Object Category Recognition with Projected Texture

Avinash Sharma and Anoop Namboodiri
Center for Visual Information Technology
International Institute of Information Technology, Hyderabad, India
{avinash_s@research., anoop@}iiit.ac.in

Abstract

Recognition of object categories from their images is extremely challenging due to the large intra-class variations, and variations in pose, illumination and scale, in addition to lack of depth information of the object. Recovering the depth information from multiple images or from image cues such as variations in illumination or focus, is both computationally intensive and error prone. In contrast, the appearance based approaches are more robust and computationally efficient. However, they lack the potential accuracy of 3D feature based approaches due to the lack of shape information. We propose the use of structured lighting patterns projected on the object, which gets deformed according to the shape of the object for recognition. Since our goal is object classification and not shape recovery, we characterize the deformations using simple texture measures, thus avoiding depth recovery step. Moreover, the shape information present in the deformations is implicitly used for classification. We show that the information thus derived can significantly improve the accuracy of object category recognition from arbitrary-pose images.

Figure 1. Variation in appearance in three different objects with change in pose.

1. Introduction

The shape of a three dimensional object can be thought of as the variations in depth over the object, looking from a particular view point. For rigid objects, the nature of these variations is deterministic, while that of a 3D surface texture is stochastic in nature. Even for rigid objects, a particular shape can result in different depth profiles depending on the view. During imaging, the depth information is lost as the camera captures only image cues such as variations in shading, focus, texture, silhouette, etc., which are induced by the shape and lighting.

Figure 1 shows, the variation in appearance of three different objects from different view points. Note that objects with different shapes may appear quite similar from certain view points (first row), while the same object might look very different as the view changes (columns of figure 1). Clearly, it is extremely difficult to recognize such objects from their appearance only. The problem of object category recognition has to deal with the additional variations introduced by the differences in shape of objects within a category (such as a cup), and the similarity in appearance of different object classes (see bowl and saucer in figure 1). The ability to recognize objects and object classes plays an important role in enabling a variety of applications such autonomous robots for human assistance and automated search and rescue operations, factory automation, mobile camera based and information retrieval, automated surveillance, etc.

Object category recognition algorithms that use the object shape explicitly, try to recover the depth information from images using a variety of methods. Shape recovery methods using correspondence information from multiple

978-0-7695-3476-3/08 $25.00 © 2008 IEEE
DOI 10.1109/ICVGIP.2008.62

images or additional cues from a single image such as shading, focus, texture, etc. [21] could fail in presence of large variations in the texture and scale of different objects. Methods using a projector-camera pair for structured lighting can be more robust in presence of varying object appearance. Structured lighting can be used for recovering 3D shape [18] and for improving correspondence [22, 13, 10] in multiple images. In [12], encoded structured light was used for recognition by improving the correspondence. We note that the process of depth recovery is both computationally intensive as well as error prone as the nature and shape of objects vary considerably, as in the case of object category recognition. Moreover, for the purpose of recognition, one needs to characterize the recovered depth, invariant to the pose variations, which is not a trivial task [14].

An alternate approach is to recognize the object classes directly from appearance based features. Riesenhuber and Poggio [15] explores various view based representations of objects for recognition. Image-based object recognition algorithms usually compromise the depth information for higher computation efficiency and robustness. They characterize the objects directly based on image features such as edges, shading [2] and texture [11]. The challenge here is to come up with a representation that is invariant to illumination and pose, thus providing robust classification. In addition, different objects of a category can vary considerably in appearance [8].

Popular approaches to object recognition use part and structure models [9] to represent an object. Examples of such models include the constellation model [3]. The structure can also be learned with probabilistic models from examples [19, 20]. Current algorithms employ powerful machine learning approaches to learn the object structure in terms of object part descriptors from large number of examples [7, 5, 8]. Such approaches have become popular both due to its simplicity and robustness to object and pose variations within a category. However, the classification power of image based approaches is limited as some of the 3D information is lost during the imaging process. Detailed surveys of appearance based approaches for object and category recognition may be found in [21, 8].

