PROJECTED TEXTURE FOR 3D OBJECT RECOGNITION

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (by Research)
in
Computer Science

by

Avinash Sharma
200505004
avinash_s@research.iiit.ac.in

International Institute of Information Technology
Hyderabad, India
July 2008

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Projected Texture for 3D Object Recognition" by Avinash Sharma, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Dr. Anoop. Namboodiri

To my parents

Acknowledgments

I would like to thank Dr. Anoop Namboodiri for his support and guidance during the past two and half years. I gratefully acknowledge Dr. Anoop Namboodiri for long hours of discussions on the problems presented in the thesis. I am also thankful to all other faculty members of our lab and institute for providing their invaluable support during this work.

I am also grateful to fellow lab mates at the CVIT, IIIT Hyderabad for their stimulating company during the past years. Above all I am thankful to my family members and friends for their support and love.

Abstract

Three dimensional objects are characterized by their shape, which can be thought of as the variation in depth over the object, from a particular view point. These variations could be deterministic as in the case of rigid objects or stochastic for surfaces containing a 3D texture. These depth variations are lost during the process of imaging and what remains is the intensity variations that are induced by the shape and lighting, as well as focus variations. Algorithms that utilize 3D shape for classification tries to recover the lost 3D information from the intensity or focus variations or using additional cues from multiple images, structured lighting, etc. This process is computationally intensive and error prone. Once the depth information is estimated, one needs to characterize the object using shape descriptors for the purpose of classification.

Image-based classification algorithms try to characterize the intensity variations of the image for recognition. As we noted, the intensity variations are affected by the illumination and pose of the object. The attempt of such algorithms is to derive descriptors that are invariant to the changes in lighting and pose. Although image based classification algorithms are more efficient and robust, their classification power is limited as the 3D information is lost during the imaging process.

Our problem is to find an image-based recognition method, which utilize the shape of the object, without explicitly recovering the 3D shape of the object. This implicitly avoids the high computational cost of shape recovery while achieving high accuracies. The method should be robust to view variation, occlusion and also should invariant to scale and position of the object. It should also handle partially specular and a texture-less object surfaces.

We propose the use of structured lighting patterns, which we refer to as projected texture*, for the purpose of object recognition. The depth variations of the object induces deformations in the projected texture, and these deformations encode the shape information. The primary idea is to view the deformation pattern as a characteristic property of the object and use it directly for classification instead of trying to recover the shape explicitly. To achieve this we need to use an appropriate projection pattern and derive features that sufficiently characterize the deformations. The patterns required could be quite different depending on the nature of the object shape and its variation across the objects.*

Specifically, we look at three different recognition problems and propose appropriate projection patterns, deformation characterizations, and recognition algorithms for each. The first category of objects are of fixed shape and pose, where minor differences in shape are to be used for discriminating between classes. 3D hand geometry recognition is taken as the example of class of objects. The second class of recognition problem is that of category recognition of rigid objects from arbitrary view points. We propose a classification algorithm based on popular bag-of-words paradigm for object recognition. Third problem is that of 3D texture classification, where the depth variation in surface is stochastic in nature. We propose a set of simple texture features that can capture the deformations in projected lines on 3D textured surfaces.

The above mentioned approaches have been implemented, verified, tested, and compared on various datasets collected as well as available on the Internet. The analysis and comparative results demonstrate significant improvement over the existing approaches, in terms of accuracy and robustness.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this new world of technological advancement, where machines are playing an active role in real life, enabling them with the capability to perceive an object is an important task. Object recognition is an interesting problem that finds its roots in Artificial Intelligence. Since we visually perceive most of the objects around us, Computer Vision is one the most promising tools to address this problem. The human object recognition process is only partially understood, and the research community has tried various approaches to mimic this extraordinary ability using machines.

We have a long history of partially successful and encouraging attempts to tackle this problem. Initial attempts concentrated on classification based on structural descriptions from segmented objects. As the computation power increased, approaches that explicitly recover the 3D information of the objects became feasible. One such approach involved the use of structured lighting during imaging to aid the shape recovery. The idea of structured light based approaches is to use a controllable light source to project a custom patten on the target surface.

As mentioned above, previous attempts to use structured light for object recognition concentrated on explicit recovery the 3D of shape of the object and using it for recognition. Recovering shape is an computationally intensive and error prone process. In this thesis, we propose a novel and efficient way of using structured lighting to solve challenging problem of object recognition.

1.1 Role of Shape in Object Recognition

Our objective is to use the inherent shape information of any 3D object/surface for the purpose of classification/recognition. The problem of recognition using 3D shape is well attempted in past, still an optimal solution to the problem is not yet proposed. Although the proposed solutions are reasonably successful for a certain classes of objects, most of them lacks the computational efficiency and robustness of direct image based classification approaches.

Shape is an important clue for recognition/classification of 3D objects. In Figure 1.1, first row shows shows three distinct objects with similar appearance in the image, but different 3D shape. In contrast, when we see the same figure column wise, three images of the same object with view variations, looks very different. Clearly, it will be difficult to recognize such objects from their appearance only.

This thesis proposes a novel method of recognition of 3D object/surface using the shape information, efficiently, avoiding the complex and error process of recovery of

Figure 1.1: Example showing variation in appearance in three class of objects having different 3D shapes. Column wise each object look different, showing high intra class variation. In first row all 3 objects looks remarkably similar, showing 'inter class similarity.

3D model of the object/surface. The emphasis is on minimizing the computational complexity and financial cost while making the approach robust to view variations and occlusion.

1.2 Application of Object Recognition

Humans have the ability to recognize objects irrespective of environmental conditions like viewpoint,illumination, occlusion and clutter. Other than appearance and context information, the 3D shape of an object is the most important information that we use for recognition, thanks to our natural stereopsis and also focus adjustment capability of the eye lenses. Human have developed the ability to efficiently use both appearance and shape information for robust recognition of more than 10,000 categories of objects [22].

With new generation digital imaging technology, it is both easy and inexpensive to capture and store the visual data in electronic form. Computer provides a powerful means of processing such data. But the biggest problem from the point of view of object recognition is the loss of shape information as well as the appearance variations introduced during the imaging process. The way a human perceive the 3D world around him and store is extremely sophisticated and difficult to imitate with existing technology. In order to equip the computer with a good recognition algorithm we need a simple representation of 3D shape information, avoiding the computationally complex process of shape recovery, storage and comparison. There are many real-world applications for machine vision even with the current state of advance in recognition technology. An improvement in the robustness and accuracy of the current recognition algorithms can greatly benefit many such applications.

- Pick & Place Robots
  Many industries are using automated robotic arms for various assembly line applications. It is really important for a robotic arm to have the capability to recognize the industrial part robustly, especially when an arm has to deal with

multiple parts. A robust algorithm that can identify the objects is essential in such a situation to reduce manufacturing errors, as well as to avoid damage to the parts and the robot itself.

- Robot Navigation
  Autonomous mobile robotics is another important area, where recognition of objects is critical for success. It can be battle field or a home servant such machines could be deployed in a wide variety of situations. Inferring about the 3D world around is one the most important factors while navigating through real world environment.

- Biometric Authentication
  Security is a major concern in current global scenario, and Authentication plays and important role in security. The most comprehensive way of establishing one's identity is through biometric authentication, where a person's physical characteristics are used for establishing his or her identity of an individual. In civilian and commercial applications, one of the most important concerns is the ease of use and acceptability among users. Hand geometry based authentication is one such application, where the shape of the palm and fingers are used to identify a person.

- Surveillance
  Automated surveillance is another aspect of security, where identification of objects plays an important role. Algorithms that can monitor a large set of security cameras and signal the presence of specific objects/people in a scene can be crucial to a good surveillance system.

- Automated billing
  Supermarket billing can be another interesting real world application where 3D object recognition can be used. Veggie Vision [23], is one such attempt to automatically recognition vegetables in a supermarket basket.

There are large number of applications where machine vision can be applied in real world. Many solutions to the problem of object recognition have been proposed. However, most of them are not practical, due to large number of inherent problems explained in next section.

1.3 Challenges in Object Recognition

Some of the prominent challenges in the area of object recognition are listed below :

- Reflectance of the surface
  Reflectance of any surface is an important physical property of any object. The surface with high reflectivity causes specularity problem, both in normal image based algorithms and structured light based algorithms (see Figure 1.2).

- Lack of texture
  Lack of texture on object surface is another challenge that makes recognition a difficult task. Inferring the object shapes through other means can make a critical difference to the success of an algorithm (see Figure 1.3 for texture less objects).

Figure 1.2: Example of highly specular surface. Courtesy [1]

Figure 1.3: Texture less objects, Courtesy [2]

- Transparent and Dark surfaces
  This is another variant of the previous challenges. When an object surface totally or partially pass/absorb the light, it is difficult to infer the shape or identity of the object (see Figure 1.4 for translucent objects).

- Illumination Variation
  Lighting change is one the common problem in object recognition. Variation in environmental illumination causes large variations in the intensity values of pixels. Some common issues related to illumination variation are formation of shadow, non-linear transformation in pixels, scaling and shifting due to the

Figure 1.4: Transparent and Translucent surfaces with occlusion

Figure 1.5: Effect of Illumination Variation on face. Courtesy [3]

change in position of light source. Figure 1.5 illustrate face images with wide range of illumination variation.

- Viewpoint Variation
  Viewpoint is an important factor and causes different transformations like in-plane transform such translation, rotation, scaling and skew, and out-of-plane transformation like projective transform. Knowledge of the viewpoint can also be exploited while handling certain applications. Figure 1.6 shows large view variation in a 3D object.

Figure 1.6: Coffee cup with large view variation.

- Occlusion
  Visibility of some part of object can be hindered due to some other object in vicinity of the current object or due to other parts of same object. The latter phenomenon is known as self-occlusion.

- Noise
  There could be noise in image acquisition process, resulting a degraded image causing loss of information. Noise in 3D shape recovery is another important factor for shape based recognition approaches.

- Scale Variation
  Variation in scale of an object can be another factor and is directly related to viewpoint. It can also be there due to inherent scaling in shape of the object. (see Figure 1.7, shows scale variation in balls)

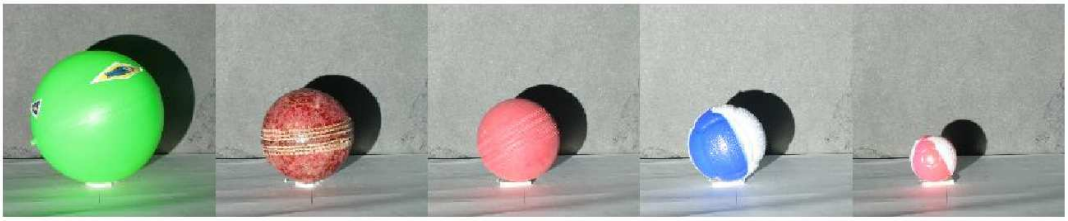


Figure 1.7: Scale variation in balls due to different sizes

- Background Clutter
  Segmentation of object is difficult many times due to highly complex background. This is critical for algorithms that rely on properties of segmented objects for recognition. Multiple object with occlusion creates the clutter more difficult to handle.

- Intra-class Variation
  There could be large variations among the samples of same class making it difficult to define class boundaries. Each column in Figure 1.1 shows high intra class variation for each of object class.

- Inter-class Similarities
  High inter class similarity is another problem when you have large number of classes with objects having very similar appearance with objects of other classes. First row in Figure 1.1 shows similar appearance for three different class of objects.

We need approach that can address most of these challenges at the same time keeping the solution computationally efficient so as to make it practically usable. Most of the existing approaches discard the 3D information because it is computationally inefficient to use it. We will outline the possible and attempted approaches in this area with and without 3D information in next chapter.

1.4 Our Contributions

In this thesis, we have proposed a solution to the 3D object recognition problem using *Projected Texture*. Figure 1.9 shows an hierarchical representation of the sub classes in object recognition. We have attempted three different subclass of object recognition as marked with double boundary in figure and proposed three different features to capture deterministic as well as stochastic deformations. Short note on each of the contributions is given below :

- Projected Texture for Recognition
  We proposed the concept of *Projected Texture*. We project pattens on 3D surface of interest for the purpose of recognition/classification. Combination of projected pattern and original texture of the target surface results in a deformed texture. These deformation in resultant texture essentially encode depth variation as well as information about physical property (reflectivity) of the target surface. Figure 1.8 shows a typical projected pattern setup.

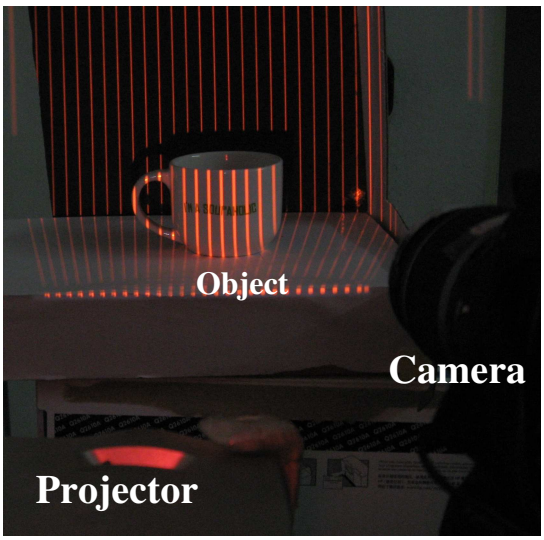


Figure 1.8: Experimental Setup for projecting a pattern on object and capturing the deformed texture image.

- Gabor filter based Feature for Fixed Pose Object Recognition
  We have proposed a window based Gabor feature for the class of objects, where the pose of the object is considered to be fixed. First, the image of the object with deformed projected texture is divided into a set of sub-windows. Then we employed Gabor filter to characterize each of these sub-window, which captures the local frequencies and their orientations.

- Fourier Domain Feature for Arbitrary Pose Object Recognition
  The solution to category recognition problem is proposed for rigid objects with arbitrary pose by extracting a 2-D Fourier transform based feature on top of concept of "bag of words" approach. 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.

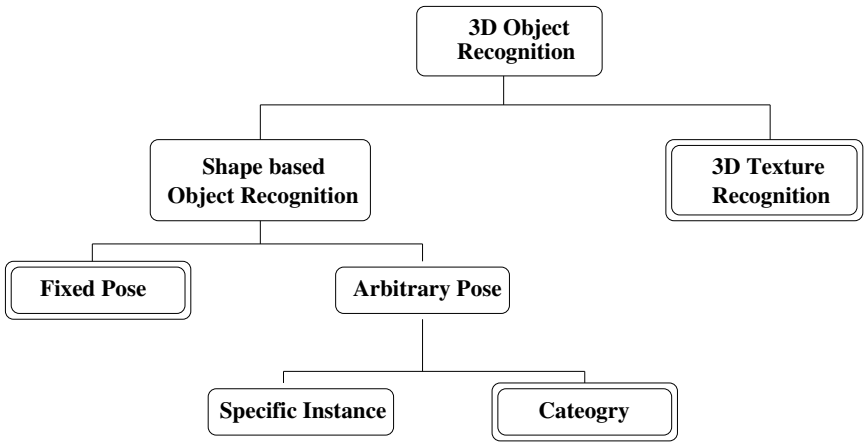


Figure 1.9: Hierarchy of 3D Object Recognition

- NHoGD Feature for 3D Texture Classification
  We also propose a set of texture features that captures the deformation statistics from stochastic depth variations present in 3-D textured surfaces.

1.5 Thesis Overview

The remainder of the thesis is organized as follows. In chapter 2, we have presented a detailed survey of the literature in the area of object recognition, shape recovery, texture classification and shape based biometrics. Chapter 3 gives a detailed description of proposed concept of projected texture. It includes the original idea and mathematical foundations of measuring projection deformations. In next three chapters , we present three different features for each of the different classes of object recognition, mentioned before. First, we demonstrate fixed pose base object recognition task with hand geometry based authentication, and later the proposed solution for recognition of 3D rigid object and classification of 3D texture. Chapter 7 presents the conclusions drawn from this thesis and also explores some of the possible avenues for future work.

Chapter 2

Preliminaries and Related Work

In this chapter we have presented a survey of existing techniques in area of *object recognition* and *texture classification*. We have discussed prominent works in these areas with a brief summary of historic background. Detailed literature survey on both the areas can be found in PhD Thesis of Robert Fergus [24] and Manik Varma [11]. We attempt to provide only the basic overview and closely related works, as the literature in this area is quite extensive.

We start with a summary of depth recovery techniques used for explicit recovery of object shape, for model based object recognition. Then, an introductory section on structured light approaches is presented. Later on, we have discussed important approaches to object recognition and texture classification.

2.1 Depth Recovery for Model based Recognition

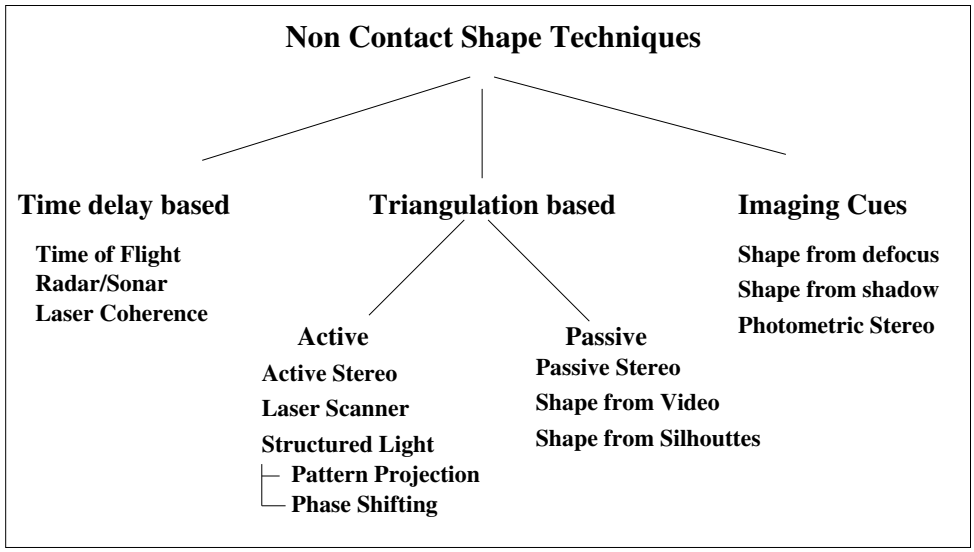


Figure 2.1: Hierarchy of various 3D Shape Recovery Approaches

There are different approaches that recover the shape of any object/scene. Figure 2.1 shows a hierarchical representation of the existing approaches for 3D reconstruction. There are mainly three classes of depth recovery approaches:

- The first category of approaches are time delay based approaches, where a trans-receiver system computes the delay or any deterioration in reflected signal to infer the depth at a point. Sonar and Laser coherence based depth computation are examples of such approaches.

- A second class of approaches works on a geometric formulation to infer depth, known as *triangulation*. A large number of active and passive triangulation approaches exist, and the prominent ones are shown in Figure 2.1.

- The last one are the shape from X approaches, where X can be stereo, texture, shading, motion and defocus. A detailed description of these approaches can be found in Appendix A.

We now take a more detailed look at the structured light based approaches.

2.1.1 Structured Light based Approaches

Structured light based approach involves use of a controllable light source with a camera, instead of the traditional stereo kind of setup, where two cameras are used. The projector can be considered as inverse camera, obeying exactly the same projection and distortion model as a regular camera. There are many problems in this inverse camera assumption. One of the major issue is that in contrast to a camera sensor, which will react almost equally for all pixels in case of a perfect uniform illumination, in a projector there is a strong (typically a factor of 2 or more in luminance) center to fringe fall-off, also known as hot-spot effect. As a consequence, a uniform pattern will not generate a spatially uniform illumination. Structured light applications can be divided into following classes:

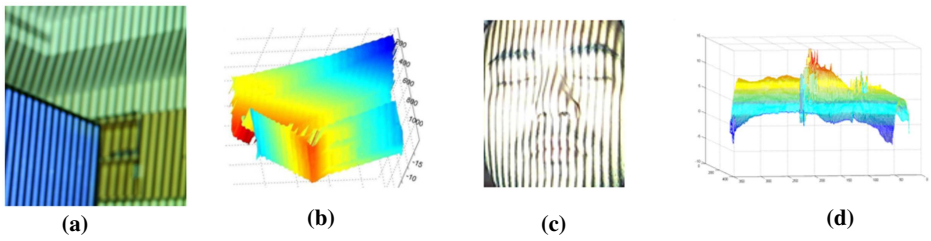


Figure 2.2: Reconstruction Results with FTP [4]. (a) the corner of a roof (b) recovered 3D depth map of the roof corner (c) the face of a person (d) reconstructed depth map of the face

- 3D Reconstruction
  This class of approaches tries to recover the 3D model of the object. Takeda *et al.* [4] proposed the technique of Fourier transform profilometry, which exploits the phase variations in a deformed grating to generate a depth map of the object under illumination. Figure 2.2 shows the reconstruction results of obtain from our FTP implementation. Figure 2.2(a), Figure 2.2(c) shows original scene/object and Figure 2.2(b) and Figure 2.2(d) shows corresponding reconstruction. The approach suffers from occlusions and is also dependent on the ability to single out the projected frequency even in presence of object texture.

Figure 2.3: Real-time hand reconstruction using multi-pass dynamic programming. Courtesy [5]

- Improving Stereo Correspondence
  Large number of approaches have been proposed in this class. Here the goal is to solve the correspondence problem in stereo by projecting structured light. The difference in pattern projection techniques lies in the way in which identification of every point is performed in the pattern. The specific type of codewords involved and the axis encoded by them is crucial in these techniques. Mainly three classes of pattern projection methods for correspondence improvement exist:

  – Time-multiplexing based
  – Spatial neighborhood based
  – Direct coding based

  In time-multiplexing techniques (proposed in [25, 26]), structure of every pattern can be very simple, since codeword generation is done by projecting a sequence of patterns, one after the other. Neighborhood coding techniques project a pattern on the scene such that any neighborhood in the image receives a unique pattern of colors. The patterns used have nearly same complexity as that of previous one [27, 28, 5].

  Figure 2.3 shows the shape acquisition results of the neighborhood coding method illustrated in[5]. Finally, direct coding techniques uses gray level or color value of pixel to define a codeword for that pixel. Salvi and Pags [29] has given a thorough account of different coding scheme for generating various patterns.

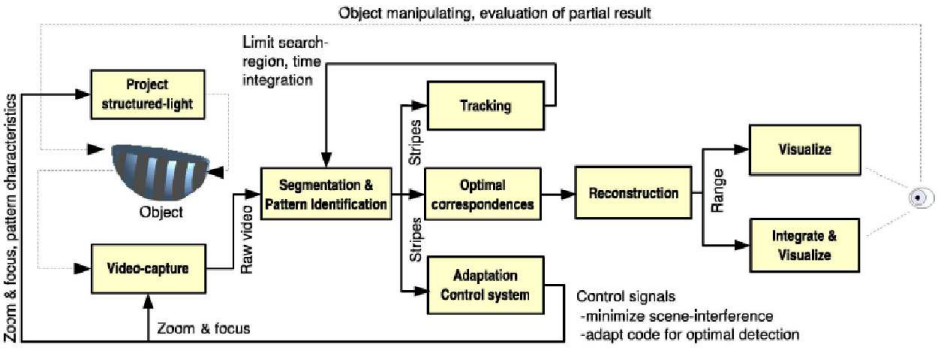


Figure 2.4: Single frame adaptive structured light proposed in [6].

- Adaptive Structured Lighting
  In many situations, the nature and complexity of the projected pattern might depend on the nature of the objects in the scene, which can be tackled by adapting the pattern to the scene under consideration. The primary idea here is to project a structured light and use a feedback mechanism to improve the pattern design, so as to achieve better performance of the reconstruction or correspondence computation system. A detailed study of the area can be found in PhD work of Koninckx [6]. Figure 2.4 shows an overview of the single structured light system proposed in [6].

- Feature Projection
  Attempts have also been made, where the structured light was used to project some features on the surface lacking natural texture. This can be used for for different task like augmented reality, tracking [30], visual servoing [31] and recognition. Here, the central theme is to create an artificial texture and then use traditional image based features for different tasks that require object feature points. Note that this is different from our approach as we do not require the detection of any feature point on the object.

2.2 Approaches to Single-Instance Object Recognition

In this section we will discuss some of the important object recognition methods in the literature. First, the traditional methods are presented. Then we have discussed the "Part Based Model" in detail. We have used a similar approach in chapter 5, while proposing solution to 3D object recognition for rigid objects with arbitrary pose. At the end, we will give a brief introduction to the texture-based recognition techniques.

2.2.1 Traditional Approaches

Initial work on object recognition was focused on single object instance recognition. At first, the general approach was to recover 3D and use it for direct comparison with a stored model. For example, Brady *et al.* [32] used Gaussian curvature to extract edges and segment the image into planar regions for matching with a 3-D CAD model.

(a) (b) (c)

Figure 2.5: Recognition results by Lowe [7]. (a) Object image, (b) Query image and (c) Image with recognized object with represented by rectangle boundary

Geometric Approaches

Later on, approaches directly worked on intensity images and used a geometric representation like extracting edge contours from the interior and exterior of the object. This enables us to tackle the problem of illumination changes and avoids the computation of 2-D or 3-D pose. Some methods try to find geometric invariant solutions, where the feature vector derived can be shown to be independent of the pose. However, the class of features that possess this property is limited. A viewpoint independent descriptor was proposed using a small set of points in image in [33]. These points used as key for hashing, while searching a database of models.

Other proposed a skeleton-based representations like aspect graphs and geometric primitives, and the use of stereo for recovering 3-D wire-frame models (*e. g.* , Pollard *et al.* [34]). Some of the good attempts for better alignment techniques and improving search presented in [35, 36]. Representing the 3D model was attempted as a mixture of 2-D models in [37] or with set of primitives like cylinders and cones in [38]. Reisenhuber *et al.* [39] explores the view based representation. Rothwell *et al.* [40] proposed a projective invariant recognition system for planar objects.

Global Appearance based Approaches

The Global appearance based approaches models the overall appearance of the object. The use of histograms over the joint statistics of the local appearance as the vectors of local shape descriptor is proposed by Schiele *et al.* [41]. These histogram based representations also incorporate spatial location of the descriptor. A 3D version of the eigenspace methods was proposed by Murase *et al.* [42]. Other approaches include the use of standard classifiers such as Support Vector Machine (SVM) on COIL dataset [43], proposed by Pontil *et al.* [44].

In summary, most of the global appearance approaches are simple and sensitive to background clutter and occlusion. Although sub-window type measures can be used to overcome their susceptibility, modeling power is wasted in some sense by modeling the-not-so useful part of the scene.

2.2.2 Texture Region based Recognition

In texture region based approach, texture regions were selected using a region selector and represented with an appropriate of descriptor. Now these descriptors can be used for matching with previously learned database of objects of interest. Operators like Harris points detector were used for region selection at multiple scales, as proposed in [45].

Lowe [7] proposed a real-time, robust and flexible quasi- affine invariant recognition scheme. A difference of Gaussian (DOG) feature detector was used to extract a large set of regions from the query image. Similarity invariant SIFT features were used to represent these regions to make the approach robust [46]. A Hough transform based voting scheme is used with Nearest Neighbor classifier to do matching with trained set of descriptor values. Figure 2.5 shows the results of [7]. Later, [47, 48, 49, 50, 51, 52] refined the concept for better performance. Some refinements were proposed by using invariance to affine instead of similarity transform and by incorporating spatial information of vectors within the image into the matching process. Texture based approaches are superior than geometry based approaches and are robust to

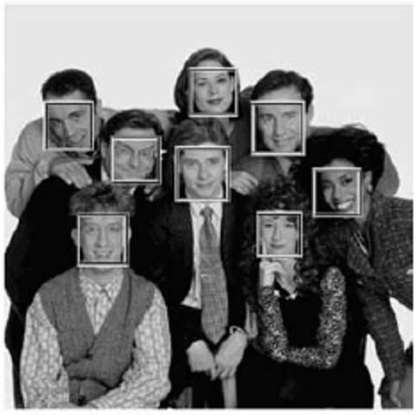


Figure 2.6: Real-time face detection results as presented by Viola and Jone [8].

background clutter, utilizes modeling power as compared to global appearance based approaches. However, they are texture dependent.

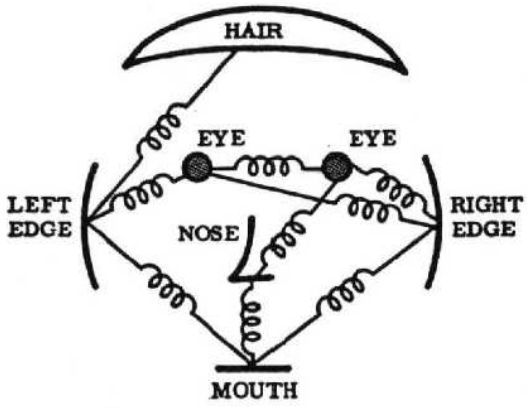


Figure 2.7: Part of structure model proposed by Fischler *et al.* [9].

2.3 Approaches to Category Level Object Recognition

Part and Structure Model, proposed in 1973 by Fischler *et al.* [9], consist of series of small templates as the parts, arranged in some geometric configuration as the structure. Here, model fitting was a cost minimization process comprising the local fit for each of the parts plus a global deformation term, measuring the deviation from a rest position. Work by Schneiderman *et al.* [53] presented a multi-view capable wavelet based approach where a histogram of wavelet transform coefficients were used for probabilistic classification. Viola nd Jones [8] gave an important contribution by successfully applying *boosting* a machine learning concept, on an exhaustive set of features to get a strong classifier. The cascaded structure proposed here helps to do selective feature/classifier selection. Figure 2.6 shows results of face detection presented in [8]. Recently, Dalal *et al.* [54] presented Histogram of Oriented Gradient

in the same direction. Mikolajczyk *et al.* [55] proposed a probabilistic assembly of robust part detectors to find humans, where body parts like face, torso and limb detectors are trained together in a discriminative manner. Their combination of detectors results in state-of-the-art performance.

Most of these approaches work only for a limited set of categories as compared to the human capability of identifying around 10,000 categories [22]. Recent work focuses on algorithms which can be applied to all categories by modeling the intra-class variability. This can be done initially in a constrained viewpoint manner, and can be extended to 3D recognition using a series of 2D models [37]. Databases like Caltech [56] or the UIUC [57] reflects a shift in effort in this direction. Most of the recent work represents the object as a collection of textured patches, each with varying details like number of patches, the detection and representation of parts and their spatial location, and how the variability in appearance is handled in a matching algorithm. These methods falls in three broad categories: generative, discriminative or hybrid. These approaches apply feature selection for automatic selection of discriminating features. Another choice that differentiate the approaches are using object localization or classifying image as a whole.

Appearance only methods

While considering appearance only methods, Csurka *et al.* [58] proposed a straightforward "Bag of Keypoints" model. On the other hand Opelt *et al.* [59] adopted AdaBoost, for category learning. All these methods are robust to noise and clutter as well as provide view invariance, but required identification of key points or object localization.

Incorporating shape information

Originally, the idea was presented in [9] on part and structure model as shown in Figure 2.7. The challenge here is to find sufficient location information to incorporate into model to make it useful without introducing computational complexity. Parona *et al.* [60, 61, 62] proposed a working model of this idea in form of a Constellation Model. This model is a probabilistic model that also handles missing features and background clutter with minimum supervision.

Weber *et al.* [10] developed their approach on top of Burl *et al.* [63] to make the training process unsupervised and achieved best results so far reported with high level of robustness to clutter and occlusion. They automatically obtained a set of potentially useful pixel patches by running an interest operator on the training set; chopping out patches around each interest point and then using k-means clustering on the patches. Mixture models are introduced in [62] for automatic category recognition. Figure 2.8 shows the overview of method proposed in [10].

Later, Fei-Fei *et al.* [64] applied a powerful machine learning method to the Constellation Model by introducing a hierarchical Bayesian version of the Constellation Model. This model is able to incorporate priors into the learning procedure in a principled manner and thus considerably decreasing the number of training images required. Felzenszwalb *et al.* [65] make it more efficient at level of image features matching with parts in the model. A tree-structured model was used to model the dependencies in the spatial relationship between parts. They used it for detecting people in images.

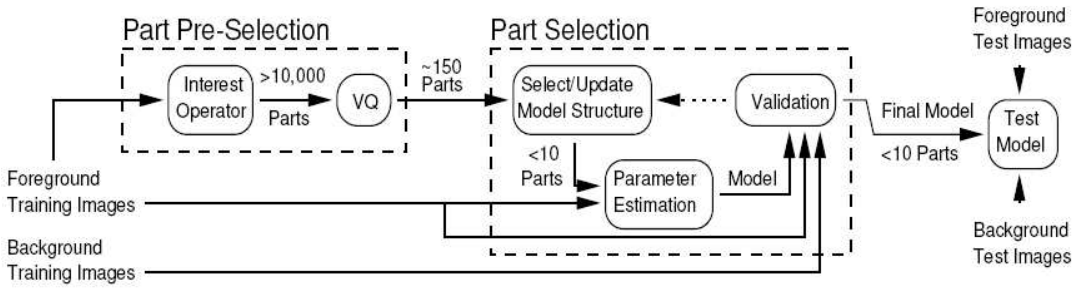


Figure 2.8: Overview of Weber's Approach [10]

In this work, we have used the 'Bag of Keypoints' model proposed by Csurka *et al.* [58] for learning the categories of rigid object with arbitrary pose in chapter 5. In [58], during training, a large set of regions are extracted from each image and vector quantized to a set of predetermined clusters. The image is then described by a feature vector listing the number of regions belonging to each cluster. This vector is labeled according to their class (object present/ object absent) and, along with vectors from all other training images, used to train an SVM. Recognition proceeds in the same manner, except that in the last step, the SVM is used to predict the class label. They got surprisingly good results even though intuitively, location information would seem to be an important part of recognition. Figure 2.9 shows overview of a "Bag of Words Model" proposed in [64].

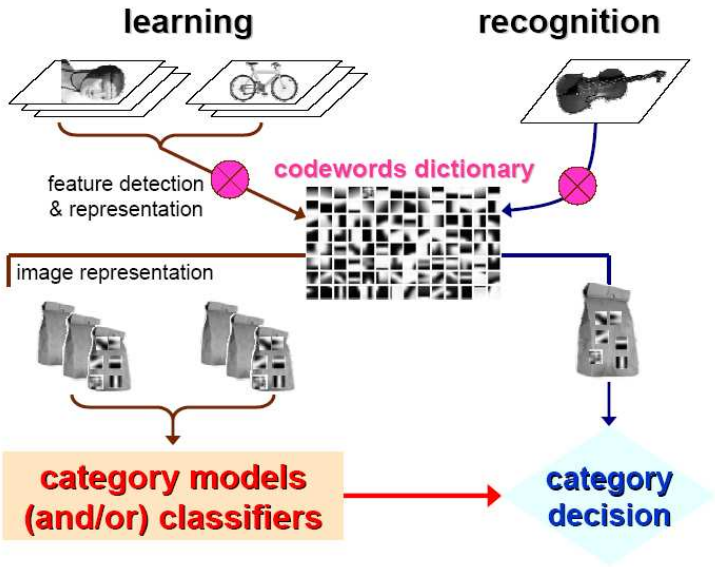


Figure 2.9: Overview of Bag of Words Model (Courtesy Li Fei-Fei CVPR07 tutorial)

Crandall *et al.* [66] used efficient method of Felzenswalb *et al.* [65] to find relatively simple and low computationally complex shape based part and structure model. They tested it on the classes other than humans. Agarwal *et al.* [67] demonstrated car classification using sparse-network-of-winnows *SNOW* classifier [68] and tested it on the UIUC dataset [57]. Borenstein *et al.* [69] combined object classification and segmentation to propose a scheme, which modeled the object with small set of image fragments.

Leibe *et al.* [70] demonstrated a leading approach, beating many methods on a wide range of datasets [71]. Work by Murphy *et al.* [72] allowed scaling of object recognition up-to many hundreds of categories. Berg and Malik proposed correspondence computation between model and features as integer quadratic programming problem for deformable shape matching.

Work by Robert Fergus [24] builds on top of Weber *et al.* [10] and model the appearance using probability density functions, giving a probability for each match. More importantly, the entire representation is made probabilistic, enabling the learning and recognition tasks to be posed as machine learning problems.

2.4 Texture Recognition/Classification

Classification of 3D textures is another problem that we have attempted in this thesis. We will now look at the important works in 2D and 3D texture classification. At the end we will give a brief summary of the state of the art work in this field by Varma *et al.* [11].

2.4.1 2D Texture Classification

Early work on texture starts with Julesz's conjecture that "two textures were perceptually indistinguishable if they had identical second order statistics" [73], and later replaced by a statistical theory of textons [74]. The theory postulated that textons were fundamental texture primitives. In late eighties, filter banks were experimentally being used for texture analysis [75, 76, 77, 78, 79, 80, 81]. The work in early nineties calculated filter responses at all possible orientations and scales from a small basis set [82, 83, 84].

The computational limitations of initial era put the early filter bank based methods constrained, to use lower order moments in-order to characterize the distribution of filter responses. Rectification, energy measurement or conversion to a rotationally invariant frame, was done on top of filter responses, instead of direct using filter response as a feature vector . A classifier of choice was then trained on the feature vectors and used to classify novel images. Some examples are works by [85, 86, 87].

From the mid nineties onwards, filter bank and wavelet based methods became increasingly successful for solving texture classification and synthesis problem. The representations were richer because, full filter response distributions were used and the joint distribution, or co- occurrence, of filter responses was learned. Also the number of filters and wavelets used, kept increasing so as to compute features at multiple scales and orientations.

As a significant contribution, Leung *et al.* [88, 89] gave an operational definition of a texton, based on filter responses and clustering. They defined a 2D texton as a cluster center in filter response space. This not only enabled textons to be generated automatically from an image, but also opened up the possibility of a universal set of textons for all textures.

In [90, 91] Dana *et al.* proposed a system that addressed some of the major shortcomings of Leung and Malik's algorithm. They demonstrated that 2D textons (learned from filter responses of single images instead of image stacks) could themselves be used for uncalibrated, single image classification without compromising on performance.

Zisserman *et al.* [92, 93] developed somewhat a similar approach. For the problem of reducing the number of models required to represent a texture. One was the Geometric approach, which was focused on building an invariant texture descriptors. A global normalization by maximizing the weak isotropy of its second moment matrix was also proposed. Full invariance can then be achieved by using a scale and rotation invariant filter bank to extract features. Some approach of model reduction uses Machine Learning to select a subset of the models, while maximizing some criteria of classification and generalization. Work by [94] shows the use of the nearest neighbor classifier used in [93], being replaced by a Support Vector Machine (SVM).

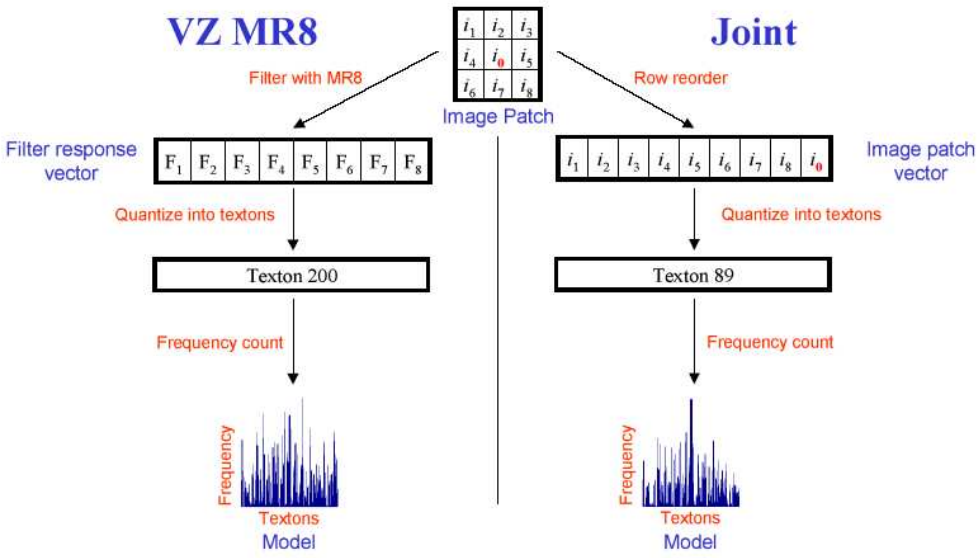


Figure 2.10: Overview of VZ and Joint Classifier Algorithm proposed by Varma *et al.* [11]. The left side of image illustrate steps of VZ algorithm, which uses filter bank response, while the other part of image shows the patch based Joint algorithm.

Using Optimal Filtering

Instead of choosing a filter bank heuristically, an optimal filter bank can be obtained by optimizing specifically for the given classification task. For example, [95, 96, 97, 98, 99, 100, 101, 102, 103, 104] methods tried to optimize the filter bank by optimizing different functions. Discriminant analysis is another way of finding optimal filter bank. [105, 81, 106] presented three different optimization criteria by making different assumption about the underlying distribution, which generated the filter responses. On the other hand [107, 108] proposed optimization methods, which are embedded into a neural network framework.

Recent Progress

Recently, filter bank and wavelet methods are challenged by MRF and image patch methods. [109, 110] used MRF successfully for texture synthesis. In [11], Maximum Response (MR) filter based approach is proposed. Also alternative patch based

Joint classifier and MRF classifier (uses neighborhood property) were proposed and reported good results. Later many heuristic and machine learning based approach were proposed to reduce the number of models used to represent a texture class. Figure 2.10 shows the overview of MR8 filter response based VZ algorithm and patch based Joint classifier algorithm. Some details of this work are discussed in later part of this work.

2.4.2 3D Texture

Until mid nineties synthesis and classification algorithms treated textures as pure albedo patterns painted on a plane surface. According to this, a single image could completely characterizes all the possible variations of a texture patch. Later it became apparent that such 2D texture models were not very physically plausible, as they ignored all 3D effects including surface normal variations, BRDF variations, illumination changes, scale and perspective, etc.

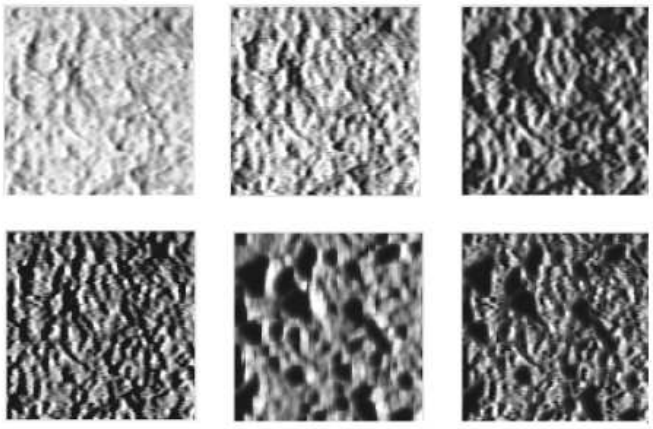


Figure 2.11: Synthesized texture results presented by Dana *et al.* [12].

Nayar *et al.* [111] proposed and validated physical models that predict a texture's intensity distribution under varying viewpoint and illumination. Later, Dana *et al.* [112] predicted the change in correlation length of the textured rough surface with viewing direction. Work in [113, 114, 115] provided valuable theoretical insights into how the variances of filter responses change with the illuminant's tilt and showed that a statistical description of surface roughness was also sufficient to estimate the illuminant's tilt direction from single images. All these models are theoretically appealing but they did not translate into practical classification algorithms because of their restrictive assumptions like uniform albedo, Lambertian surfaces, inability to model shadows, occlusions, specularity, etc. [116, 117] proposed physical BRDF models and lead to good results in synthesis.

In [12] Wang and Dana proposed concept of Hybrid Texton; a texture representation that integrates the appearance-based information from the sampled bidirectional texture function (BTF) with concise geometric information inferred from the sampled BTF. The model is a hybrid of geometric and image- based models and has key advantages in a wide range of tasks, including texture prediction, recognition, and

synthesis. Figure 2.11 shows synthesis results from [12].

2.5 Shape based Biometrics

A third problem that we address in this thesis is that of fixed pose object recognition. The example that we choose is that of hand geometry based person authentication. We now take a brief look at the existing hand geometry based recognition algorithms as a basis of comparison of our approach.

Measurements of the human palm, such as the length and width of fingers and the 3D palm profile are known to have some amount of identity information. Sildauskas [118] patented the first electronic hand geometry based identity verification apparatus, and several commercial systems have been developed since then. Jain *et al.* [13] outlined the challenges in such an authentication system and proposed a simple set of hand measurements, inspired by the previous work. Even the most recent hand geometry algorithms [14] used extensions of the set of features outlined in [13]. The research in 2D hand geometry based authentication has progressed primarily in three different directions:

The first set of algorithms tried to include additional measurements of palms such as area, perimeter, distances between specific feature points on the palm, etc. [14] to improve the verification accuracy. Even though the results showed improvements on the prior art, the comparisons are limited. A second direction was to integrate hand geometry along with other biometric traits to achieve high recognition performance. Fingerprints [119] and Palm prints [120, 121] are ideal candidates for this due to their ease of acquisition along with the hand geometry. A third set of algorithms look at generic techniques to improve the classification process used for verification, such as feature discretization [122], use of error correcting codes, use of more powerful classifiers [14], etc.

The use of 3D information in hand-geometry based authentication is limited to adding partial depth information computed from the profile view, usually captured using a slanting mirror[13]. The use of depth information of the hand has the potential to improve the recognition and verification performance of hand geometry based systems. The most promising approach for hand geometry based authentication is to use structured lighting to recover the hand shape [123, 124]. However, this is both, time consuming and error prone, and requires the development of shape descriptors for the purpose of classification.

Cofer*et al.* [125] proposed the use of dot patterns for computing the correspondence, and hence the depth, at specific points on the palm. The recovered depth is used along with silhouette features for recognition. Faulkner *et al.* [126] proposed the use of light stripes instead of dot pattern in-order to compute the correspondences. However, both of the above approaches aim to recover partial depth information, which in turn is used along with 2D object features for authentication.

Texture measures inherent in a biometric traits such as palm prints [120], fingerprints [127] and iris patterns [128] have been used extensively for identity verification. We proposed a similar approach for hand geometry based authentication with projected texture.

2.6 Representation Schemes

Up to this point, we have looked at object recognition algorithms from its totality. Each algorithm uses their own feature representation that characterizes the object. We provide an overview of the representation methods that are popularly used, along with their strengths and weaknesses. Variants of some of them are used to derive the features proposed in our work.

2.6.1 Simple geometric Measures

There are many approaches that employ simple geometric features like Euclidean distance between specific points of object, area, perimeter and edges for representation purpose. Two such representation scheme proposed in [13, 14] is implemented in chapter 4 for the purpose of comparison for hand geometry based authentication.

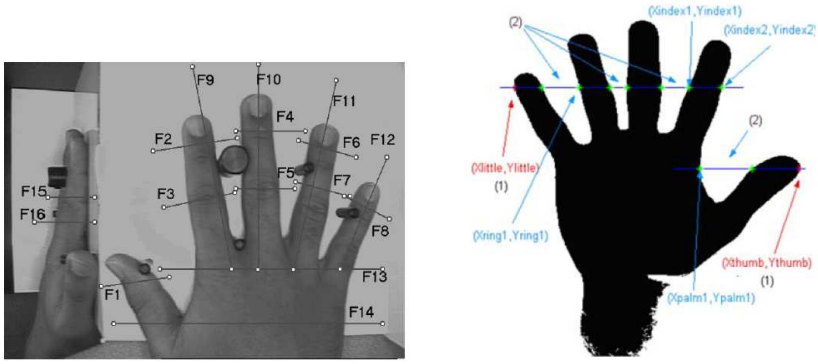


Figure 2.12: Axis defined to capture hand geometry by [13, 14]

First work extracted 16 features from 16 predefined axes as shown in Figure 2.12. The five pegs helps in choosing these axes. The hand is represented as a vector of these feature values. In [14] extra features were devised as pixel distance between a set of points on extracted contour of the hand. Feature of both the approaches is shown in Figure 2.12.

2.6.2 Filter Bank Responses

This is an important representation scheme while dealing with textures. In terms of signal processing, a filter bank is an array of band-pass filters that separates the input signal into several components, each one carrying a single frequency sub-band of the original signal. This is also called frequency analysis, as the filter bank serves to isolate different frequency components in a signal. In field of image processing and computer vision, images are thought of as a 2D signal, thus a 2D filter bank can be applied to an image in-order to detect and match the feature pattern in the image. Here, the combined strength of the filter responses at an image patch is an indicator of the similarity of the image patch to the given filter. Hence, if we assume the intensity to be mean normalized, the filter responds most strongly to patches that are scalar multiples of itself and responds least strongly (zero filter response) to patches which are orthogonal to it. Such behavior gives a analogy of dot product to filtering, which

in turn, is convolution of image with filter. This is equivalent to projecting all the patches in the image onto the vector representations of the filter.

As we have implemented texture analysis work by Varma *et al.* [11] for comparison against the proposed texture classification algorithm, we give a brief introduction to filter-bank used by them. They proposed Maximum Response 8 (MR8) filter bank from Base Filter Set (BFS) by recording only the maximum filter response across all orientations for the two anisotropic filters. Measuring only the maximum response across orientations reduces the number of responses from 38 (six orientations at three scales for two anisotropic filters, plus two for Gaussian and Laplacian of Gaussian, both also known as isotropic filters) to 8 (three at each scale for two anisotropic filters, plus two for isotropic filter). Thus, the MR8 filter bank consists of 38 filters but only 8 filter responses. Figure 2.13 shows the Maximum Response (MR8) filter bank used in [11] for texture analysis.

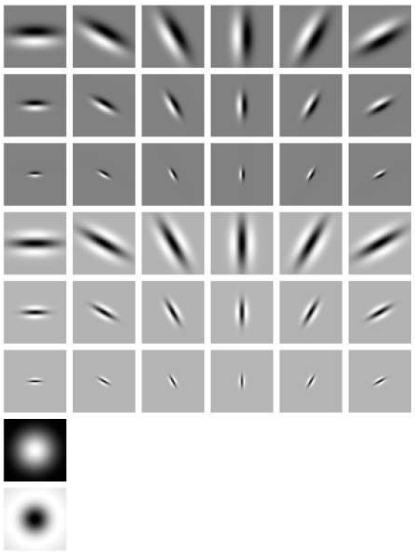


Figure 2.13: MR8 Filter Bank as proposed in [11]

2.6.3 Texture Patch Representation

Recent approaches has directly used a texture patch (the intensity values from a patch of the image) and presented comparable results with traditional filter-bank based techniques. Varma *et al.* [11] successfully demonstrated use of texture patch for texture classification by proposing Joint classifier and MRF classifier algorithms. Figure 2.10 shows the steps of patch based Joint classifier in comparison with VZ algorithm.

2.6.4 Statistical Measures

Statistical measures are one the most popular class of representation scheme in literature. A simple gray image is in true sense is a 2D stochastic signal with repeated intensity value or we can define an image as 2D distribution of data points. From simple properties of these data point like mean and variance to a complex manifold

learning all can be used as a representation scheme. The Histograms are one of the basic representation scheme in this class. It provides a compact summarization of the distribution of data in an image. The work by Manik Varma discussed in previous section also use histogram as descriptor of texture image. We will discuss two popular descriptors *HOG* and *SIFT*, which uses histogram of the filter responses or representations.

Histogram of Oriented Gradient

Histogram of Oriented Gradient descriptors (or HOG descriptors) are the feature descriptors first introduced by Dalal *et al.* [15]. The technique counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, scale-invariant feature transform (SIFT) descriptors, and shape contexts, but differs in that it computes on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved performance. Figure 2.14 demonstrates structure of HOG feature.

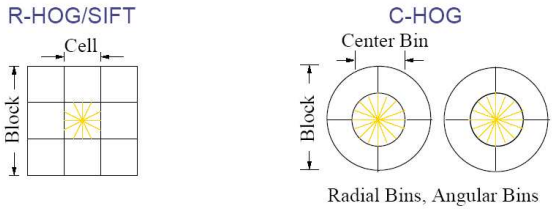


Figure 2.14: Overview of Histogram of Oriented Gradient feature proposed by Dalal *et al.* [15]

SIFT

As proposed in [129], SIFT (Scale Invariant Feature Transform) when applied to a region finds its gradients and then normalizes for orientation by finding the dominant orientation, rotating the region so as to make it axis aligned. Then 8-bin orientation histograms are formed of the gradients in each cell of a 4×4 spatial grid overlaid on the region, thus giving a $4 \times 4 \times 8 = 128$ dimensional vector descriptor. Figure 2.15 demonstrate method of computing SIFT. The gradient computation helps to achieve illumination invariance, while the loose grid gives a little bit of slack to handle minor translation and scale offsets due to inexact feature detection.

2.7 Summary of Literature

The problem of object recognition is an interesting and challenging problem in field of Computer Vision and Artificial Intelligence. The vast increase in the computation power as well as technology advancement, enabled the recent attempts to take up top-down approach for recognition. Early work either recovered 3D shape and then do a direct or improved matching with stored shape model. Stereo, shape from X and structured light techniques were used for shape recovery.

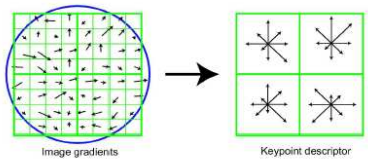


Figure 2.15: A SIFT descriptor of [16]. On the left are the gradients of an image patch. The blue circle indicates the Gaussian center-weighting. These gradients are then accumulated over 4×4 subregions, as shown on the right, the length of the arrow corresponding to the sum of the gradient magnitudes in that direction. Instead of 4×4, a 2×2 descriptor array is shown here.

Later, research community shifted to efficient image based recognition by defining efficient and robust features on 2D image. Meanwhile, part and structure models were also presented to given in-order to shift the orientation of approaches from bottom-up to top-down frameworks. Recently, more of the machine learning concepts are proposed to introduce learning into recognition framework. At the same time, attempts have been made to model the surface properties for recognition. Most of the existing work either explicitly recover a 3D shape or altogether neglect it and use only the 2D information, with few exceptions, which try to use both together.

Although a lot of attempts have been made still the goal of developing a generic, real time, robust and efficient object recognition system is too far to achieve. Our approach presented in next chapters is very much different from existing work, as we are encoding the shape information into spatial domain and then proposed specialized feature to capture these deformation for characterizing the surface. We hope that this will open up a whole set of possibilities in a variety of applications, as the 2D feature representation is enriched by the shape information. Note that one could use many of the representation schemes and classification approaches described above in conjunction with projected textures.

Chapter 3

Projected Texture

Three dimensional object are characterized by their shape, which can be thought of as the variation in depth over the object, from a particular view point. These variations could be deterministic as in the case of rigid objects or stochastic for surfaces containing a 3D texture. These depth variations are lost during the process of imaging and what remains is the intensity variations that are induced by the shape and lighting, in addition to the variations in focus. Algorithms that utilize 3D shape for classification try to recover the lost 3D information from the intensity, focus variations or using additional cues from multiple images, structured lighting, etc. This process is computationally intensive and error prone. Once the depth information is estimated, one needs to characterize the object using shape descriptors for the purpose of classification.

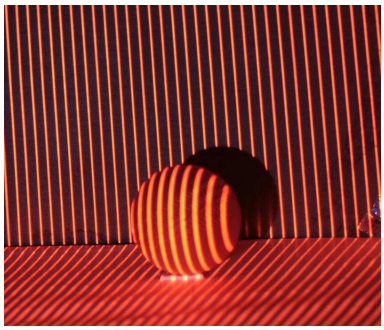


Figure 3.1: Deformation in Projected-Texture due to overall depth variation in object shape.

A second important property that characterizes an object is the inherent texture (color and albedo variations) present on the surface of the object. The object texture, along with the intensity variations introduced due to lighting and pose in an image determines the appearance of an object. Image-based classification algorithms try to characterize such intensity variations present in the image of the object for recognition. As we noted, the intensity variations are affected by the illumination and pose of the object. Such algorithms attempts to derive descriptors that are invariant to the changes in lighting and pose. Although image based classification algorithms are more efficient and robust, their classification power is limited, as the 3D information is lost during the imaging process.

An overview of the existing algorithms reveal that the shape of an object is a

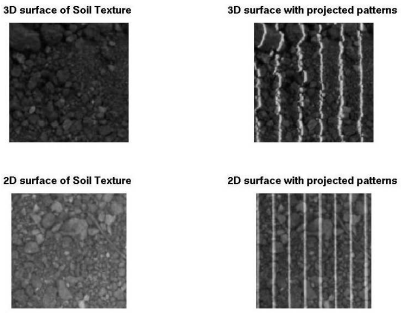


Figure 3.2: Deformation in Projected-Texture due to depth variation in surface of soil. In top row the line patterns were projected onto 3-D surface of soil, thus giving rise to deformation in pattern, as can be seen from image. In bottom row the same patterns were illuminated on image of soil texture, essentially a 2-D soil texture which does not contribute to deformation in Projected Texture.

robust characteristic, which is difficult to estimate; and the object's appearance is easier to characterize, while being sensitive to a variety of factors such as pose and illumination. To overcome these difficulties, we propose to introduce textural features in the image that are dependent on the object's shape. The primary idea is to use a structured lighting pattern, projected on to the 3D object during the imaging process, which we refer to as projected texture. The depth variations of the object induces deformations in the projected texture, and these deformations encode the shape information. One can view the deformation pattern as a characteristic property of the object and use it directly for classification instead of trying to recover the shape explicitly. To achieve this we need to use an appropriate projection pattern and derive features that sufficiently characterize the deformations. The patterns required could be quite different depending on the nature of object shape and its variation across different object classes.

3.1 Deformations of Projected Texture

The resultant deformation in projected texture can be thought of as transformation in the originally projected patterns due to depth profile of the target surface, neglecting the physical property of the surface, such as reflectance.

3.1.1 Deformation due to depth variation

The transformations of a ray of projected light due to variations in surface depth can be primarily classified into two categories:

- *Pattern Shift*: The position where a particular projected pattern is imaged by the camera depends on the absolute height from which the pattern in reflected. Figure 3.3 illustrates this with a cross section of a projection setup. Note that the amount of shift depends on the height difference between the objects as well as the angle between the projector axis and the height axis.
- *Pattern Deformation*: Any pattern that is projected on an uneven surface gets deformed in the captured image depending on the change in depth of the surface

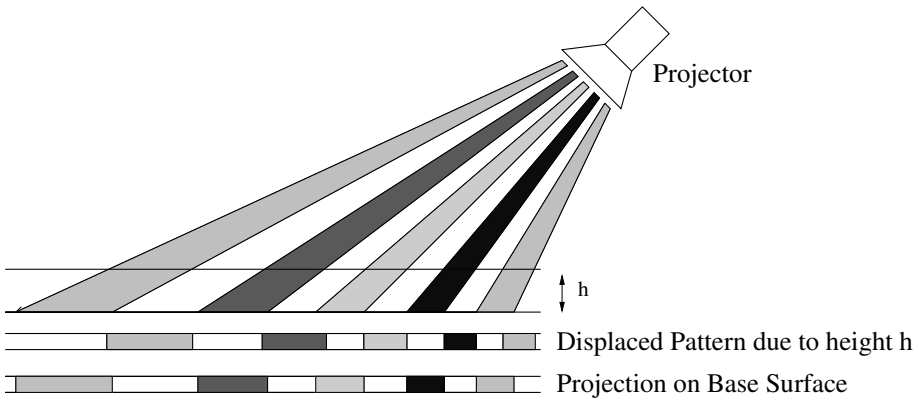


Figure 3.3: Shift in projected pattern due to uniform height difference in target surface.

(see Figure 3.2 and 3.1). 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.

3.1.2 Deformations due to physical properties of the surface

Other than depth variations, there are other factors such as reflectance and natural texture that contribute to changes in the projected texture. The physical property that mainly affects the final deformed pattern is the reflectance of the surface. The various problems that can arise due to reflectance are :

- Specular surfaces
  Specular surfaces have always been a challenge in the field of recognition and reconstruction. The primary issue with partially specular surfaces is that the appearance of the object depends on the environment, which is reflected on the object, in addition to its shape and other surface properties. For purely specular objects (mirror-like surfaces), the projected texture at any point gets reflected in a direction depending on the direction of projection and the object normal, and hence is not captured by the camera. However, with partially reflective surfaces, this is mostly not an issue as the light captured by the camera comes from lambertian reflection.

- Transparent and Translucent surfaces
  Transparent and translucent objects pose another similar challenge. The projected texture based approach allows us to capture the patterns generated due to inter-reflection of light within the object, in the case of translucent objects, and hence recognize them in many cases. Purely transparent objects would need a different approach, where the projected texture is captured after refraction through the transparent medium. This refracted texture could either be captured directly, or using a planar screen.

- Natural Texture
  Natural texture that is present on the surface of the object affect the reflectance at various points. This in turn introduces variations in the amount of projected

light that is reflected from the object surface. Note that this affects the intensity of reflection and not its location or pattern shape. Moreover, this variation does not include any information regarding the shape of the object, and hence need to be nullified while characterizing the object.

- Dark surfaces
  Purely black surfaces does not reflect any light back and hence are not observable in theory. However, most dark surfaces reflect some amount of light, and hence can be dealt with by calibrating the projection light and the camera's sensitivity.

We note that, projected texture can incorporate the physical properties of surface in addition to its shape, which is often useful in recognition/classification. This is a simple and efficient way to implicitly use these inherent properties of surface, instead of complex approaches like modeling the surface albedo for obtaining a 3D model of the surface/object.

3.2 Pattern Deformation and Projector Camera Configuration

In this section, we will discuss how does the projector camera configuration affects the resultant deformation in projected texture. As mentioned previously, the deformation due to depth variation depends on uniform height variation and slope in object surface. We now derive the relationship between the relative configuration of the projector, the camera and the object, and the amount and nature of deformation introduced to the projected pattern. This relationship allows us to determine the type of projected pattern and the projector-camera configuration to be used for a particular application.

3.2.1 Quantifying Deformation

Figure 3.4 shows a planar object being illuminated by a sheet of light (a line in the 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

$$x \cot \phi + z - b = 0 \tag{3.1}$$

Object surface plane can be represented by

$$z - y \tan \theta = 0 \tag{3.2}$$

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

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

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

or,

$$\vec{n_3} = [\ -\tan\theta \quad \cot\phi \quad \tan\theta\cot\phi]^T \tag{3.3}$$

One point common to both plane say p can be obtained by solving equation 3.1 and 3.2

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

Hence equation of 3D line can be written as

$$\vec{r} = [\ b\tan\phi - s\tan\theta \quad s\cot\phi \quad s\tan\theta\cot\phi]^T, \tag{3.4}$$

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

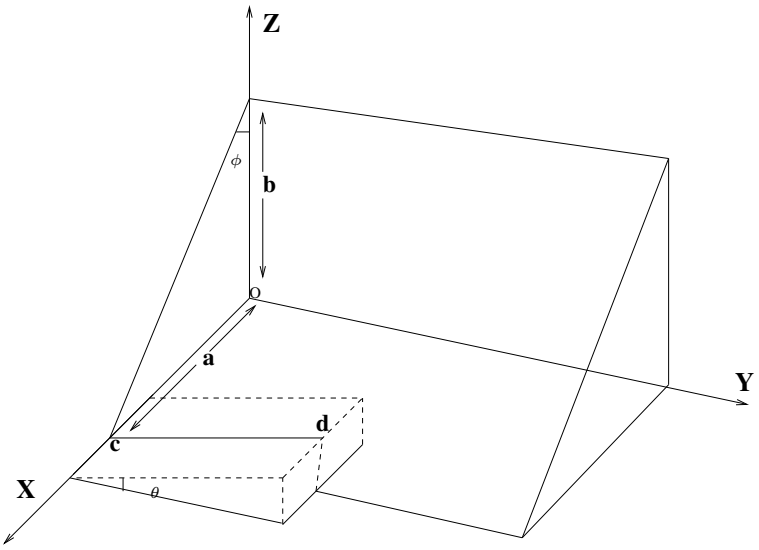


Figure 3.4: Geometric representation of our projector camera configuration. Line cd in XYZ coordinate system, which is formed due to intersection of Projector plane and object surface, is imaged in image plane of camera and thus define a deformation

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.

$$Q_1 = [\ b\tan\phi - l_1\tan\theta \quad l_1\cot\phi \quad l_1\tan\theta\cot\phi]^T \tag{3.5}$$

$$Q_2 = [\ b\tan\phi - l_2\tan\theta \quad l_2\cot\phi \quad l_2\tan\theta\cot\phi]^T \tag{3.6}$$

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

$$q_1 = \begin{bmatrix} R_1(b\tan\phi - l_1\tan\theta) + R_2 l_1\cot\phi + R_3 l_1\tan\theta\cot\phi + t_1 \\ R_4(b\tan\phi - l_1\tan\theta) + R_5 l_1\cot\phi + R_6 l_1\tan\theta\cot\phi + t_2 \\ R_7(b\tan\phi - l_1\tan\theta) + R_8 l_1\cot\phi + R_9 l_1\tan\theta\cot\phi + t_3 \end{bmatrix} \tag{3.7}$$

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

$$q_1 = \begin{bmatrix} X_1 & Y_1 & Z_1 \end{bmatrix}^T, \tag{3.8}$$

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.

$$q_2 = \begin{bmatrix} X_2 & Y_2 & Z_2 \end{bmatrix}^T \tag{3.9}$$

In homogeneous coordinate system q_1 and q_2 can be represented as

$$\overline{q_1} = \begin{bmatrix} \frac{X_1}{Z_1} & \frac{Y_1}{Z_1} \end{bmatrix}^T, \overline{q_2} = \begin{bmatrix} \frac{X_2}{Z_2} & \frac{Y_2}{Z_2} \end{bmatrix}^T \tag{3.10}$$

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

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

or,

$$\vec{L} \ : \ X(Z_1Y_2 - Z_2Y_1) - Y(Z_1X_2 - Z_2X_1) - X_1Y_2 + X_2Y_1 = 0 \tag{3.11}$$

$$m = (Z_1Y_2 - Z_2Y_1)/(Z_1X_2 - Z_2X_1) \tag{3.12}$$

From the equation of line it can inferred that the slope m of the line in the image computed in equation (3.12) will depend upon X_1, Y_1, Z_1 and X_2, Y_2, Z_2, which cab be further expanded in terms of b, ϕ and θ. Thus, the slope of object surface directly affects orientation of the projection of a 3D line onto the image plane. This change in orientation of lines in the projected texture in camera plane directly affects responses of a filter bank. Thus distinct height profiles create different responses to localized filters, which can be used as a signature of the object.

3.2.2 Setup Details

The experimental setup consists of a projector and camera arranged such that their field of views overlap. The camera is fixed at a height h the reference surface, pointing down, such that a top view of the object is captured. The projector setting is such that it is tilted at an angle ϕ and placed at a height b. The projector and the camera are focused at the center of the object so that the whole object is in acceptable focus. A typical setup is shown in Figure 3.5.

Figure 3.5: Experimental Setup used for proposed Projected Texture Approach

The relative position of the projector, camera and the 3D object/surface are very important and should not change throughout the process. The ambient illumination is kept constant for views captured for image-based recognition algorithms, for comparison. We can characterize the deformation in a position and illumination in an invariant manner to overcome these restrictions.

3.2.3 Capturing Disparity in Multiple Directions

In the previous derivation, we have assumed only one orientation of projector, for the sake of simplicity. This allows us to capture the disparity (*i. e.* , depth profile), only in one direction, although depth variation in other directions also contribute to the shape. One option could be to use multiple projectors or a single projector that can be placed in different direction around a fixed camera. Figure 3.6 shows such a possible setup configurations. Another possibility is to rotate the object/surface itself and capture the deformation. In both cases we can capture the deformations in multiple directions and thus can get more information for the recognition task. In our experiments, we have rotated the object instead of moving projector.

3.3 Design of Projected Texture

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

1. For the deformation to be visible in the captured view at any point in the image, the gradient of the texture at that point should not be zero in the direction of

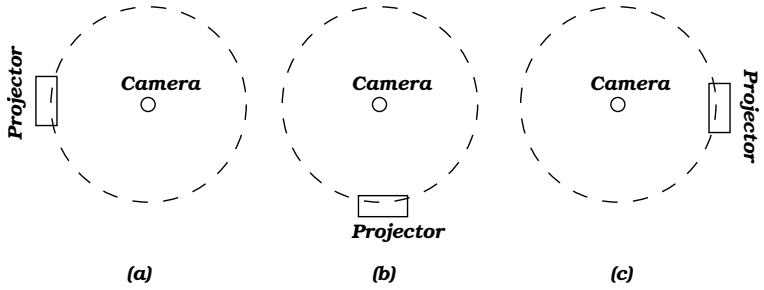


Figure 3.6: Experimental setup configuration with projector at multiple position around fixed camera, in order to capture disparity in different directions.

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 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.

3.3.1 Fixed and Adaptive Patterns

For the purpose of 3D texture recognition, we use a set of parallel lines with regular spacing, where the spacing is determined based on the scale of the textures to be recognized. For hand geometry based authentication, we have selected a repetitive star patten that has gradients in eight different directions. The width of the lines and the density of patterns in the texture were selected experimentally so that it captures the height variations between the palms at the angle of projection selected.

Another possibility is to have a feedback mechanism that allows to improve the design of patterns iteratively and in-turn improves the performance of recognition system. One could think of a optimization framework based on a discriminant function to design a pattern that best differentiates between a given set of objects/surfaces.

3.4 Summary of Projected Texture

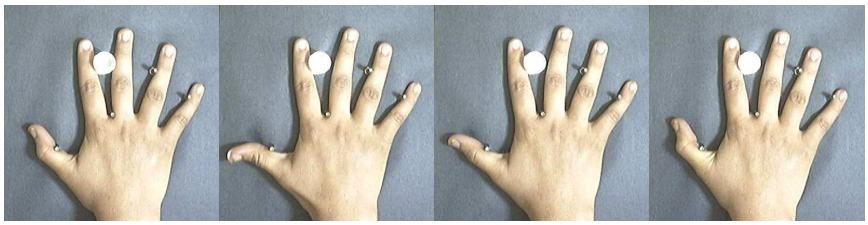
In this chapter we have introduced the concept of *Projected Texture*, which introduces the important shape information into an image for the purpose of classification. To the best of our knowledge, this is the first attempt to encode the shape information in deformed texture and process it as a texture for classification/recognition problem. This helps us to avoid the computationally complex and error prone process of recovering 3D models. One of the properties that is not explored in detail is the interaction of physical properties of an object material, such as transparency and

specularity, with projected texture. This requires detailed modeling of the object material and is beyond the scope of this thesis.

Chapter 4

Recognition of 3D Object with Fixed Pose

Object recognition is a challenging problem in computer vision (Chapter 1). For recognition algorithms that rely on object based features, the segmentation and localization of the object of interest are the most difficult parts in preprocessing. Also, the pose variations are difficult to deal with as the appearance changes considerably with pose. These problems are relaxed in the case of objects with fixed pose. The object recognition task can work directly on the images as the appearance is does not vary much across instances of the same object. There are many scenarios in which object pose may be assumed to be static. Biometric authentication is one such scenario. In this chapter we have proposed an approach based on projected texture to characterize fixed pose objects for recognition. We have demonstrated our approach by developing a hand geometry based person authentication system.



(a) **(b)** **(c)** **(d)**

Figure 4.1: Examples of hand images which shows non rigid nature of hands

4.1 Hand Geometry based Authentication

The problem here is to differentiate similar objects with a fixed pose. Note that we are dealing with the problem of recognition of a specific instance and not a category. Defining a fixed pose for a category is not always trivial, and hence is dealt with as an arbitrary pose recognition problem. We will tackle the problem of arbitrary pose category recognition in next chapter.

We have taken hand geometry based person authentication as the example problem in this class. Our aim is to authenticate (verify the claimed identity of) an individual on the basis of their 3D hand geometry, which is an important biometric characteristic

for civilian applications. Note that many people are not comfortable with providing their fingerprints due to the criminal stigma associated with it. In such case hand geometry can be good feature to differentiate between users. Primarily, we have focused on the problem of authentication, although the approach can be extended (as we show) to recognition also. Authentication involves, given a set of reference templates corresponding to a person, and a test sample claimed to be of the same person, verifying whether the claim is correct or not. It means we have a claimed identity, and we need to match the current biometric sample with the reference samples of the claimed user.

In the case of recognition we do not know the owner of the given biometric sample, and our task is to find the best match among the enrolled set of users. In the later part of this chapter, we will also demonstrate our results on person recognition.

The major challenges in hand geometry based authentication problem are:

- Similar Shapes
  The hand shapes of the two individuals are very similar in appearance. This requires the proposed feature to have fine resolution in order to authenticate two people.

- Non Rigid Objects
  The nature of human hand is not rigid. Thus final appearance of same hand can be different at different times, even when we ask the person to put the hand in same pose. Figure 4.1 demonstrates the non rigid nature of human hand.

- Varying Pose
  Pose variation is an other important challenge. Even if we have a peg based system to capture fixed pose data, different people put their hand with varying pose at different instances of time. Although there is no major variation in pose still these minor variation are significant while dealing with hand geometry, due to the similarity in shapes of different hands. Figure 4.2 shows examples of the hand of a single person with varying poses.

(a) **(b)** **(c)** **(d)**

Figure 4.2: Examples of hand images of a person with varying pose.

4.2 Projected Texture for Hand Geometry based Authentication

Our task is to utilize the 3D shape of hand to differentiate between samples from different users. We now propose a solution to this problem in framework of projected texture proposed in Chapter 3. Thus, our problem reduces to characterizing the

deformed patterns in such a way that it captures 3D information in reliable and efficient manner.

4.2.1 Feature for Characterizing Deformations

As noted in the previous chapter, there are two types of transformations: shift and deformation. We design a projection pattern that can capture both for the purpose of authentication. Wavelet methods for texture analysis has been well accepted as a good feature to characterize local frequency components, in our case the deformations in projected texture. Since the pose of the object is fixed we can use the spatial location of the projected patten as a clue for recognition.

For the purpose of hand geometry based authentication, we have selected a repetitive star patten that has gradients in four different directions. This will allow us to capture depth variations in different directions within a window. The width of the lines and the density of patterns in the texture were selected experimentally so that it captures the height variations between the palms at the angle of projection selected.

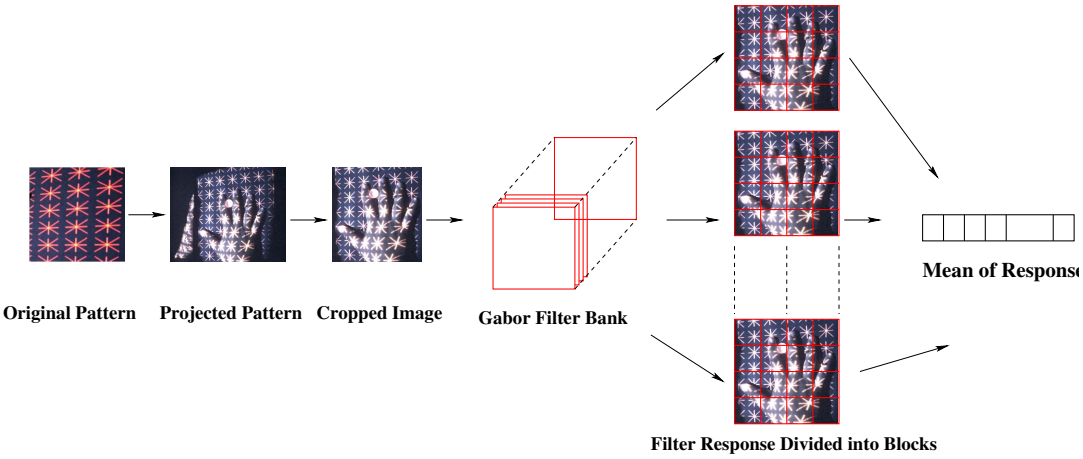


Figure 4.3: Computation of the proposed projected texture based features.

Window based Gabor feature

Once the pattern is selected, we need to characterize the deformations induced by the height variations of the object. For hand-geometry based verification, we divided the image into 64 sub-windows (a 8×8 grid). Each sub-window is then characterized by the responses of Gabor filters that captures the local frequencies and their orientations.

A Gabor function is a Gaussian function, modulated by a complex sinusoid. A simplified form of the filter, $G(x, y)$, may be written as:

$$\begin{aligned} G_{\sigma,\phi,\theta}(x,y) &= g_\sigma(x,y) \cdot e^{2\pi j\phi(x\cos\theta + y\sin\theta)} \\ g_\sigma(x,y) &= \frac{1}{\sqrt{(2\pi)}\sigma} e^{-(x^2+y^2)/2\sigma^2}, \end{aligned}$$

where θ is the orientation of the sinusoid with frequency ϕ, and $g_\sigma(x, y)$ is a Gaussian

with scale parameter σ. The feature vector representing a sample image is computed as follows (see Figure 4.3).

4.2.2 The Classifier

Since proposed feature captures discriminative information efficiently, we have used a simple nearest neighbor classifier for feature matching. This helps to evaluate the representation power of our feature.

4.3 Experiments

In this section we have given a detailed account of experiments we have conducted including dataset collected, implementation details or our and compared approach.

4.3.1 Dataset Details

To analyze the discriminative power of the projected texture based features, we check the verification performance on a dataset of 1341 hand images collected from 149 users. Each user provided 9 images with the projected texture and 9 with uniform illumination to serve as a comparison dataset for traditional 2D approaches.

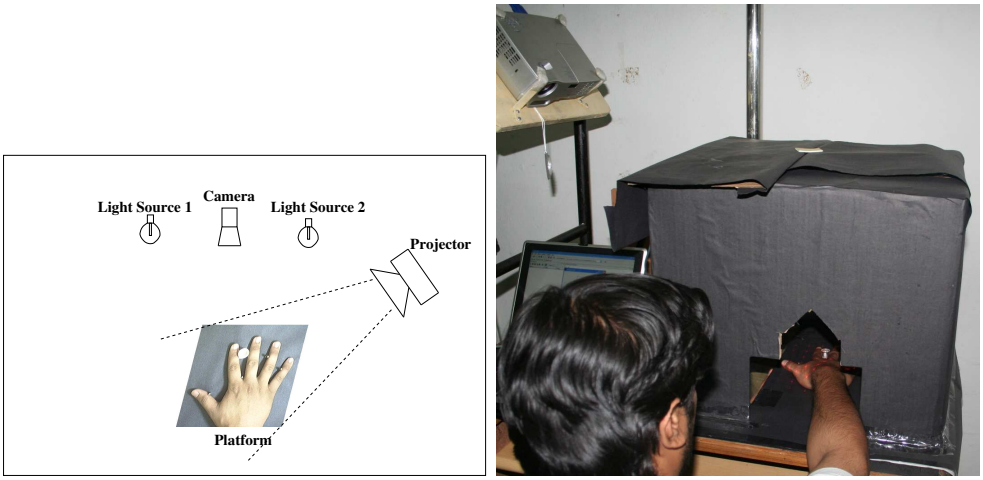


Figure 4.4: Imaging setup for projecting a pattern and capturing deformations due to hand geometry

The image capturing setup is similar to that discussed in Jain *et al.* [13], where pegs are used to guide the placement of the palm. However, unlike the popular peg-based datasets, where the placement of the hand is controlled, we encouraged the users to vary the hand pose within the peg limits to make the dataset more realistic as in unsupervised scenarios. The surface of placement of the palm was darkened to facilitate the segmentation process for 2D image analysis. The illumination over the area of the palm could be either uniform or from a projector that is placed at an angle to the palm (see Figure 4.4). The images are captured by a camera located directly above the palm with its optical axis perpendicular to the palm surface. A reflecting mirror fixed by the side of the imaging surface helps to capture a side view of the palm, and thus to include thickness of fingers as described in [13]. Each user

provided 9 hand images with uniform illumination, as well as with projected texture (see Figure 4.5). The users were asked to remove their hand and replace it for each image captured, with limited variations in hand pose.

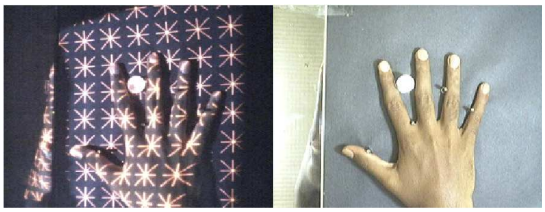


Figure 4.5: The hand under structured illumination as well as normal lighting. The mirror to the left enables us to compute height of palm at specific points.

4.3.2 Implementations

In our experiments, we use a bank of 24 filters with 8 orientations ($\theta = 0, \pi/8, 2\pi/8, \cdots, 7\pi/8$) and 3 radial frequencies, controlled by the frequency of the sinusoid.

The image is converted to gray scale and the area of the image that contains the palm is cropped. The pixel values are then normalized to have a specific mean and variance for the image. The resultant image is then convolved with each of the 24 Gabor filters and the mean of the filter responses are computed for each sub-window. In our experiments, we divided each response image into 64 sub-windows (8×8). This resulted in a feature vector of dimension 1536.

For the purpose of comparison, we compute two different 2D feature sets from the samples with uniform illumination, in addition to those from the projected texture. The feature sets used are:

- *Feat-1*: The set of 17 feature proposed by Jain *et al.* [13], computed from the width and length of fingers and palm as well as the height of the index finger computed from the image reflected on the mirror.
- *Feat-2*: A set of 10 features proposed by Faundez-Zanuy *et al.* [14], including 5 finger lengths, area of the palm, the contour length and distance between specific points on the palm contour.
- *Feat-3*: The proposed projected texture based features, computed from filter responses from 64 sub-windows.

Comparisons with 3D hand geometry approaches such as Cofer and Hamza [125] and Faulkner [126] could not be carried out as the patents does not provide sufficient information about the exact nature of feature extraction. Moreover, as we mentioned before, our approach does not require depth computation or even segmentation of the palm, and hence is comparable to the image based approaches in spirit and complexity.

Due to the variations in hand pose, our dataset is much more challenging than popular ones using peg based approach. Figure 4.6 demonstrate 9 hand samples in our dataset and a bar representation of corresponding feature vector (57 dimensional selected feature vector is shown in figure instead of full 1536 dimensions). Note that users introduced considerable variations in the pose, even when constrained by

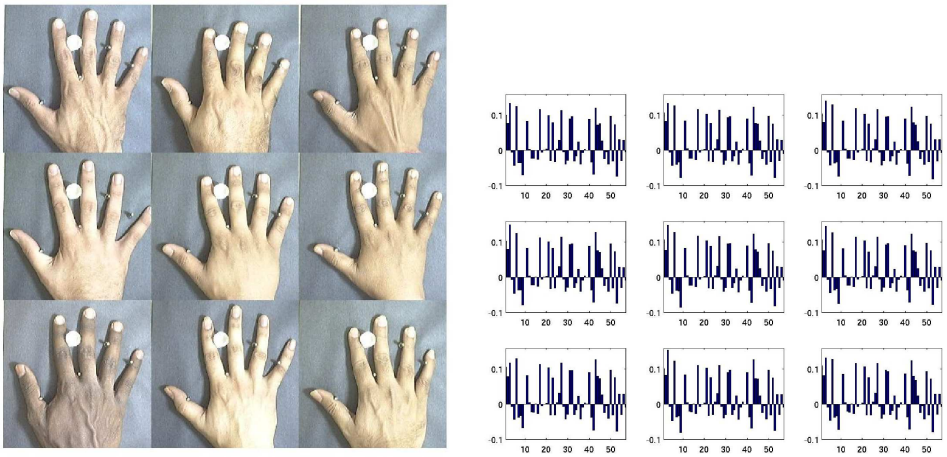


Figure 4.6: Sample image of 9 users and corresponding histogram representation

the pegs. Similar variations were introduced for the projected texture dataset also. Due to these variations in pose, traditional approaches for 2D feature extraction that assumes peg-based imaging will fail in many samples. Hence we have tracked the finger locations and computed the features appropriately. We also verified and manually corrected the 2D features that were incorrectly computed due to pose variations.

The primary aim of the experiments is to compare the proposed feature set to the traditional image based features. For this reason, we have avoided complex classifiers or post processing techniques as proposed in Faundez-Zanuy *et al.* [14]. One of the best indicators of the discriminating power of a feature set is the ROC curve induced by distances computed in the corresponding feature space. The ROC curve indicates the level of separation between the genuine and imposter distance distributions. We used a simple Euclidean distance to compute the distance between feature vectors in all three cases. Note that the performance of the classifiers in each case could be improved by more complex classifiers or post processing techniques. Hence the accuracies reported here should be used only for comparison of the feature spaces, and not as an indicator of the absolute discrimination power of any of the feature sets.

4.4 Results and Discussion

Figure 4.7 gives the ROC curves obtained from the three feature sets mentioned above. The Equal Error Rate (EER), or the rate at which false rejects equals false acceptance rate, is a single indicator that can be computed from the ROC curve. The EERs for the *Feat-1* and *Feat-2* were 4.06% and 4.03% respectively, while the proposed feature set achieved and EER of 1.91%. Clearly the projected patterns induce a large amount of discriminating information into the computed features. In addition to the equal error rate, we note that the genuine acceptance rate continues to be above 80%, even at false acceptance rates of 0.001% for the proposed features, while the performance of the 2D image based features degrade considerably at this point.

We also conducted an experiment in feature selection to choose a subset of the 1536 features that would help us in reducing the computations required. We note

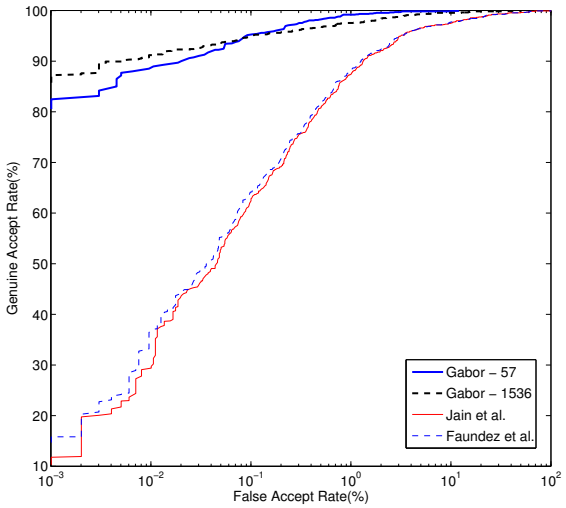


Figure 4.7: ROC curves for two 2D feature based, and the proposed projected texture approaches.

that even with just 57 features out of 1536, the ROC curve is similar to that of the complete feature set. Moreover, the equal error rate improves to 0.84% with the reduced feature set. This is possible as the selection process avoids those sub-windows, where the intra-class variations in pose are high.

Figure 4.8: The red color square patches represents the windows corresponding to selected features

Figure 4.8 shows the windows corresponding to the most discriminative 12 features. Note that the features belong to windows that are at the edges of the fingers as well as on the palm surface, which indicates that the depth information of the palm is also used for authentication, in addition to the shape of the fingers. The presence of a window outside the palm region could be because it encodes the relative brightness of projected pattern, which in turn encodes the skin color.

Another interesting observation is that a weighted combination of the distance scores from 2D and texture-based features did not improve the performance. Evidently, the projected texture encodes most of the information that is contained in the

2D features, in addition to the 3D information of the hand surface.

We have also applied 2D Fast Fourier Transform (FFT2) feature proposed in next chapter for comparison purpose. Figure 4.9 shows the ROC curve for FFT2 feature. We have achieved EER of 1.53% with FFT2 feature which is a 5000 dimensional

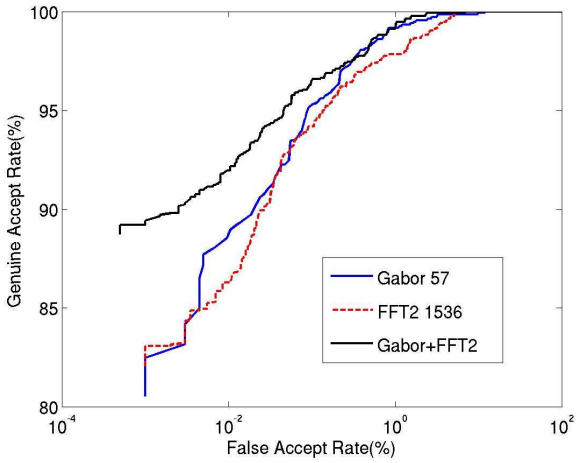


Figure 4.9: ROC curves for FFT2 based, Gabor based and both combined approach.

vector in our case. While combining weighted combination of 57 dimensions Gabor with FFT2, we have achieved a EER of 0.90% which is close to what achieved by selected Gabor feature. Although EER of combined feature is slightly worse than single Gabor feature, it performs well at lower false acceptance rates. Thus, it can be a useful feature when we need high performance on limited false acceptance rate.

Figure 4.10 shows examples of misclassification, where a sample from one user get misclassified as another user.

4.5 Hand Geometry based Recognition

We have also run the experiment to obtain the recognition results using the same feature vectors used for authentication. We have performed column-wise normalization of the feature vector to obtain the recognition performance using 1-Nearest Neighbor classifier. Table 4.5 shows the different approaches and their recognition accuracies.

	Jain *et al.*	Neural Hand	Gabor	FFT2	Gabor_sel	Gabor_sel+FFT2
1-Nearest Neighbor	12.45	12.08	0.97	0.75	0.45	0.30

Table 4.1: Recognition Error Rates

We can see from table that the combination of selected Gabor features and the FFT2 feature perform the best with an error rate of 0.30%. These results shows a significant improvement as compared to 2D image based approaches.

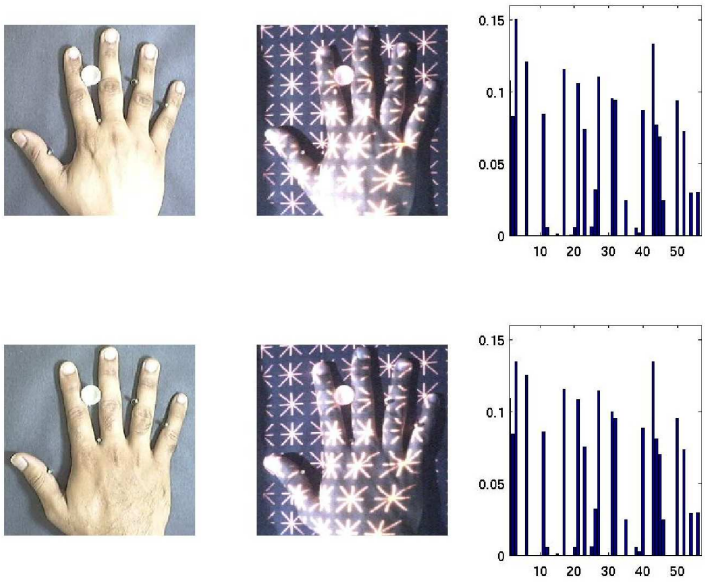


Figure 4.10: Example of misclassification: normal image, projected pattern image and a bar representation of the feature vector.

4.6 Summary

We have proposed a new feature that characterizes the deformed pattern projected on the hand, efficiently capturing the depth information encoded in those deformations. We have demonstrated a robust hand geometry based person authentication system that works on a feature vector that computes local frequencies in fixed windows. We note that the computation of textural features from specific local windows can yield a feature vector that is far more discriminative than the traditional 2D object features used for hand geometry based authentication. The approach is computationally efficient and the time taken is comparable to that of the 2D image based authentication. Moreover, the approach is robust to occlusions and noise as opposed to 3D hand geometry systems that need to explicitly compute a depth map of the hand. However, several issues still remain unaddressed. Temporal variation in hand geometry is one such issue, as the shape of a person's hand can vary with age, weight, etc. For this we need to collect data for a longer period of time and introduce learning at the classifier level to accommodate changes in hand geometry with time. Similar problem exists with most of the existing system also. Note that the same feature can also be applied to other fixed pose recognition problems as well.

Chapter 5

Recognition of 3D Object with Arbitrary Pose

Another related problem in object recognition is that of recognizing the category of an object. For example, one might want to recognize that the given image is that of a car instead of identifying it as a specific model and make. Image based category recognition approaches often follow a part based approach where the appearance of selected windows of the image are used to identify the object category. However, as in the previous case, the lack of depth information continues to be a bottleneck here.

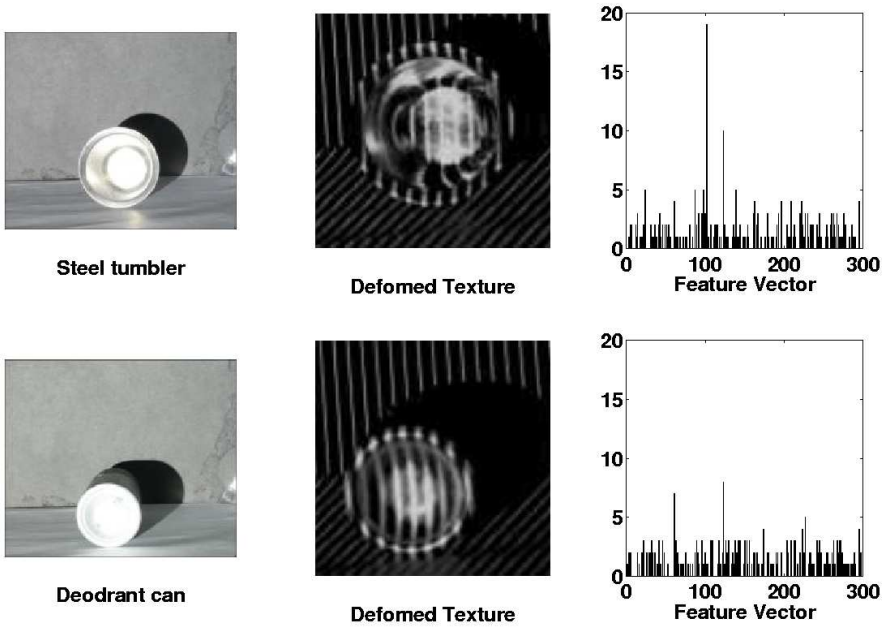


Figure 5.1: Variations in deformation on similar looking objects.

In this chapter we have given an efficient and robust 3D object category recognition approach under the Projected Texture framework presented in chapter 3. We have proposed a 2D Fourier based feature on top of bag of words approach, which can efficiently capture the deformation in projected texture and model an object.

5.1 Challenges in Object Category Recognition

Recognition of object categories 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. Due to large variation in reflectance property, object from same category are difficult to recognize. Translucent and transparent as well as dark objects fall in this category. Another problem is lack of natural texture in object category recognition. A detailed discussion can be found in Chapter 1.

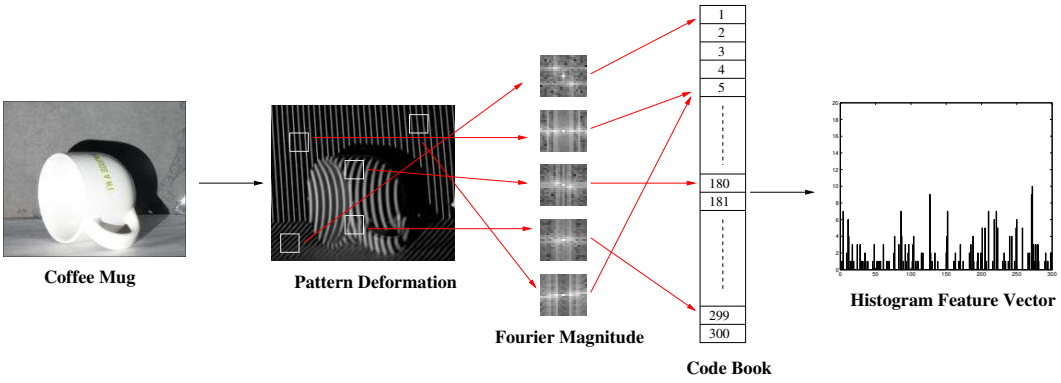


Figure 5.2: Computation of window based FFT2 feature vector.

We have seen in chapter 3 that slope of height variation of the object directly affects orientation the imaged line. This indicates that characterization of an image patch in terms of the angle of the imaged lines can capture the surface height variations at that point. The exact relationship enables 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 specularity.

5.2 Characterizing Deformations

The primary concerns in developing a representation for object category is that the description should be invariant to both shape and pose of the object. 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 [24]. 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.

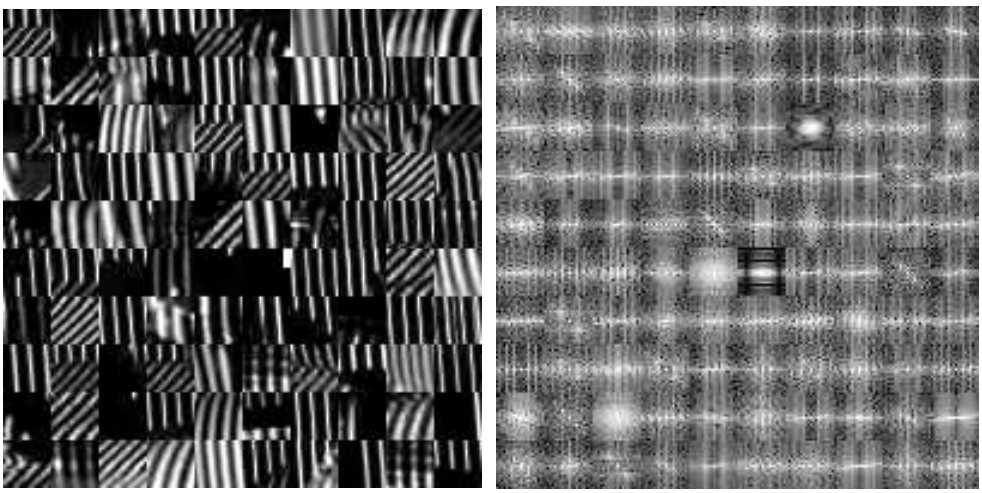


Figure 5.3: Spatial and corresponding Spectral representation of 100 Words from our codebook

5.2.1 2D Fourier Transform based Feature

There are two primary concerns to be addressed while developing a parts based shape representation, one is local shape descriptor and the other is identifying the proper locations to compute local shape descriptors. 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 direct use of the natural 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 should exploit the periodic nature of projected patterns. Fourier representation is an effective descriptor for periodic signals. 2D discrete Fourier transform (DFT) of an image function $f(x, y)$ can be described by equation [5.1]. $|F(u, v)|$ represents the magnitude of 2D DFT and Φ represents phase information as defined in equation (5.2).

$$F(u,v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \exp^{-j2\pi(\frac{xu}{M}+\frac{yv}{N})} \tag{5.1}$$

$$\begin{aligned} |F(u,v)| &= \sqrt{F_r(u,v)^2 + F_i(u,v)^2} \\ \Phi &= \tan^{-1}(\tfrac{F_r(u,v)}{F_i(u,v)}) \end{aligned} \tag{5.2}$$

Since we are interested in the nature of deformation and not its exact location, we compute magnitude or the absolute value of the Fourier coefficients (AFC) corresponding to each of the window patch as our feature vector. This magnitude map of a local patch will capture the change in orientation of the periodic projected pattern, which is critical to represent the local object shape. The phase map will capture

mainly shift in projected pattern, which will change with positioning of the object. As explained earlier we are using a codebook approach and thus finding local orientation deformation in projected pattern, for which the magnitude of the FT is a better choice of representation. To make comparisons in a Euclidean space for 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.

Figure 5.2 illustrates the computation of the feature vector from a scene with projected texture. The main step of feature extraction process are:

- The images in the training set are divided into a set of overlapping windows of size 20×20 (decided experimentally).
- Each window is represented using the magnitude of Fourier representation in logarithmic scale. This results in a 200 dimensional feature vector (due to symmetry of Fourier representation) for each window.

The features are extracted during training as well as testing phase.

- Training Phase:
  - Project pattern on object surface and capture deformed texture image with camera.
  - Perform normalization operation on image to achieve illumination invariance.
  - Divide the image into windows and extract 200 dimensional Fourier transform based feature vector for each window.
  - Apply K-means clustering in 200 dimensional feature space for all windows over all training data, for all category of objects. This allows us to identify the dominant pattern deformations, which forms a codebook for the classification problem. (see figure 5.3 for our codebook).
  - Find the closest match of each of the feature vector corresponding to each window into codebook and then represent each object category training image with histogram of these codebook index.
- Testing Phase:
  - Perform normalization operation on image to achieve illumination invariance.
  - Divide the image into windows and extract 200 dimensional Fourier transform based feature vector for each window.
  - Find the closest match of each of the feature vector corresponding to each window into existing codebook learned during training and then represent each object category training image with histogram of these codebook index.
  - Apply Nearest Neighbor or any other classifier to find the matching the histogram representation of test image from stored category models.

As mentioned above, during the testing phase, the feature representation of the windows in an image is 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 5.2 the patches that are part of the background maps to one location in codebook. Thus codebook can isolate the words that captures maximum information for defining an object category. Detailed comparison of recognition results of new features and existing one are presented in the next section. 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.

5.3 Experimental Details

The pre-processing step includes the 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.

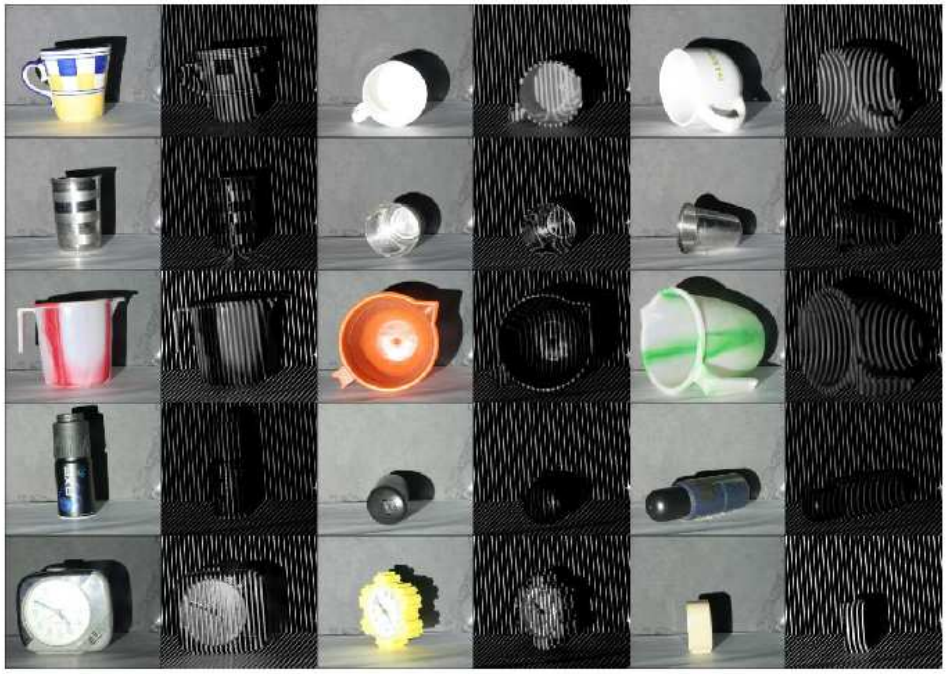


Figure 5.4: 3D object category Dataset

5.3.1 Dataset Description

We have collected dataset with total 5 object categories: i) *Coffee Cup*, ii) *Steel Tumbler*, iii) *Plastic Mug*, iv) *Deodorant Spray*, and v) *Alarm Clock*. The categories were chosen to introduce challenging similarities between object categories, and 5 objects of each category were chosen to have large intra-class variations (see Figure 5.4). 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.

5.3.2 Implementation Details

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 classifier. All results reported are the mean error rates based on 4 Fold cross validation. 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 same feature used by [24] on our dataset without the projected patterns. 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 difference is in the local patch representation.

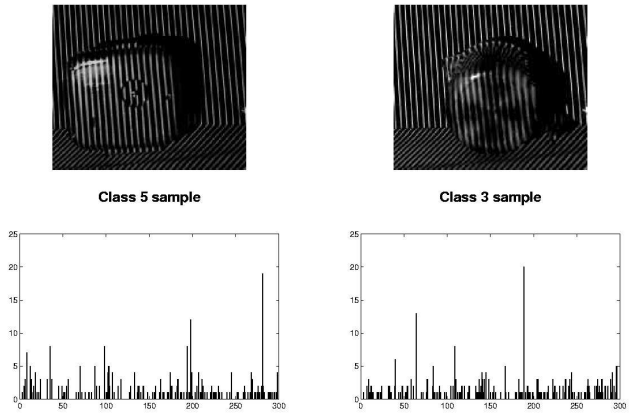


Figure 5.5: Miss classification example in which instance of object class 5 (*i. e.* Clock) get misclassified as instance of class 3(*i. e.* Plastic Mug)

5.4 Experimental Results and Analysis

	LAFC	SIFT	GABOR
MLP	1.33	21.33	15.11
1-NN	5.73	20.09	14.22

Table 5.1: Recognition Error Rates

Table 5.1 presents the mean error for three features. The first one is the SIFT based feature set proposed by [24], for object category recognition, which is considered to be the state of the art for image based recognition. The Fourier magnitude features proposed here (LAFC) is the second set. Since the Gabor filter based features proposed in the previous chapter can also encode some of the shape information, we have included it in the experiments for comparison purposes.

The recognition results clearly show the superiority of our approach over the state-of-the-art. The error rate is only 1.33% in our case. Table 5.2 shows the confusion matrix for our experiment. There are only three misclassifications, which includes

	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 5.2: Confusion Matrix of the our approach.

two between classes 2 and 4, and one between classes 3 and 5. Figure 5.5 shows one of the misclassification example with corresponding feature vector.

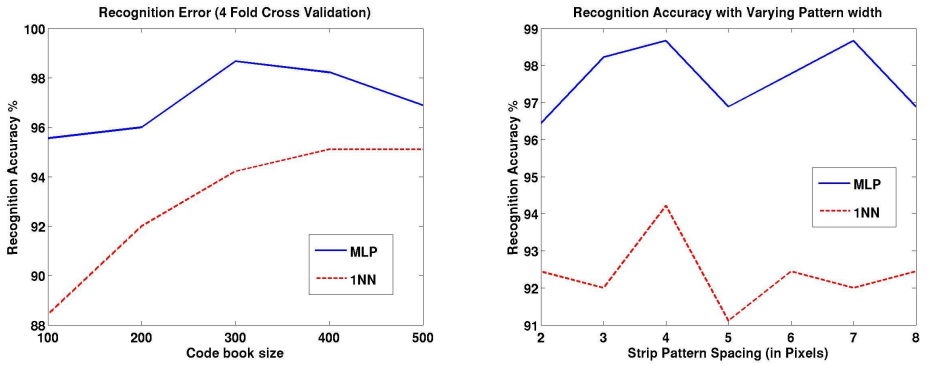


Figure 5.6: Performance with variation in Codebook size and Pattern width

We also conducted experiments with different codebook sizes and pattern variations. Figure 5.6 shows the graph of accuracy vs size of code book and variation in performance with change in width of projected pattern.

5.5 Summary

We have presented a novel approach for object category recognition within the previously proposed framework of Projected Texture, using a Fourier transform based feature, which efficiently captures the deformation in projected periodic patterns. A patch based representation of the object categories is used, where each patch is characterized by a frequency domain representation of deformations. The effectiveness of the approach is demonstrated on a small but challenging dataset, which demonstrates a significant improvement in recognition rates.

Chapter 6

Classification of 3D Texture

Texture classification is an important problem in object recognition field as most of the real world object surface has some amount of texture. Moreover, the object category here is a different nature, where the stochastic variations in the depth characterizes a surface. Most of the approaches in literature define the texture as albedo variation and handled it as stochastic intensity variation in 2D. In recent past, research community has also started working on 3D textures, where subtle variations in the height creates a textured surface. Such textures are abundant in real world, such as the surfaces of bricks, leather, concrete, and sand. Thus proposed solutions in the literature tries to model the surface reflectance from ambient light, or try to recover the surface normal with stereo image pairs, and then try to recognize/classify the surface.

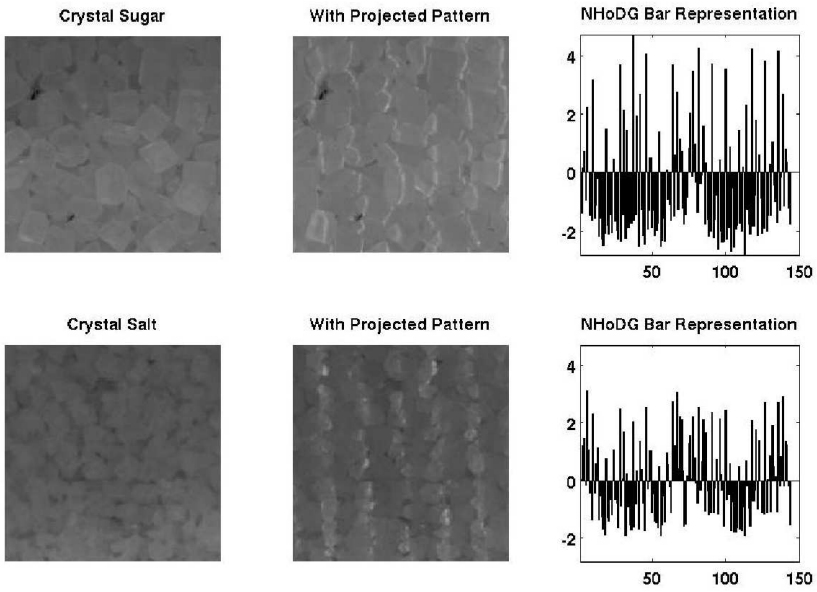


Figure 6.1: Salt and Sugar crystal with and without projected texture and the corresponding feature representations.

Within the framework of projected texture that we have presented in chapter 3, we have proposed a statistical feature for texture classification. The inherent property of 3D texture makes depth variation stochastic. Statistical features such as histograms are best suited for capturing these variation.

6.1 Projected Texture for 3D Texture Classification

The idea is to use projected texture to encode the shape information of the surface and characterize the 3D texture surface for classification. Since the variations in depth are stochastic in nature, a histogram based approach can be used to represent a particular variation.

The key challenges as explained in chapter 1 are intra and inter class variations, inter-reflections, scaling etc. Figure 6.1 shows an example to two similar texture classes (salt and sugar crystals), under regular illumination and with projected texture, along with the computed feature vectors. One can clearly see the difference in feature representation with projected texture, while the image based feature sets look similar. One should note that an approach using structured lighting has it limitations also as it requires some amount of control of the environment.

However this approach has wide range of applicability. Robot navigation is one such scenario, where a robot has to adjust it's navigation parameter according to the nature of surface on which it is moving. Since recovering 3D is complex and error prone method, our approach fits well for such scenario. Other scenario could be the classification in a supermarket counter or an industrial conveyor belt, where we need to characterize the object based on its surface properties.

6.2 Deformation Characterization

An effective method for characterization of the deformations of the projected texture is critical for its ability to discriminate between different object. We propose a set of texture features that captures the statistics of deformation in the case of 3D textures. As we noted before, the projection pattern used for 3D texture classification was a set of parallel lines. The feature set that we propose (NHoDG), captures the deformations in the lines and computes the overall statistics.

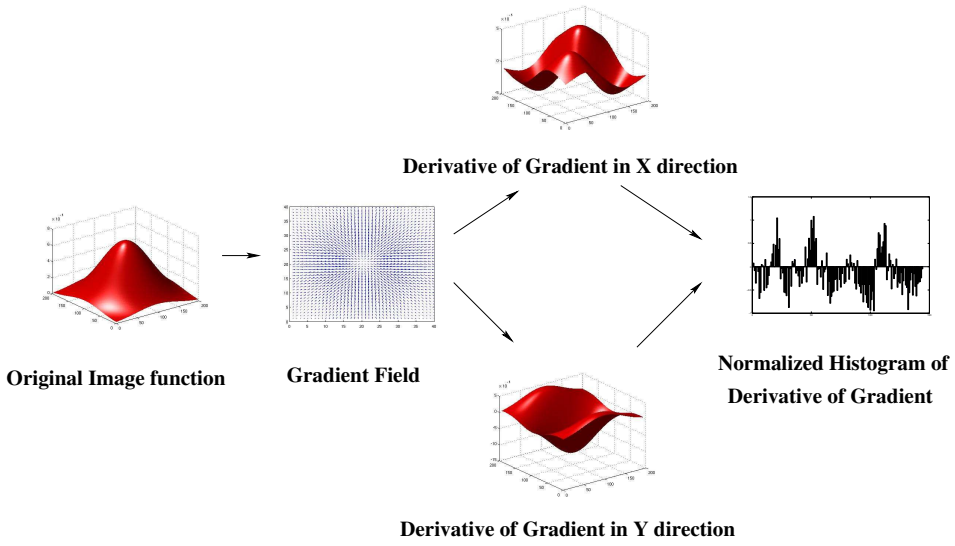


Figure 6.2: Computation of NHoDG feature vector.

6.2.1 Normalized Histogram of Derivative of Gradients directions (NHoDG):

The primary idea is to capture the stochastic variations in the surface orientation, which in turn affects the orientation of projected lines. Hence we propose a feature that computes a histogram of the derivatives of the line directions (gradient directions). The Gradient directions θ in images, computed in equation (A.5), are the directions of maximal intensity variations. In our scenario, the derivative of gradient directions at a pixel can indicate the direction of the deformed projected patterns at that pixel. In equation (A.7) we compute the differential of the gradient directions θ'_1, θ'_2 in both x and y axes to measure the rate at which the surface height varies. The derivatives of gradient directions are computed at each pixel of the projected texture image, and the texture is characterized by a Histogram of the Derivatives of Gradients (HoDG).

The gradient direction derivative histogram is a good indicator of the nature of surface undulations in a 3D texture. For classification, we treat the histogram as a feature vector to compare two 3D textures. As the distance computation involves comparing corresponding bins from different images, we normalize the counts in each bin of the histogram across all the samples in the training set. This normalization allows us to treat the distance between corresponding bins between histograms, equally, and thereby enabling the use of Euclidean distance for comparison of histograms.

$$\begin{aligned} x' &= \tfrac{\delta}{\delta x} I(x,y) \\ y' &= \tfrac{\delta}{\delta y} I(x,y) \end{aligned} \tag{6.1}$$

$$\theta = \tan^{-1}(y'/x') \tag{6.2}$$

$$\begin{aligned} \theta'_1 &= \tfrac{\delta}{\delta x}\theta \\ \theta'_2 &= \tfrac{\delta}{\delta y}\theta \end{aligned} \tag{6.3}$$

The Normalized histograms, or $NHoDG$ is a simple but extremely effective feature for discriminating between different texture classes. Figure 6.2 illustrates the computation of the NHoDG feature from a simple image with bell shaped intensity variation.

We compare the effectiveness of this feature set under structured illumination in the experimental section using a dataset of 30 3D textures.

6.3 Experimental Results and Analysis

The experimental setup consists of a planar surface to place the object samples, an LCD projector fixed at an angle to the object surface, and a camera located directly above the object with its axis perpendicular to the object plane (see setup Figure in chapter 3).

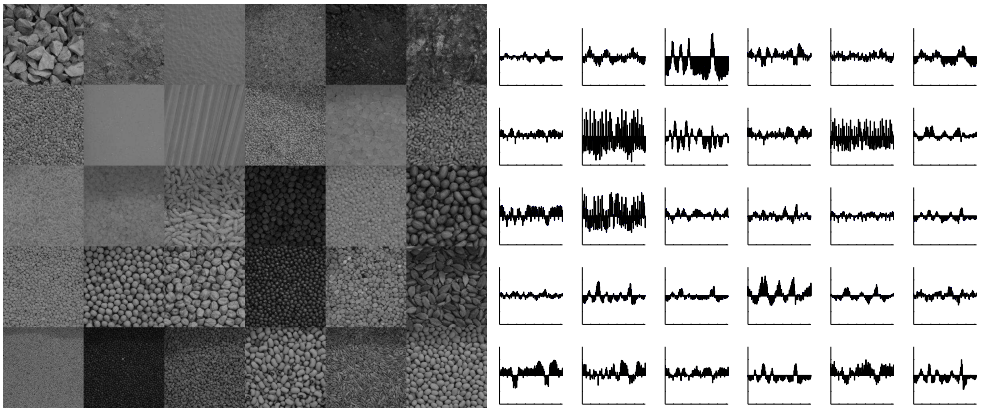
6.3.1 Dataset

We considered a set of thirty 3D textures which has considerable depth variation. Details of each texture is given in Table 6.1 Total 3600 images were collected, with 120 samples for each of the 30 classes. The 120 samples consist of 24 different object

Class Id	Texture Name	Class Id	Texture Name	Class Id	Texture Name
01	Pebble	11	Crystal Sugar	21	Chick Peas
02	Concrete	12	Wheat grain	22	Green gram
03	Thermocol	13	Rice grain	23	Red gram /Pigeon Pea
04	Sand	14	Crystal Salt	24	Cardamom
05	Soil	15	Puffed Rice	25	Poppy seeds
06	Stone	16	Black Gram	26	Mustard seeds
07	Barley	17	Sago	27	Fenugreek seeds
08	Sponge	18	Ground Nut	28	Soybean seeds
09	Ribbed Paper	19	Split Gram beans	29	Fennel/Aniseed
10	Sesame Seed	20	Green Peas	30	White beans

Table 6.1: List of 3D texture surfaces used in our experiments. We have used to set of grains and pulses to create surfaces with similar scale of depth variations, which makes the classification problem, challenging.

samples, each taken with different projected patterns and illumination conditions. The projected patterns are parallel lines having uniform spacing of $5, 10, 15$ and 20 pixels between them. We call these patterns as W5,W10,W15 and W20 for rest of our experimental validation part.



(a) (b)

Figure 6.3: Examples of textures from the 30 classes and their NHoDG representations.

Our data has large scale variation across the textures, while having several surfaces with similar depth variation profiles, thereby making the recognition task very challenging. However, we have not varied the pose of imaging as the application under consideration require controlled illumination conditions. The illumination variations are also limited due to this fact. Images of the thirty different classes is shown in Figure 6.3. Their NHoDG feature representations are shown in Figure 6.3.

For the purpose of comparison, we have also tested the proposed feature set on a standard dataset (Curet [130]). However, note that we could perform the tests only without projected texture and hence the results are not indicative of the power of the proposed approach.

6.3.2 Basic Implementation

We have run our experiments with and without projection patterns, as well as using the proposed and traditional 2D features. We have also implemented the patch based approach proposed in [131]. As the texton dictionary [131] that can be obtained from filter response or image patch is one of the best performing 2D image feature sets, we have used it for comparison with our approach. We have included two more filters with higher scale into original MR8 to make it MR12, so as to improve the results of texton dictionary on our dataset with higher scale variation. A maximum 50 iteration were used for k-means clustering. A brief description of the texton based feature set can be found in Chapter 2 (details can be found in [11]).

Table 6.2: Error rates of classification using NHoDG, MR, and Image Patch features on the PTD and Curet datasets (in %).

		NHoDG	MR	Image Patch		
Dataset	Projection			3x3	5x5	7x7
Curet	Without	12.93	**3.15**	4.67	4.38	3.81
PTD	Without	2.36	**1.18**	3.51	1.53	1.46
	With	1.15	**0.76**	1.60	1.18	0.90
	Combined	**0.07**	0.31	1.70	0.66	0.62

Table 6.2 gives a detailed view of the results using MR12, image patch based and the proposed NHoDG feature set on our dataset as well as the Curet dataset [130]. However, note that the performance on the Curet dataset is without projected patterns. All the results are based on a 4-fold cross validation, where the dataset is divided into non-overlapping training and testing sets, which is repeated 4 times and the average results are reported.

We have also tested the stability of our approach by varying various parameters like the histogram bin size and the width of projected strip pattern. The first plot in Figure 6.4 shows the variation in classification performance as the histogram bin sizes vary, while the second graph shows classification performance when the pattern separation are varied. We note that the performance is consistently good, and we select a bin size of 5 degrees and pattern separation of 5 pixels for the rest of the experiments.

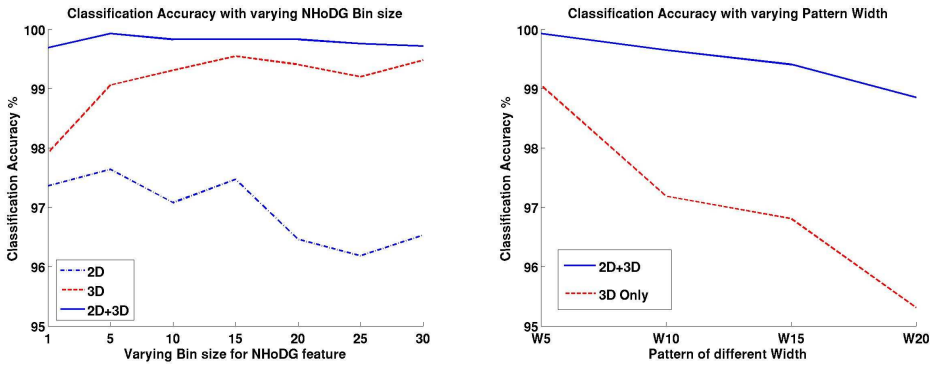


Figure 6.4: Classification performance with varying histogram bin sizes and with varying pattern separation

We note that the 2D image based approach achieves an error rate of 1.18% (*i.e.*, 34 misclassification on our dataset of 2880 samples). In comparison, the projection based approach with NHoDG features achieves an error rate of 0.07% when combined with 2D images, which corresponds to just 2 samples being misclassified. Also, the error rate achieved by patch based approach is 1.46% on natural texture set and 0.62 on combined 2D and projected texture set, which is a clear improvement in such a closer range of error rates. Thus, the results presented in Table 6.2 and **??** shows superiority of our approach on top of state of the art filter-bank and patch based texton approach on our 3D texture dataset.

Figure 6.5 shows one of the misclassified samples, and the corresponding NHoDG and Texton features. We also note that the proposed feature set is primarily intended for use with projection and does not perform well on datasets such as Curet, without projected patterns.

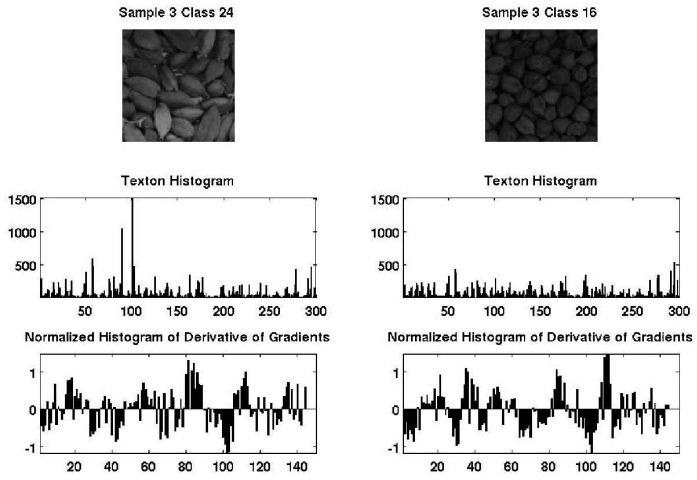


Figure 6.5: One of the two misclassification in the dataset using NHoDG feature set.

Although there is a little pose variation our dataset still it is challenging enough, as the performance of texton approach proves this fact.

6.4 Summary

We have proposed an efficient and robust feature on top of idea of projected texture. The feature captures the stochastic depth variation in surface and represents it by histogram of derivative of gradient directions. Our results show superiority of our approach over the state of the art results and we have also contributed to area by collecting a large dataset for 3D texture.

Chapter 7

Conclusions

7.1 Primary Contributions

Our primary contribution as explained in previous chapters is the proposed framework of *Projected Texture*, which allows us to incorporate shape information of object/surface implicitly for the process of object/surface recognition/classification. The central idea is to project certain structured light patterns on 3D surfaces which get deformed due to the geometry, physical property like reflectance and natural texture of the surface. This resultant texture is referred to as projected texture, which essentially encode the depth profile as well as other properties of the surface. Now we can use this texture as an identity of the surface and can apply generic texture based recognition approaches for object recognition. This is a novel attempt to characterize a surface by projecting structured light and treating the resultant texture as identity texture for the surface, as most of the previous surface tried to use structured light for modeling the surface.

Along with this framework, we have also proposed three different features for three different class of 3D object recognition; 3D objects with Fixed Pose, Arbitrary Pose, and 3D Textured surface (see chapter 4, 5 and 6 for details). The three features proposed in this thesis are :

- Gabor Feature for Fixed Pose Object Recognition
  A window based Gabor feature is proposed for fixed posed objects which captures local frequencies in multiple orientations and represent them as a mean value over the window.

- Fourier Feature for Arbitrary Pose Object Recognition
  A 2-D Fourier transform based feature on top of concept of "bag of words" is proposed. The feature captures the local deformations and represent them as logarithm of magnitude of Fourier cofficient of a window path defined over object, and

- NHoGD Feature for 3D Texture Classification
  A statistical feature is proposed which captures the stochastic deformations by computing derivative of gradient directions and build a histogram on top this.

7.2 Limitations and Future Works

Like any other approach our's also has some limitations. We will discuss some of the important limitations and will see how can we overcome them as future extensions to the current work.

- The use structured light constrains the applicability of this approach as a controlled environment is required. This is a common limitation to most of the structured lighting based systems. This can be partially overcome by using Infrared and other non-visible spectra of light that are not present in the environment.
- Highly specular and transparent surface are still a problem for this approach, although within some limit these properties of the object surface are utilized for recognition in proposed framework. A possible approach to handle such special surfaces could be to project patterns on to the worlds which are reflected by the specular surfaces, or measure the refraction of the patterns by transparent objects.
- Another important direction to explore is the use of adaptive patterns, which can be designed for a specific set of object/category. It can be done within discrimination framework to get an optimal performance of classification/recognition.
- Use of more complex classifiers is another possible work, as we have shown the results using a simple Nearest Neighbor classifier to demonstrate the efficiency and robustness of our approach. We strongly feel that by introducing advanced learning methods, the recognition performance can be increased significantly.

Appendix A

Appendix

A.1 Imaging Process with Pinhole Camera Model

The process of image formation in camera models is inspired by human visual system. Light emanating from a light source get reflect from the object surface and falls on the image plane after passing eye lens. Lens helps to control the image formation parameter known as focal length. In this section we will consider the pin-hole camera where the assumption is, the lens is a infinitely small opening (also known as pinhole) through which only single ray can pass. In reality, the pinhole has a finite (albeit small) size, and each point in the image plane collects light from a cone of rays subtending a finite solid angle. This idealize a extremely simple model of imaging geometry, also known as *pinhole perspective* projection model.

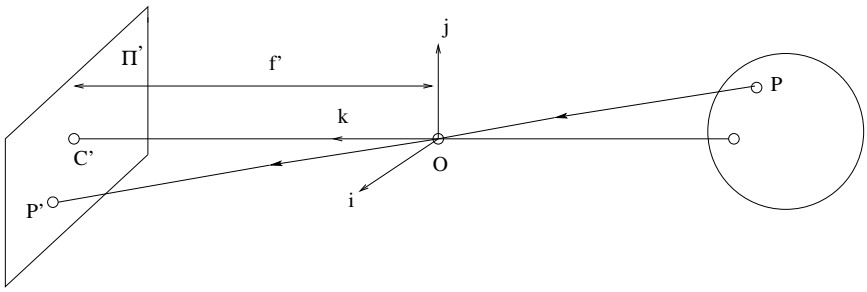


Figure A.1: Perspective Projection Geometry (Courtesy[17]).

Figure A.1 Let P be one a scene point with coordinates (x, y, z) and P' denote its image with coordinates (x', y', z'). Since P' lies in the image plane, we have $z' = f'$. Since the three points P, O, and P' are collinear, we have $\vec{OP'} = \lambda \vec{OP}$, for some number λ, so

$$\lambda = \frac{x'}{x} = \frac{y'}{y} = \frac{z'}{z}$$

and therefore,

$$x' = f'\frac{x}{z} \qquad y' = f'\frac{y}{z} \tag{A.1}$$

In reality equation (A.1) is valid only when all distance are measure in camera's reference frame and image coordinates have their origin at the principal point where the axis of symmetry of the camera pierces its retina. In practice, the world and camera coordinate system are related by a set of physical parameters, such as the

focal length of of th lens, the size of pixels, the position of the principal point and the position and orientation of the camera. We will not derive these parameters, but will give a final expression for them. Figure A.2 illustrate normalized and physical image coordinate system representations used for deriving intrinsic and extrinsic parameters.

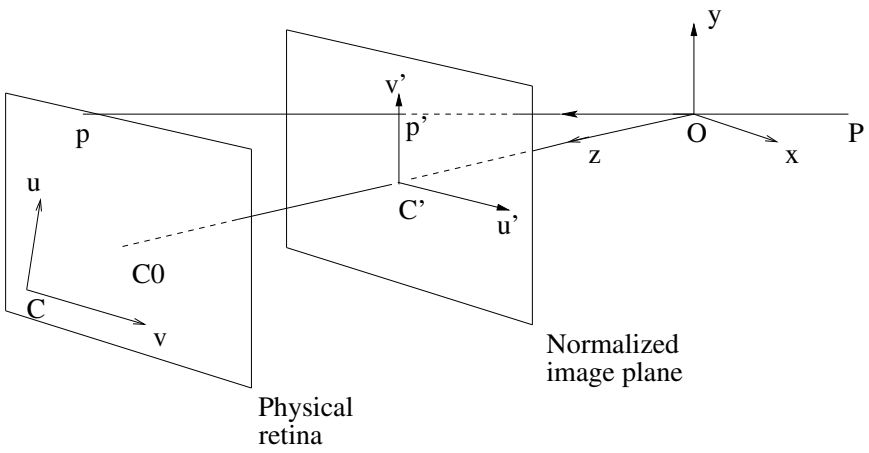


Figure A.2: Imaging Coordinate System for a Pinhole camera (Courtesy[17]).

Intrinsic Parameter

Intrinsic parameters are the parameters which relate the camera's coordinate system to idealize coordinate system explained in start of the section. Total 5 intrinsic parameter are derived. These can be represented as a matrix K which is a part of final camera matrix M.

$$K = \begin{bmatrix} \alpha & -\alpha \cot\theta & u_0 \\ 0 & \frac{\beta}{\sin\theta} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{A.2}$$

Where $\alpha = kf, \beta = lf$ are magnification parameter derived from scaling factor $kandl$. $u_0 and v_0$ are the coordinates of principal point C_0, which is not coincide with center of real world image plane (*i.e.* CCD matrix). The parameter θ is the skew factor.

Extrinsic Parameter

Extrinsic parameters relates camera's coordinate system to a fixed world coordinate system and specify its position and orientation in space. There are 6 extrinsic parameters derived, three from rotation $(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ and three from translation (t_1, t_2, t_3).

$$R = [\mathbf{r}_1 \mathbf{r}_2 \mathbf{r}_3]^T$$

$$t = [t1t2t3]^T$$

Thus camera matrix M can be written as combine effect of intrinsic and extrinsic parameters

$$M = K(R, \mathbf{t}) \tag{A.3}$$

and imaging of real point P into image plane point p as

$$p = \frac{1}{z} MP \tag{A.4}$$

Thus, imaging process can be understood as a transformation of points from real world to image plane using the camera matrix M. A detailed discussion on imaging process can be found in [17].

A.2 Shape from X

In computer vision, the techniques to recover shape are called *shape-from-X* techniques, where X can be stereo, texture, shading, motion and defocus. We will briefly discuss each of the method one by one. A detailed literature survey on these methods can be found at [132]

A.2.1 Shape from Stereo

Stereo vision refers to the ability to infer information on the 3-D structure and distance of a scene from two or more images taken from different viewpoints. From a computational standpoint, a stereo system must solve two problems. The first, known as correspondence, consist of finding location of image of a part in two or more image planes. A rather subtle difficulty here is that some parts of the scene are visible only in one image. Therefore, a stereo system must also be able to determine the image parts that should not be matched.

The second problem that a stereo system must solve is reconstruction. Disparity is the shift in position of image of a real world point in two image plane. The disparities of all the image points form the so-called disparity map, which can be displayed as an image. If the geometry of the stereo system is known, the disparity map can be converted to a 3-D map of the viewed scene (the reconstruction).

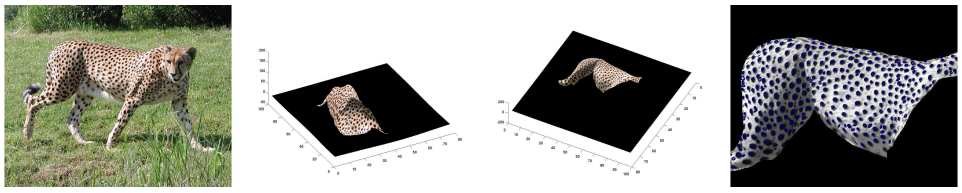


Figure A.3: Shape from Texture (Courtesy [18]).

A.2.2 Shape from Texture

In this technique shape of an 3D object is recovered from 2D images by using the texture information. Although human is capable of realize patterns, estimate depth and recognize objects in an image by using texture as a cue, development of an automated system which mimic human behavior is far from trivial. Texture is defined as pattern formed by repetitive occurrence of a basic element known as texels (TEXture ELement).

The basic idea is to first identify texels and obtain the surface normal at texel position corresponding to the deformation in each of the texel element. The method exploits the perspective distortion, which makes objects far from the camera appear smaller, and foreshortening distortion, which makes objects not parallel to the image plane shorter. The amount of both distortions can be measured (shape distortion and distortion gradient) from an image. A map of surface normals specifies the surface's

orientation only at the points where the normals are computed. But, assuming that the normals are dense enough and the surface is smooth, the map can be used to reconstruct the surface shape. Figure A.3 illustrates some results on shape recovery from texture, obtained by [18].

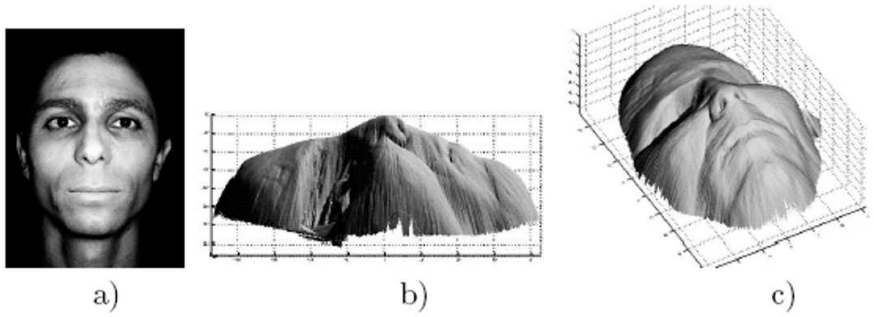


Figure A.4: Shape from Shading Results (Courtesy [19]).

A.2.3 Shape from Shading

The basic principle behind Shape-from-Shading (*SFS*) is to recover depth from a gradual variation of shading in the image. The idea is quite old and used by artists to convey vivid illusions of depth in paintings. It is important to study how the images are formed and reflectivity of surface in-order to understand SFS approach. The Lambertian model is a simple model, in which the gray level at a pixel in the image depends on the light source direction and the surface normal. In SFS, given a gray level image, the aim is to recover the light source and the surface shape at each pixel in the image. However, real images do not always follow the Lambertian model. Even if we assume Lambertian reflectance and known light source direction, and if the brightness can be described as a function of surface shape and light source direction, the problem is still not simple. This is because if the surface shape is described in terms of the surface normal, we have a linear equation with three unknowns, and if the surface shape is described in terms of the surface gradient, we have a non-linear equation with two unknowns. Therefore, finding a unique solution to SFS is difficult it requires additional constraints. Figure A.4 shows shape from shading results presented in [19].

A.2.4 Shape from Motion

In this section we will discuss the Shape from Motion where the shape of a scene is extracted from the spatial and temporal changes occurring in an image sequence. This technique exploits the relative motion between camera and scene and can be divided into subprocess of finding the correspondence from consecutive frames and reconstruction of the scene. In this approach the relative 3D displacement between the viewing camera and the scene is not necessarily caused by a single 3D transformation. Also the average displacement in two consecutive frames is much more smaller than typical stereo images.

Two kinds of methods are commonly used to compute the correspondence computation. One is the Differential methods that uses estimates of time derivatives

and require therefore image sequences sampled closely. This method leads to dense measurements. Second type of Matching methods use Kalman filtering to match and track efficiently sparse image features over time. This method produces sparse measurements.The Reconstruction is more difficult in motion than in stereo. Frame-by-frame recovery of motion and structure turns out to be more sensitive to noise to small baseline between consecutive frames. For reconstruction we can use the motion field of the image sequence. The motion field is the projection of the 3D velocity field on the image plane. One way to acquire the 3D data is to determine the direction of translation through approximate motion parallax. Afterwords, we can determine a least-squares approximation of the rotational component of the optical flow and use it in the motion field equations to compute depth.

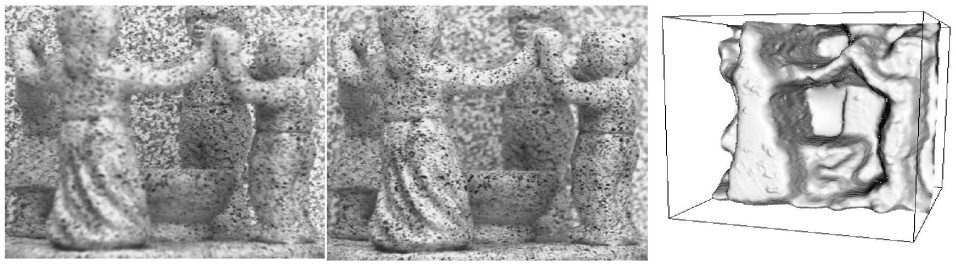


Figure A.5: Shape from Defocus results (Courtesy [20]).

A.2.5 Shape from focus/Defocus

In depth from focus/defocus approaches recovers the 3D shape from two or more images of the scene, which are obtained by same position but with different imaging parameters like focal setting or the image plane axial position. The difference between depth from focus and depth from defocus is that, in the first case it is possible to dynamically change the camera parameters during the surface estimation process, while in the second case this is not allowed. These can also be called active and passive depending upon whether whether it is possible or not to project a structured light onto the scene. The basic ideas is to use real aperture cameras instead of pin-hole cameras. They have a short depth of field, resulting in images which appear focused only on a small 3D slice of the scene. In this cased the image process formation can be explained with optical geometry. The lens is modeled via the thin lens law. Figure A.5 presents depth map computed using shape from defocus by [20].

A.3 Fourier Transform

In this section, we present some of the basics of Fourier transforms in brief. For more details, one can refer to [133, 134]

The theory of Fourier series is based on the idea that most signals, and all engineering signals, can be represented as a sum of sine waves. One of the interesting property of a sine wave is that it allows to do many natural operations on set of different frequencies as if each signal were processed individually (they are linear with regard to frequency). This essentially allows us to apply these operations to individual sine waves and merely add and multiply to look at the effect on the entire signal.

Fourier representation helps us to obtain an overall picture of the content in a signal. It helps to separate high and low frequency components. This is very useful in edge detection as edges appear on fast changing boundary, which is associated with high frequency components in the 2D image signal. So by removing low frequency components from the Fourier representation, an edge map of the image can be recovered. It also helps to eliminate noise of known frequency from the data. All these operations can be done on the frequency map of signal, also known as the spectrum. The spectrum of a signal shows the strength of the individual frequencies in the signal. A spectrum representation has two sides: the negative side on the left, and the positive side on the right. The negative side contains negative frequencies. For real signals (with no imaginary part), like audio signals, the negative side of the spectrum is always a mirrored version of the positive side.

A.3.1 One-dimension Fourier Transform

The Continuous Fourier Transform, for use on continuous 1D signals $f(x)$, is defined as follows:

$$F(w) = \int_{-\infty}^{\infty} f(x) e^{-2\pi wix} dx$$

And the Inverse Continuous Fourier Transform, which allows you to go from the spectrum back to the signal, is defined as:

$$f(x) = \int_{-\infty}^{\infty} F(w) e^{2\pi wix} dw$$

$F(w)$ is the spectrum, where w represents the frequency, and $f(x)$ is the signal in the time, where x represents the time. The similarity between the forward and inverse transforms indicates the duality between a signal and its spectrum.

A computer works in discrete domain with finite number of discrete points sampled from a continuous signal. One of the properties of the Fourier Transform and its inverse is, that the FT of a discrete signal is periodic. Since for a computer, both the signal and the spectrum must be discrete, both the signal and spectrum will be periodic also. So in the discrete case, it is assumed that signal is infinitely repeated for a 1D signal; or infinitely tiled for a 2D image. One important and helpful property of the transform is that both the signal and the spectrum will have the same number of discrete points.

Since the signal is finite in time, the infinite borders of the integrals can be replaced by finite ones, and the integral symbol can be replaced by a sum. So the DFT for one-dimensional discrete signal is defined as:

$$F_n = \sum_{k=0}^{N-1} f_k e^{-2\pi ink/N}$$

And the inverse DFT as:

$$f_k = \frac{1}{N} \sum_{n=0}^{N-1} F_n e^{2\pi ikn/N}$$

A.3.2 Properties of The Fourier Transform

A signal is often denoted with a small letter, and it's Fourier transform or spectrum with a capital letter. The relation between a signal and its spectrum is often denoted with $f(x) \Longleftrightarrow F(w)$, with the signal on the left and it's spectrum on the right.

Linearity

$$f(x) + g(x) \Longleftrightarrow F(w) + G(w)$$
$$a * f(x) \Longleftrightarrow a * F(w)$$

This means that if you add/subtract two signals, their spectra are added/subtracted as well, and if you increase/decrease the amplitude of the signal, the amplitude of it's spectrum will be increased/decreased with the same factor.

Scaling

$$f(a * x) \Longleftrightarrow (1/a)F(w/a)$$

This means that if you make the function wider in the x-direction, it's spectrum will become smaller in the x-direction, and vice-versa. The amplitude will also be changed.

Time Shifting

$$f(x - x_0) \Longleftrightarrow e^{-iwx_0}F(w)$$

Since the only thing that happens if you shift the time, is a multiplication of the Fourier Transform with the exponential of an imaginary number, you won't see any difference of a time shift in the amplitude of the spectrum, but only in its phase.

Frequency Shifting

$$e^{-iw_0x}f(x) \Longleftrightarrow F(w - w_0)$$

This is the dual of the time shifting.

Duality

if $f(x) \Longleftrightarrow F(w)$
then $F(x) \Longleftrightarrow f(-w)$

For example, apart from some scaling factors, the spectrum of a rectangular pulse is a sinc function and at the same time the spectrum of a sinc function is a rectangular pulse.

A.3.3 2D Fourier Transform

One-dimensional Fourier transform can be directly extended to two dimensions. The 2D Fourier Transform of a two-dimensional continuous function $f(x, y)$ can be expressed as,

$$F(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)e^{-i2\pi(ux+vy)}dxdy$$

Similarly expression for Inverse Fourier Transform is,

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{i2\pi(ux+vy)}dudv$$

In discrete domain, where two-dimensional function $f(x,y)$ represents an image, the expression of 2D Discrete Fourier Transform (DFT) is,

$$F(u,v) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2\pi(ux/M+vy/N)}$$

and Inverse DFT cab be expressed as,

$$f(x,y) = \sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F(u,v)e^{j2\pi(ux/M+vy/N)}$$

where M and N are the dimensions of the image. Note that the value of the transform at $(u,v) = (0,0)$ for a gray scale image is, the average gray level of the image:

$$F(0,0) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)$$

All other properties are similar to the one-dimensional Fourier Transform.

A.3.4 Difference in Amplitude and Phase Spectra

Fourier transform of a function is a function of complex variables. We can separate real and imaginary part using polar representation. Let $F(u)$ represents the Fourier transform of a continuous one-dimensional function $f(x)$. Then $F(u)$, which is a complex function, can be represented in polar form as

$$F(u) = A(u)e^{j\phi(u)}$$

where

$$A(u) = |F(u)|$$

is the amplitude of Fourier response $F(u)$ and

$$\phi(u) = arg(F(u))$$

is the phase of Fourier response $F(u)$.

Similarly, for a two-dimensional function, we will get two-dimensional amplitude and phase spectra. In case of Fourier transform, the amplitude gives the strength of each frequency component in the image, while the phase information of each of the frequency component encodes the precise location of objects in the image.

A.3.5 Convolution Theorem

The most fundamental relationship between the spatial and frequency domains is established by a well-known result called the *convolution theorem*. Formally, the two-dimensional discrete convolution of two functions $f(x, y)$ and $h(x, y)$ of size $M \times N$, denoted by$f(x, y) * h(x, y)$, is

$$f(x, y) * h(x, y) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n)h(x - m, y - n)$$

The basic steps of convolution process are : (a) Flip one function about the origin, (b) Shifting that function *w.r.t.* to the other by changing the value of (x, y), and (c) Compute a sum of products over all values of m and n for each displacement.

Letting $F(u, v)$ and $H(u, v)$ denote the Fourier transform of $f(x, y)$ and $h(x, y)$, respectively, the convolution theorem states that $f(x, y) * h(x, y)$ and $F(u, v)H(u, v)$ constitutes a Fourier transform pair. Or formally stating, convolution in spatial domain is multiplication in frequency domain.

$$f(x, y) * h(x, y) \Longleftrightarrow F(u, v)H(u, v)$$

The theorem also states that, multiplication in spatial domain is convolution in frequency domain.

$$f(x, y)h(x, y) \Longleftrightarrow F(u, v) * H(u, v)$$

A.4 Gabor Filter

Nobel laureate Dennis Gabor first proposed that a signal can be represented as the combination of elementary functions, now known as *Gabor functions*. Since then Gabor functions have played an important part in various fields of theory of communication. Specifically, wavelet based image analysis became popular, Gabor wavelets have been used extensively in image characterization and processing. In this section, we will discuss the process of local frequency characterization using Gabor filters. A detailed description can be found in work by Joni-Kristian Kmrinen [135] and Ville Kyrki [21].

Time and frequency are two fundamental domains. They are physically measurable quantities, but still idealizations if one is considered from the other's perspective. Frequency can be represented as a simple waveform in the time domain. In order to define it precisely, the wave must be infinite in time domain. Similarly, for an event to be precisely localized in time domain, an infinite set of frequencies are required. Thus a duality between the domains hold. Real world functions do not follow either of the above phenomena exactly, but shares properties from both. They certainly have some frequency characteristics, and are bounded in a range of time. So a real world signal needs a description that allows combination of time-frequency representation. The Gabor function allows the optimum trade-off between localization in time and frequency domains among all possible representations.

A.4.1 Gabor Elementary Function

Gabor elementary function enables combined analysis of time and frequency aspects of a signal. Let the effective widths of a signal in time and frequency domains be represented by, Δt and Δf. These widths are defined based on the variance of the signal in the corresponding domains. This leads to the definition of minimal uncertainty as the product of the two:

$$\Delta t \Delta f \geq \frac{1}{2}$$

Gabor then showed that the signal that minimizes the product $\Delta t \Delta f$, thus turning the inequality into an equality, is:

$$\psi(t) = e^{(-\alpha^2(t-t_0)^2)} e^{(-j(2\pi f_0 t + \phi))}, \tag{A.5}$$

where j is the imaginary unit. The above function represents a complex sinusoidal wave modulated by a Gaussian function. α, t_0, f_0 and ϕ are constants that denote the sharpness of the Gaussian, center of the Gaussian in time domain, the frequency of the sinusoid, and its phase shift.

Applying Fourier transform to equation A.5 gives

$$\Psi(f) = \frac{\sqrt{\pi}}{\alpha} e^{-\left(\frac{\pi}{\alpha}^2\right)(f-f_0)^2)} e^{(-j(-2\pi t_0(f-f_0)+\phi)}, \tag{A.6}$$

which also is a complex sinusoid modulated by a Gaussian. The function in equation A.5 is known as the Gabor elementary function (GEF). This is used to analyze signals by decomposing them into a number of GEFs. The decomposition includes the Fourier analysis and the time-domain analysis as a special case.

$$\alpha \to 0 \Rightarrow \psi(t) \to sinusoid \Rightarrow Fourier \;\; analysis$$

$$\alpha \to \infty \Rightarrow \psi(t) \to Dirac - delta \Rightarrow time \;\; description$$

The Gabor decomposition is a predecessor of multi-resolution and wavelet analysis. The decomposition of signal is computationally difficult when used as wavelet basis as the GEF's are not orthogonal. The solution to this problem is that instead of decomposition, signals can be analyzed by convolving them with GEFs. The filter response to input $\xi(t)$, that is the convolution, can be defined as:

$$r_\xi(t) = \int_{-\infty}^{\infty} \psi(t-\tau)\xi(\tau)d\tau$$

A.4.2 Two-dimensional spatial-frequency space

In 1978, Granlund presented a two-dimensional counterpart of GEF as a general picture processing operator, arguing that this operator could detect and describe structure at different level. However, the resemblance of two-dimensional GEF with the receptive fields of simple cells in a mammalian visual system, was the main reason of increased the popularity of Gabor analysis in the computer vision community. When the GEFs are used as filters, the filter can be centered at origin and the phase and time shift parameters, ϕ and t_0, dropped. The filtering will be considered in continuous domain for simplicity.

The two-dimensional Gabor filter is a complex sinusoidal plane wave modulated by an elliptical Gaussian probability density function. Following Gabor's formulation for the one-dimensional GEF, a two-dimensional Gabor filter $\psi(x, y)$ can be defined as :

$$\psi(x, y) = e^{-\left(\alpha^2 x'^2 + \beta^2 y'^2\right)} e^{j2\pi f_0 x'}, \tag{A.7}$$

where

$$x' = x\cos\theta + y\sin\theta,$$
$$y' = -x\sin\theta + y\cos\theta,$$

f_0 is the frequency of the sinusoidal plane wave, θ is the anti-clockwise rotation of the Gaussian and the plane wave, α sharpness of the Gaussian along the axis parallel to the wave, and β is the sharpness along the axis perpendicular to the wave. The 2-D Gabor filter representation in equation A.7 is not in the most general form of the 2-D Gabor elementary function but the orientation of the elliptical Gaussian is the same as the orientation of the plane wave. In addition, the shifts in space and phase have been dropped because the function is used as a convolution filter. The response to input image $\xi(x, y)$ is then

$$r_\xi(x, y) = \int\int_{-\infty}^{\infty} \psi(x - x', y - y')\xi(x', y')dx'dy'$$

The filter in equation A.7 can be normalized by fixing the ratio of the frequency of the wave and the sharpness values of the Gaussian, *i.e.* , $\gamma = \frac{f_0}{\alpha}, \eta = \frac{f_0}{\beta}$. This fixes the number of waves in spatial filter to a constant value. This formulation makes the DC-response identical for all frequencies as it fixes the behavior of the response regardless of the frequency. It is desired that the DC-response is small, otherwise the average image intensity affects the response. This can be controlled by setting parameter γ large enough. To make the area under the Gaussian unity, a normalization factor $\frac{\alpha\beta}{\pi}$ has to be used. Thus, a normalized filter can be presented as

$$\psi(x, y) = \frac{f_0}{\pi\gamma\eta} e^{-\left(\frac{f_0^2}{\gamma^2}x'^2 + \frac{f_0^2}{\eta^2}y'^2\right)} e^{j2\pi f_0 x'} \tag{A.8}$$

With application of Fourier transform, equation A.8 can be represented in the frequency domain as

$$\Psi(u, v) = \frac{\pi^2}{f_0^2} e^{\left(\gamma^2(u' - f_0)^2 + \eta^2 v'^2\right)} \tag{A.9}$$

$$u' = u\cos\theta + u\sin\theta$$
$$v' = -v\sin\theta + v\cos\theta$$

Thus, in the frequency domain the filter is a real Gaussian with centroid at frequency f_0 at orientation θ. The parameters of a frequency domain filter are illustrated in figure A.6.

A Gabor filter bank is a combination of several Gabor filters with varying parameters like frequency and orientation. Usually filter banks have equal orientation spacing

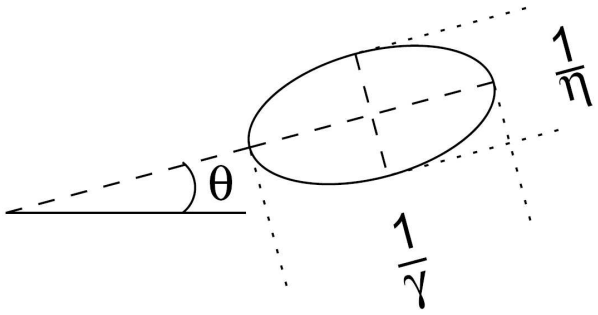


Figure A.6: Gabor filter parameters in frequency domain (Courtesy[21]).

and octave frequency spacing, while the relative width γ and η stay constant. That is,

$$\theta_k = \frac{k\pi}{n_\theta} \quad where \quad k = \{0, 1, \cdots, n_\theta - 1\}$$

$$f_l = s^{-l} f_{max} \quad where \quad l = \{0, 1, \cdots, n_f - 1\},$$

where θ_k is k^{th} orientation, n_θ is number of orientations, f_l the l^{th} frequency, f_{max} the maximal frequency, n_f the number of frequencies, and $s(> 1)$ the ratio between two consecutive frequencies. Only a half of the frequency plane needs to be covered, because the input to the filters is assumed to be real and thus its frequency representation is symmetric and Hermitian.

A.4.3 Properties of Gabor Filter

Although parameter selection is a problem, still Gabor filters are very popular due to their properties. First, they can be used to extract various kinds of visual features, including texture, edges, lines, and shapes. Similarity with the simple cells of the mammalian visual cortex is another supporting factor. Next is the robustness of Gabor responses as the amplitudes of complex Gabor coefficients are invariant under small translation, rotation, and scaling. Shiftability of the Gabor filters in spatial and frequency domains is a good explanation for it, but it requires a nonorthogonal, over-complete representation. In addition, shiftability makes it possible to interpolate responses both in spatial and frequency coordinates. Also, the Gaussian nature of the filters makes them tolerant to noise. However, primarily only the amplitude of the response is considered and not the individual real and imaginary parts.

Related Publications

- Avinash Sharma and Anoop Namboodiri, "Projected Texture for Object Classification", to appear in *European Conference on Computer Vision* (ECCV 2008), October 12-18, Marseille, France.
- Avinash Sharma, Nishant Shobhit and Anoop Namboodiri, "Projected Texture for Hand Geometry based Authentication", *CVPR Workshop on Biometrics* (CVPRW 2008), June 28, Anchorage, Alaska, USA. IEEE Computer Society 2008.
- Avinash Sharma and Anoop Namboodiri, "Object Category Recognition with Projected Texture", Submitted to *Indian Conference on Vision, Graphics and Image Processing* (ICVGIP 2008), (Results Awaited).

Bibliography

[1] M. Oren and S. K. Nayar, "A theory of specular surface geometry," *Internationa Journal Computer Vision*, vol. 24, no. 2, pp. 105–124, 1997.

[2] R. Ramamoorthi, "Modeling Illumination Variation with Spherical Harmonics," in *Face Processing: Advanced Modeling Methods*. 2006, pp. 385–424.

[3] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, ACM Press/Addison-Wesley Publishing Co., 2000, pp. 145–156.

[4] M. Takeda and K. Mutoh, "Fourier transform profilometry for the automatic measurement of 3-d object shapes," *Applied Optics*, vol. 22, no. 24, 1983.

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

[6] T. P. Koninckx, "Adaptive structured light," Ph.D. dissertation, KATHOLIEKE UNIVERSITEIT LEUVEN, May 2005.

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

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[9] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computer*, vol. 22, no. 1, pp. 67–92, 1973.

[10] M. Weber, "Unsupervised learning of models for object recognition," Ph.D. dissertation, California Institute of Technology, Pasadena, 2000.

[11] M. Varma, "Statistical approaches to texture classification," Ph.D. dissertation, University of Oxford, October 2004.

[12] J. Wang and K. J. Dana, "Hybrid textons: Modeling surfaces with reflectance and geometry," in *Proceedings of the IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, vol. 1, 2004, pp. 372–378.

[13] A. K. Jain, A. Ross, and S. Pankanti, "A prototype hand geometry-based verification system," in *Proceedings of the AVBPA'99*, Washington D.C., Mar. 1999, pp. 166–171.

[14] M. Faundez-Zanuy, D. A. Elizondo, M. Ferrer-Ballester, and C. M. Travieso-González, "Authentication of individuals using hand geometry biometrics: A neural network approach," *Neural Processing Letters*, vol. 26, no. 3, pp. 201–216, 2007.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA*, IEEE Computer Society, 2005, pp. 886–893.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Proceedings of International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002.

[18] A. M. Loh, "The recovery of 3-d structure using visual texture patterns," Ph.D. dissertation, School of Computer Science and Software Engineering, University of Western Australia, 2006.

[19] E. Prados, F. Camilli, and O. Faugeras, "A unifying and rigorous shape from shading method adapted to realistic data and applications," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 3, pp. 307–328, 2006.

[20] P. Favaro and S. Soatto, "Learning shape from defocus," in *Proceedings of the 7th European Conference on Computer Vision*, vol. 2, London, UK, Springer-Verlag, 2002, pp. 735–745.

[21] V. Kyrki, "Local and global feature extraction for invariant object recognition," Ph.D. dissertation, Lappeenranta University of Technology, 2002.

[22] I. Biederman, "Visual object recognition," *An Invitation to Cognitive Science*, vol. 2, no. 2, pp. 121–165, 1987.

[23] I. Incorporatoin. "Veggie vision, http://www.research.ibm.com/ecvg/jhc_proj/veggie.html,",".

[24] R. Fergus, "Visual object category recognition," Ph.D. dissertation, University of Oxford, 2005.

[25] J. Posdamer and M. Altschuler, "Surface measurement by space-encoded projected beam system," *CGIP*, vol. 18, pp. 1–17, January 1982.

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

[27] M. Maruyama and S. Abe, "Range sensing by projecting multiple slits with random cuts.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 6, pp. 647–651, 1993.

[28] N. Durdle, J. Thayyoor, and V. Raso, "An improved structured light technique for surface reconstructionof the human trunk," *IEEE Canadian Conference on Electrical and Computer Engineering.*, vol. 2, pp. 874–877, May 1998.

[29] S. J., J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, no. 4, pp. 827–849, 2004.

[30] A. Adan, F. Molina, A. S. Vazquez, and L. Morena, "3d feature tracking using a dynamic structured light system," in *Proceedings of the 2nd Canadian conference on Computer and Robot Vision, Washington, DC, USA*, IEEE Computer Society, 2005, pp. 168–175.

[31] J. Pags, C. Collewet, and J. Chaumette, F.and Salvi, "Robust decoupled visual servoing based on structured light," in *Proceedings of Intternational Conference on Intelligent Robots and Systems,Edmonton, Canada*, vol. 2, IEEE Computer Society, 2005, pp. 2676–2681.

[32] P. J. and B. J., "Towards a surface primal sketch," *Three dimensional machine vision*, pp. 195–240, 1987.

[33] I. Rigoutsos and R. Hummel, "A bayesian approach to model matching with geometric hashing," *Comp. Vis. and Img. Understanding*, vol. 62, pp. 11–26, 1995.

[34] S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby, "Pmf: A stereo correspondence algorithm using a disparity gradient limit," *Perception*, vol. 14, pp. 449–470, 1985.

[35] D. G. Lowe, "The viewpoint consistency constraint," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 57–72, 1987.

[36] M. Jerrum and A. Sinclair, "The markov chain monte carlo method," *Approximation Algorithms for NP-hard Problems*, 1997. In D. S. Hochbaum, editor.

[37] S. Ullman and R. Basri, "Recognition by linear combinations of models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, 1991.

[38] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, no. 115, pp. 115–147, 1987.

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

[40] C. Rothwell, D. Forsyth, A. Zisserman, and J. Mundy, "Planar object recognition using projective shape representation," *International Journal of Computer Vision*, vol. 16, no. 2, 1995.

[41] S. B. and J. L. Crowley, "Object recognition without correspondence using multidimensional receptive field histograms," *International Journal of Computer Vision*, vol. 36, no. 1, pp. 31–50, 2000.

[42] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.

[43] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-100). technical report cucs-006-96," technical report, Columbia University, 1996.

[44] M. Pontil, S. Rogai, and A. Verri, "Recognizing 3-d objects with linear support vector machines," in *Proceedings of the 5th European Conference on Computer Vision, Freiburg, Germany*, 1998, pp. 469–483.

[45] S. C. and R. Mohr, "Local greyvalue invariants for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–534, 1997.

[46] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points.," in *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, 2001, pp. 525–531.

[47] V. Ferrari, T. Tuytelaars, and L. Van Gool, "Simultaneous object recognition and segmentation by image exploration," in *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic*, vol. 1, 2004, pp. 40–54.

[48] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proceedings of the British Machine Vision Conference*, 2002, pp. 384–393.

[49] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. 257–263.

[50] P. Moreels, M. Maire, and P. Perona, "Recognition by probabilistic hypothesis construction," in *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic*, vol. 2, 2004, pp. 55–68.

[51] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets, or ̈how do i organize my holiday snaps?;" in *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, vol. 1, Springer-Verlag, 2002, pp. 414–431.

[52] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using ane-invariant patches and multi-view spatial constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 272–280.

[53] H. Schneiderman and T. Kanade, "A statistical method for 3d object detection applied to faces and cars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 1746–1769.

[54] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA*, 2005, pp. 886–893.

[55] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic*, vol. 1, Springer-Verlag, 2004, pp. 69–82.

[56] R. Fergus and P. Perona. "Caltech object category datasets. http://www.vision.caltech.edu/htmlfiles/archive.html,",", 2003.

[57] S. Agarwal, A. Awan, and D. Roth. "Uiuc car dataset, http://l2r.cs.uiuc.edu/ cogcomp/data/car,",", 2002.

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

[59] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic*, vol. 2, 2004, pp. 71–84.

[60] M. Burl and P. Perona, "Recognition of planar object classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1996, pp. 223–230.

[61] 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*, 1998, pp. 628–641.

[62] 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*, 2000, pp. 20–27.

[63] T. K. Leung, M. C. Burl, and P. Perona, "Finding faces in cluttered scenes using random labeled graph matching," in *Proceedings of the 5th International Conference on Computer Vision, Boston*, 1995, pp. 637–644.

[64] 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,*, 2003, pp. 1134–1141.

[65] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 2066–2073.

[66] 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*, vol. 1, 2005, pp. 10–17.

[67] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, 2002, pp. 113–130.

[68] A. Carlson, C. Cumby, J. Rosen, and D. Roth, "The snow learning architecture," Technical Report uiucdcsr-99-2101, Dept. of Computer Science, UIUC, 1999.

[69] E. Borenstein and S. Ullman, "Class-specific, top-down segmentation," in *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, 2002, pp. 109–104.

[70] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[71] M. Everingham, L. Van Gool, C. Williams, and A. Zisserman. "Pascal visual object challenge datasets. http://www.pascal-network.org/challenges/voc/voc/index.html,",", 2005.

[72] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC*, 2004, pp. 762–769.

[73] B. Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch, "Inability of humans to discriminate between visual textures that agree in second-order statistics -revisited," *Perception*, vol. 2, no. 4, pp. 391–401, 1973.

[74] B. Julesz, "Textons, the elements of texture perception and their interactions," *Nature*, vol. 290, pp. 91–97, 1981.

[75] J. M. Coggins and A. K. Jain, "A spatial filtering approach to texture analysis," *Pattern Recognition Letters*, vol. 3, pp. 195–203, 1985.

[76] T. Caelli and G. Moraglia, "On the detection of gabor signals and discrimination of gabor textures," *Vision Research*, vol. 25, no. 5, pp. 671–684, 1985.

[77] O. D. Faugeras, "Texture analysis and classification using a human visual model," in *Proceedings of the International Conference on Pattern Recognition, Kyoto, Japan*, 1978, pp. 549–552.

[78] I. Fogeli and D. Sagi, "Gabor filters as texture discriminator," *Biological Cybernetics*, vol. 61, pp. 102–113, 1989.

[79] K. L. Laws, "Textured image segmentation," Ph.D. dissertation, University of Southern California, 1980.

[80] M. R. Turner, "Texture discrimination by gabor functions," *Biological Cybernetics*, vol. 55, pp. 71–82, 1986.

[81] M. Unser, "Local linear transforms for texture measurements," *Signal Processing*, vol. 11, no. 1, pp. 61–79, 1986.

[82] W. T. Freeman, "Steerable filters and the local analysis of image structure," Ph.D. dissertation, MIT, 1980.

[83] P. Perona, "Steerable-scalable kernels for edge detection and junction analysis," in *Proceedings of the European Conference on Computer Vision, Ligure, Italy*, 1992, pp. 3–18.

[84] P. Perona, "Deformable kernels for early vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 488–499, 1995.

[85] H. Greenspan, S. Belongie, P. Perona, and R. Goodman, "Rotation invariant texture recognition using a steerable pyramid," in *Proceedings of the International Conference on Pattern Recognition, Jerusalem, Israel*, vol. 2, 1994, pp. 162–167.

[86] G. M. Haley and B. S. Manjunath, "Rotation-invariant texture classification using modified gabor filters," in *Proceedings of the IEEE International Conference on Image Processing, Washington, DC*, vol. 1, 1995, pp. 262–265.

[87] J. R. Smith and S. F. Chang, "Transform features for texture classification and discrimination in large image databases," in *Proceedings of the IEEE International Conference on Image Processing, Austin, Texas.*, vol. 3, 1994, pp. 407–411.

[88] T. Leung and J. K. Malik, "Recognizing surfaces using three-dimensional textons," in *Proceedings of the International Conference on Computer Vision, Kerkyra, