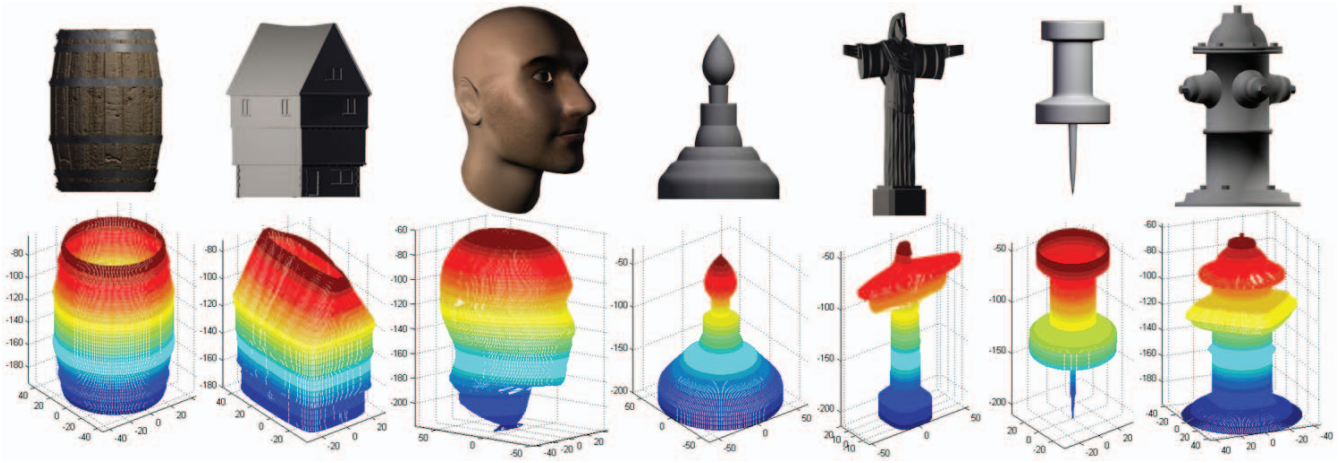

Fig. 4. 3D reference models (top row) and 3D Mean-Shift estimates (bottom row).

function of the spatial position in each slice and its corresponding Mean-shift algorithm for finding its maxima. Experimental results have shown that this new statistical approach is more accurate than the Radon transform approach where the inverse transform is computed using the filtered back-projection algorithm.

Future work will look at including priors in our statistical modelling to get more accurate 3D shape estimation. In addition, we will look at adding colour information in the framework so that segmenting the object as a pre-process will not be necessary anymore.

5. ACKNOWLEDGEMENT

This project is supported by the "Irish Research Council for Science, Engineering and Technology in collaboration with INTEL Ireland Ltd: funded by the National Development Plan" and by an Innovation Partnership between Sony-Toshiba-IBM and Enterprise Ireland (IP-2007-505).

6. REFERENCES

[1] W. N. Martin and J. K. Aggarwal, "Volumetric description of objects from multiple views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, pp. 150–158, January 1983.

[2] A. Laurentini, "How far 3d shapes can be understood from 2d silhouettes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 188–195, 1995.

[3] J.S. Franco and E. Boyer, "Efficient polyhedral modeling from silhouettes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(3), pp. 414–427, March, 2009.

[4] D. Ludwig, "The radon transform on euclidean space," *Comm. Pure Appl. Math.*, vol. 19, pp. 49–81, 1966.

[5] G. N. Ramachandran and A. V. Lakshminarayanan, "Three dimensional reconstructions from radiographs and electron micrographs: Application of convolution instead of fourier transform," *Nat. Acad. Sci*, vol. 68, pp. 2236–2240, 1971.

[6] C. Pintavirooj and M. Sangworasil, "3d shape reconstruction based on radon transform with application in volume measurement," *10th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2002.

[7] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density gradient, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, pp. 32–40, 1975.

[8] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 790–799, 1995.

[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[10] C. Shen, M. Brooks, and A. van den Hengel, "Fast global kernel density mode seeking: Applications to localization and tracking," *IEEE Transactions on Image Processing*, vol. 16, May 2007.

[11] R. Dahyot, "Mean-shift for statistical hough transform," Tech. Rep., Technical report department of Statistics, Trinity College Dublin, April 2009.