Our primary contribution to the area is the use of structured lighting patterns, which we refer to as *projected texture*, for the purpose of recognition rather than explicit shape recovery. The deformations in the projected texture are induced by the depth variations over the object, and these deformations encode the shape information. We also propose a set of simple position and pose invariant features for characterizing the deformations based on the popular bag-of-words paradigm [6] for object representation (see figure 2). Formulation of the problem as that of recognition of the object directly from the deformations rather than reconstruction allows us to concentrate on features that are

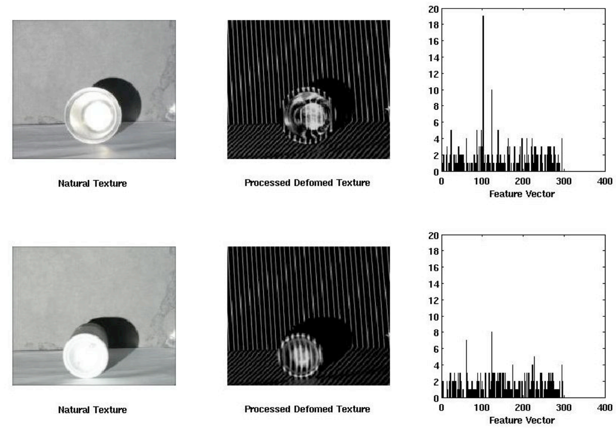


Figure 2. Variations in deformation on similar looking objects.

relevant to discrimination between object classes rather than reconstruction of their exact shape. Experimental results clearly indicate the superiority of the approach as compared to traditional image based classification algorithms.

2. Projected Patterns and Deformation

A 3D object can either be characterized by its exact shape (rigid or non-rigid objects), or the nature of stochastic variations (3D texture). In this paper, we look at the problem of recognizing rigid objects from arbitrary poses. Specifically, we concentrate on the problem of category recognition, where the learning algorithm needs to generalize from a variety of objects in a category. Note that the images of a functional category of objects such as a coffee mug can vary significantly in size, shape and pose. Our challenge is to arrive at a shape descriptor from the image that is discriminative enough to separate the functional categories, while generalizing over the within-class variations. In our case, this boils down to the characterization and representation of pattern deformations of an object category.

The object is placed in an arbitrary pose and a specific texture pattern is projected on it while imaging. The projected pattern, falling on the scene containing the object, gets transformed according to the depth map of the object under illumination. These transformations falls into two categories: *pattern shift* and *pattern deformation*. Pattern shift depends upon the absolute depth of the surface from which the projected light in reflected, in addition to the projector camera configuration. On the other hand, pattern deformation depends on the change in depth of the surface. These deformations depend on the absolute angle between the projector axis and the normal to the surface at a point as well as its derivative.

We will now look at the exact relationship between the projector camera configuration, the object surface and the amount of deformation.

Figure 3(b) shows a planar object being illuminated by a sheet of light. The slope of the line can be expressed in terms of the angle of illumination, ϕ, and the slope of the object plane, θ (see Appendix A). Thus, the slope of object surface directly affects orientation of the projection of a 3D line onto the image plane. This indicates that a characterization of an image patch in terms of the angle of the imaged lines can capture the surface height variations at that point. The above relationship enables us to predict the projector camera configuration as well as the pattern to be projected for a class of objects with a specific range of depth variations. In addition to depth deformation, one also need to take into account the reflective properties of the object surface and shadow effects while deciding on a projection pattern. In our problem we selected a set of vertical stripes as the texture, since the camera and projector are displaced horizontally in the setup. The spacing and width of patterns were selected experimentally, while intensities were chosen to reduce inter-reflections and specularities.

2.1. Selection of Projection Pattern

The choice of an appropriate projection pattern is important due to a variety of factors:

1. For the deformation to be visible in the captured at any point in the image, the gradient of the texture at that point should not be zero in the direction of gradient of the object depth.
2. One should be able to capture the deformations of the projected pattern using the texture measure employed for this purpose.
3. The density of the projected pattern or its spatial frequency should correspond to the frequency of height variations to be captured. Hence, analyzing the geometry of an object with a high level of detail will require a finer texture, whereas in the case of an object with smooth structural variations, a sparse one will serve the purpose.
4. Factors such as the color, and reflectance of the object surface should be considered in selecting the color, intensity and contrast of the texture so that one can identify the deformations in the captured image.

For the purpose of object category recognition, we use a set of parallel lines with regular spacing, where the spacing is determined based on the scale of the objects to be recognized. The width and spacing of the lines were selected experimentally so that it captures the height variations within an object at the angle of projection selected. Once the selected pattern is projected on the object of interest, we need to compute a characterization of the deformations that is appropriate for recognition. In [4], it has been shown that curvature of the projected pattern gives information about the surface type.

3. Characterizing Shape with Deformation

The primary concern in developing a representation for object category is that the description should be robust to both shape and pose of the object class. Note that the use of projected patterns allows us to avoid object texture, and concentrate only on its shape. Approaches such as 'bag of words' computed from interest points have been successfully employed for image based object category recognition [8]. Our approach is similar in spirit to achieve pose invariance. We learn the class of local deformations that are possible for each category of objects by creating a codebook of such deformations from a training set. Each object is then represented as a histogram of local deformations based on the codebook. Figure 4 illustrates the computation of the feature vector from a scene with projected texture.

There are two primary concerns to be addressed while developing a parts based shape representation:

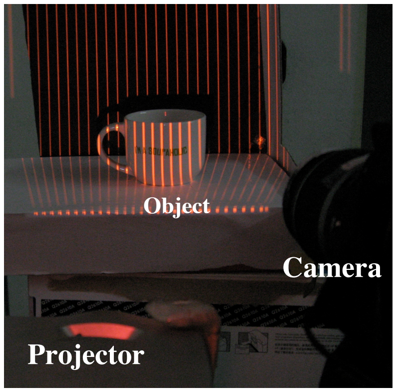
The location of points from which the local shape descriptor is computed is important to achieve position invariance. In image based algorithms, the patches are localized by using an interest operator that is computed from object texture or edges. However, in our case the primary objective is to avoid using texture information and concentrate on the shape information provided by the projected texture. Hence we choose to use a set of overlapping windows that covers the whole scene for computation of local deformations. Our representation based on the codebook allows us to concentrate on the object deformation for recognition.

The description of the local deformations should be sufficient to distinguish between various local surface shapes within the class of objects. The feature vector computed should exploit the periodic nature of projected pattern that we employ. Fourier domain representations provide us with effective descriptors for periodic signals. A 2D discrete Fourier transform (DFT) of an image function $f(x, y)$ can be described by equation [1].

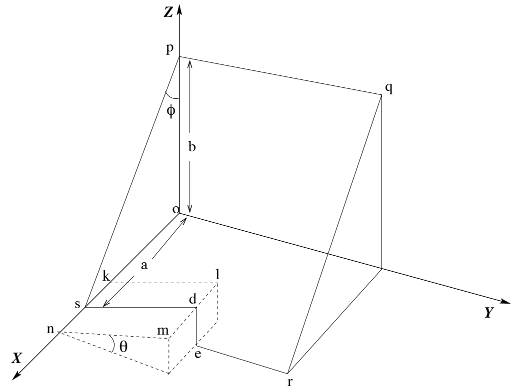
$$F(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \exp^{-j2\pi(\frac{xu}{M} + \frac{yv}{N})}, \quad (1)$$

where $|F(u, v)|$ represents the magnitude of 2D DFT and Φ represents phase information as defined in equation (2).

$$\begin{aligned} |F(u, v)| &= \sqrt{F_r(u, v)^2 + F_i(u, v)^2} \\ \Phi &= \tan^{-1}\left(\frac{F_r(u,v)}{F_i(u,v)}\right) \end{aligned} \quad (2)$$



(a)



(b)

Figure 3. The image capture setup and the pattern deformation geometry.

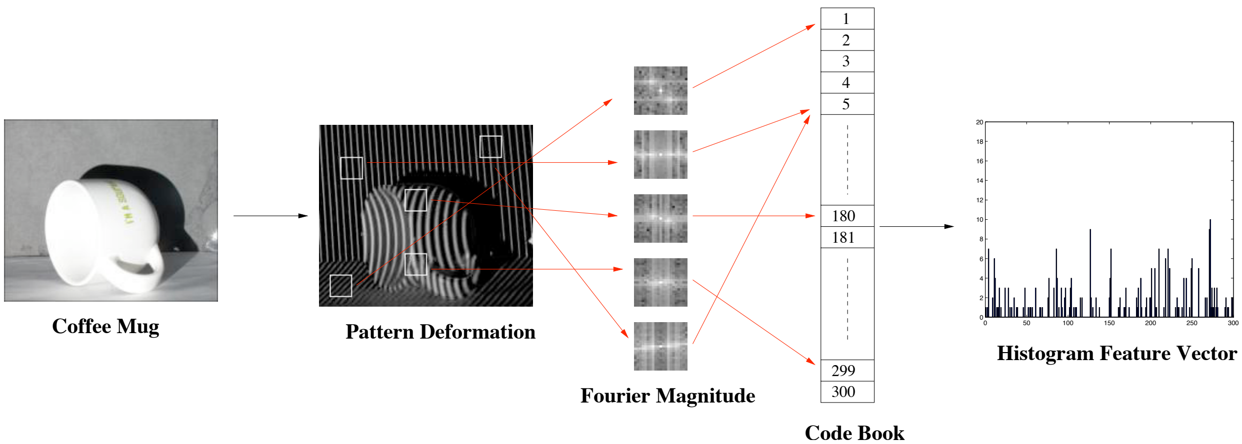
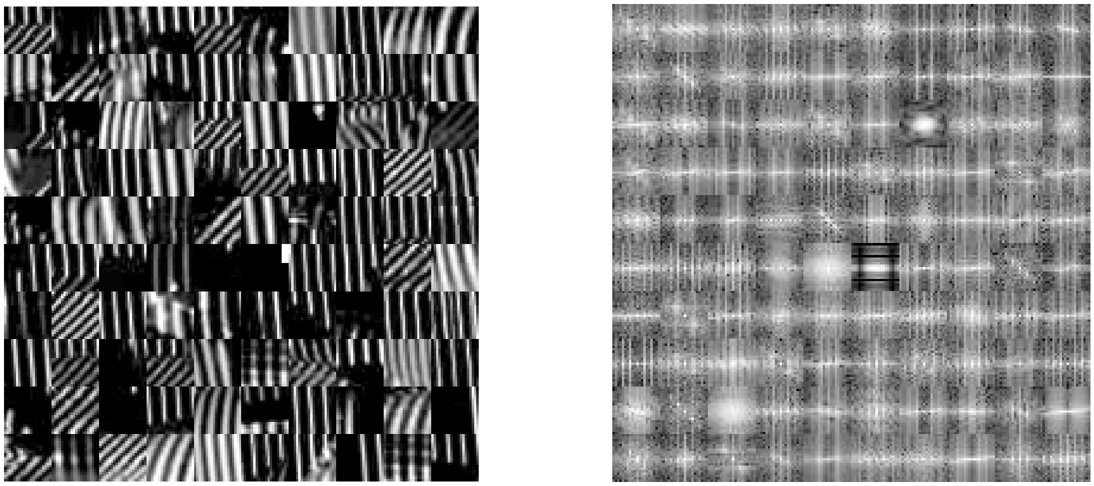


Figure 4. Computation of feature vector.

Since we are interested in the nature of deformations and not its exact location, we compute the magnitude of the Fourier coefficients (referred to as *AFC*) within each of the window patch as our feature vector. To make comparisons in a Euclidean space more effective, we use a logarithmic representation of these coefficients (*LAFC*). We show that this simple Fourier magnitude based representation of the patches can effectively achieve the discriminative power that we seek.

The feature extraction process proceeds as follows. The images in the training set are divided into a set of overlapping windows of size 20×20 (decided experimentally). Each window is then represented using the magnitude of Fourier representation in logarithmic scale (*LAFC*). This results in a 200 dimensional feature vector (due to symmetry of Fourier representation) for each window. A K-means clustering of windows in this feature space allows us to identify the dominant pattern deformations, which forms a codebook for the classification problem (see figure 5). During the testing phase, the feature representations of the windows in an image are computed as above, and each window is mapped to the nearest codebook vector. A histogram of the codes present in an image forms the representation of the object contained in it. As shown in figure 4 the patches that are part of the background maps to a few locations in codebook. Thus codebook can isolate the words that captures maximum information for defining an object category.

We note that the representation is independent of the position, while the classification algorithm achieves pose invariance due to the generalization from different poses in the training set.



(a) (b)

Figure 5. Spatial and spectral representations of 100 words from our codebook.

4. Experiment Results and Analysis

We have collected dataset of 225 images of objects selected from 5 object categories: i) *Coffee Cup*, ii) *Steel Tumbler*, iii) *Plastic Mug*, iv) *Deodorant Spray*, and v) *Alarm Clock*. Even though the number of categories is limited, they were chosen to introduce challenging similarities between the categories. Five objects of each category were chosen so that they have large intra-class variations (see Figure 6). For each object, 9 different images were collected with views around 45 degrees apart, making the dataset an challenging one. All images were captured under 8 different texture patterns with varying widths as well as under uniform illumination for comparison.

The pre-processing stage includes removal of object texture by subtracting a uniformly illuminated image of the object from the image with projection and Gaussian smoothing to reduce the imaging noise. For the purpose of classification, we have used two different classifiers: Multi Layer Perceptron (MLP), which has good generalization capabilities, and a simple Nearest Neighbor (1NN). All results reported are the mean error rates based on 4-fold cross validation with training and testing sets decided randomly. The number of hidden nodes in the MLP was set to 20 for all experiments.

For the purpose of comparison, we conducted similar experiments with SIFT based features proposed by [8] on our dataset without the projected patterns. SIFT feature might not be work well in case of object with low texture but our approach will avoid this problem as we are projecting texture on object surface. Note that the comparison is made only to show the effect of the additional information introduced by the projected patterns into the classification process and is not a testimony of the classification algorithm itself. In fact, the algorithms are remarkably similar, and the primary differences are in the selection of locations of the patches and its representation. Table 1 presents the mean error for both of the approaches, which clearly shows superiority of our approach over the state-of-the-art. The error rate is only 1.33%, which amount to three misclassifications on the whole dataset.

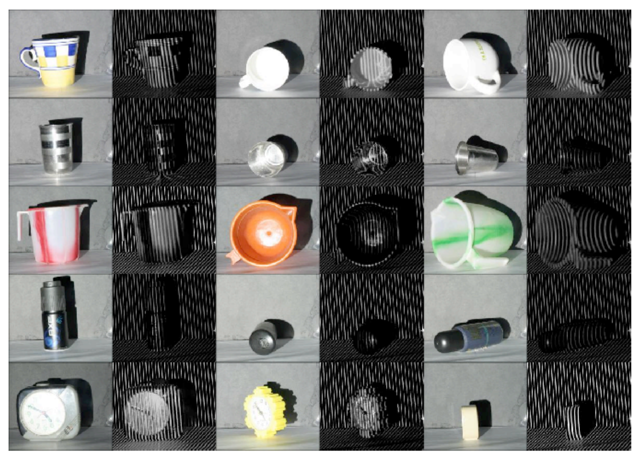


Figure 6. Selected views of objects from the dataset.

Table 2 shows the confusion matrix between the object classes, and Figure 7 shows an example of the misclassified

	LAFC	SIFT
MLP	1.33 %	21.33 %
1-NN	5.73 %	20.09 %

Table 1. Error Rates for Recognition with the proposed and SIFT feature.

	1	2	3	4	5
1	45	0	0	0	0
2	0	45	0	0	0
3	0	0	45	0	0
4	0	2	0	43	0
5	0	0	1	0	45

Table 2. Confusion matrix of object categories.

objects and its representation.

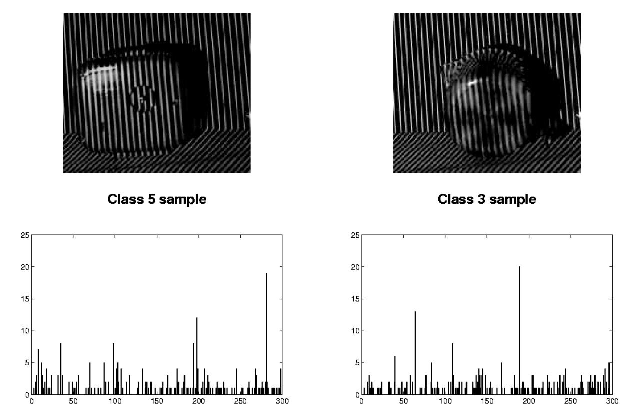


Figure 7. Miss classification example

We also conducted experiments with different codebook sizes and pattern variations. Figure 8(a) shows the graph of accuracy vs size of code book, and Figure 8(b) depicts the variation in performance with change in width of projected pattern. We note that variations in performance with variations in both the codebook size and the pattern width are not significant when compared to the gain in performance from the existing image based approaches.

We have also extended the approach for other categories of recognition such as fixed pose objects as well as 3D surface textures. With appropriate projections and deformation characterizations, the idea of projected can effectively enhance the applicability and performance in such problems as well. The results were reported elsewhere.

5. Conclusion and Future Work

We have proposed a novel approach for object category recognition based on characterization of objects from deformations of a projected light pattern. A patch based representation of the object categories is proposed, where each patch is characterized by a frequency domain representation of the deformed texture. The effectiveness of the approach is demonstrated on a small but challenging dataset, which demonstrates a significant improvement in recognition rates. The approach can also be used for other categories of objects as with fixed pose objects as well as 3D surface textures.

We are currently working on approaches to deal with more complex backgrounds as well as to verify the approach on a larger dataset of objects. Another possible direction of future work is to learn appropriate projection patterns that are best suited to discriminate different classes of objects as well as objects in a specific dataset. Another direction to improve the performance is to avoid quantization as proposed by [1]. Also a hybrid approach can be tested using natural texture in combination with projected texture. The idea of projected texture is successfully applied for biometric authentication [17] and for classifying 3d texture surface [16].We can also apply this idea for other shape based biometric application like 3d face based recognition.

Appendix

Quantifying Deformations of Projected Texture

Let θ be the slope of the plane (we will call it object surface plane) with respect to the X-Y plane. Let the angle between the plane created by the projected line (we call this, projector plane) and the Y-Z plane be ϕ. The equation of projector plane will be

$$\frac{x}{a} + \frac{z}{b} = 1,$$

where a can be expressed as

$$a = b \tan \phi$$

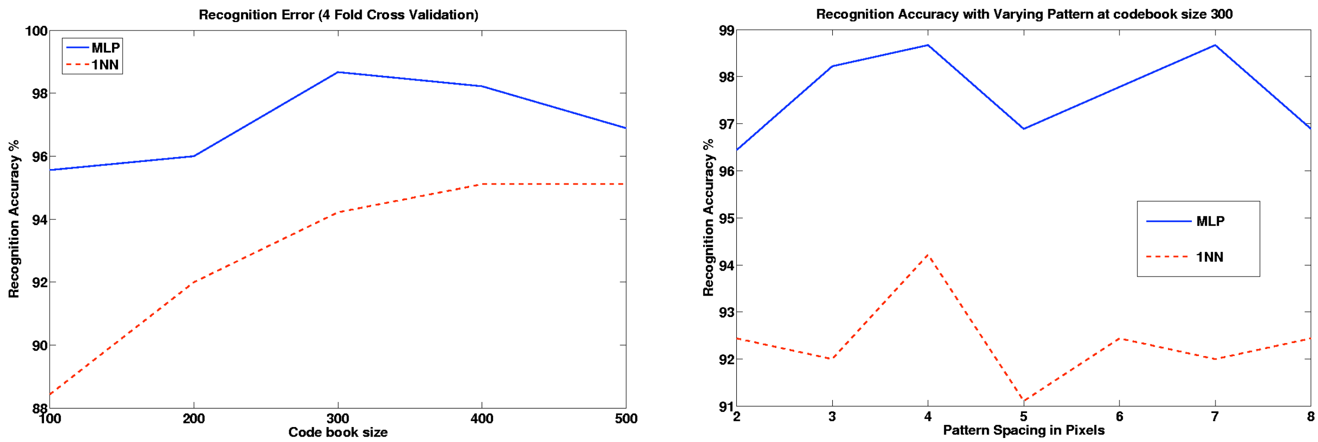
Thus, we can express the projector plane in terms of b and ϕ as

$$x \cot \phi + z - b = 0 \tag{3}$$

Object surface plane can be represented by

$$z - y \tan \theta = 0 \tag{4}$$

The line cd as shown in figure is the intersection of both of these planes, and it can be expressed by single point on



(a) (b)

Figure 8. Performance with variation in Codebook size and Pattern width.

that line and the direction vector of the line obtained by finding cross product of the normal of both intersecting planes. If $\vec{n_1}$ and -$\vec{n_2}$ represent the normals of the projector plane and object surface plane respectively, the direction of the line cd will be

$$\vec{n_3} = \vec{n_1} \times \vec{n_2}$$

$$\vec{n_3} = [\,\cot\phi \quad 0 \quad 1]^T \times [\,0 \quad \tan\theta \quad -1]^T$$

or,

$$\vec{n_3} = [\,-\tan\theta \quad \cot\phi \quad \tan\theta\cot\phi]^T \tag{5}$$

One point common to both plane say p can be obtained by solving equation 3 and 4

$$p = [\,b\tan\phi \quad 0 \quad 0]^T$$

Hence equation of 3D line can be written as

$$\vec{r} = [\,b\tan\phi - s\tan\theta \quad s\cot\phi \quad s\tan\theta\cot\phi]^T, \tag{6}$$

where s is the line parameter and different values of s will give different points on line.

In order to express a 2D projection of this 3D line onto the image plane of a camera, we consider two points on 3D line such that they are in the Field of View (FOV) of camera. Let Q_1 and Q_2 be two such points, with corresponding value of s as $s = l_1$ and $s = l_2$ respectively.

$$Q_1 = [\,b\tan\phi - l_1\tan\theta \quad l_1\cot\phi \quad l_1\tan\theta\cot\phi]^T \tag{7}$$

$$Q_2 = [\,b\tan\phi - l_2\tan\theta \quad l_2\cot\phi \quad l_2\tan\theta\cot\phi]^T \tag{8}$$

For simplicity, let us assume camera to be a pinhole camera with camera matrix $P = K[R|t]$. Let $K = I$ (*i. e.* the internal parameter matrix is unity matrix) and R and t be

$$R = \begin{bmatrix} R_1 & R_2 & R_3 \\ R_4 & R_5 & R_6 \\ R_7 & R_8 & R_9 \end{bmatrix}, t = \begin{bmatrix} t_1 & t_2 & t_3 \end{bmatrix}^T$$

The image of these points in camera plane be $q_1 = PQ_1$ and $q_2 = PQ_2$. q_1 can be represented in matrix form in terms of R_1 to R_9, l_1 and ϕ, θ

$$q_1 = \begin{bmatrix} R_1(b\tan\phi - l_1\tan\theta) + R_2 l_1\cot\phi + R_3 l_1\tan\theta\cot\phi + t_1 \\ R_4(b\tan\phi - l_1\tan\theta) + R_5 l_1\cot\phi + R_6 l_1\tan\theta\cot\phi + t_2 \\ R_7(b\tan\phi - l_1\tan\theta) + R_8 l_1\cot\phi + R_9 l_1\tan\theta\cot\phi + t_3 \end{bmatrix} \tag{9}$$

For simplifying the expressions, lets write q_1 in terms of variables X_1, Y_1 and Z_1.

$$q_1 = \begin{bmatrix} X_1 & Y_1 & Z_1 \end{bmatrix}^T, \tag{10}$$

where,

$$X_1 = R_1(b\tan\phi - l_1\tan\theta) + R_2 l_1\cot\phi + R_3 l_1\tan\theta\cot\phi + t_1$$

$$Y_1 = R_4(b\tan\phi - l_1\tan\theta) + R_5 l_1\cot\phi + R_6 l_1\tan\theta\cot\phi + t_2$$

$$Z_1 = R_7(b\tan\phi - l_1\tan\theta) + R_8 l_1\cot\phi + R_9 l_1\tan\theta\cot\phi + t_3$$

similarly q_2 can be represented in terms of R_1 to R_9, l_2 and ϕ, θ or, in term of variables X_2, Y_2 and Z_2.

$$q_2 = \begin{bmatrix} X_2 & Y_2 & Z_2 \end{bmatrix}^T \quad (11)$$

In homogeneous coordinate system q_1 and q_2 can be represented as

$$\overline{q_1} = \begin{bmatrix} \frac{X_1}{Z_1} & \frac{Y_1}{Z_1} \end{bmatrix}^T, \overline{q_2} = \begin{bmatrix} \frac{X_2}{Z_2} & \frac{Y_2}{Z_2} \end{bmatrix}^T \quad (12)$$

Thus the equation of line in 2D image plane can be written as

$$\vec{L} : \overline{q_1} \times \overline{q_2} = 0$$

or,

$$\vec{L} \; : \; X(Z_1Y_2 - Z_2Y_1) - Y(Z_1X_2 - Z_2X_1) - X_1Y_2 + X_2Y_1 = 0 \quad (13)$$

$$m = (Z_1Y_2 - Z_2Y_1)/(Z_1X_2 - Z_2X_1) \quad (14)$$

From the equation of line it can inferred that the slope m of the line in the image computed in equation (14) will depend upon X_1, Y_1, Z_1 and X_2, Y_2, Z_2, which can be further expanded in terms of b, ϕ and θ.

References

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proceedings of CVPR*, pages 1–8, 2008.

[2] M. Burl and P. Perona. Recognition of planar object classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 223–230, 1996.

[3] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proceedings of the European Conference on Computer Vision*, pages 628–641, 1998.

[4] R. Cipolla and A. Zisserman. Qualitative shape from deformation of image curves. *International Journal of Computer Vision*, 8(1):53–69, 1992.

[5] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego*, volume 1, pages 10–17, 2005.

[6] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[7] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the 9th International Conference on Computer Vision, Nice, France,*, pages 1134–1141, 2003.

[8] R. Fergus. *Visual Object Category Recognition*. PhD thesis, University of Oxford, 2005.

[9] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, 22(1):67–92, 1973.

[10] S. Inokuchi, K. Sato, and F. Matsuda. Range imaging system for 3-d object recognition. In *International Conference on Pattern Recognition*, pages 806–808, 1984.

[11] D. Lowe. Local feature view clustering for 3d object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, pages 682–688. Springer, 2001.

[12] E. Petriu, Z. Sakr, H. Spoelder, and A. Moica. Object recognition using pseudo-random color encoded structuredlight. In *Proceedings of Instrumentation and Measurement Technology Conference*, volume 3, pages 1237–1241, 2000.

[13] J. Posdamer and M. Altschuler. Surface measurement by space-encoded projected beam system. *CGIP*, 18(1):1–17, January 1982.

[14] R. Ramamoorthi and J. Arvo. Creating generative models from range images. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 195–204, New York, NY, USA, 1999. ACM Press Addison-Wesley Publishing Co.

[15] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, pages 1199–1204, 2000.

[16] A. Sharma and A. Namboodiri. Projected texture for object classification. In *Proceedings of the European Conference on Computer Vision*, 2008.

[17] A. Sharma, N. Shobhit, and A. Namboodiri. Projected texture for hand geometry based authentication. In *Proceedings of CVPR Workshop on Biometrics*, 2008.

[18] M. Takeda and K. Mutoh. Fourier transform profilometry for the automatic measurement of 3-d object shapes. *Applied Optics*, 22(24), 1983.

[19] M. Weber. *Unsupervised Learning of Models for Object Recognition*. PhD thesis, California Institute of Technology, Pasadena, 2000.

[20] M. Weber, W. Einhauser, M. Welling, and P. Perona. Viewpoint-invariant learning and detection of human heads. In *Proceedings of 4th IEEE International Conference Automatic Face and Gesture Recognition, FG2000*, pages 20–27, 2000.

[21] R. White and D. Forsyth. Combining cues: Shape from shading and texture. In *Proc. of CVPR*, volume 2, pages 1809–1816, 2006.

[22] L. Zhang, B. Curless, and S. M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. *3DPVT*, pages 24–37, 2002.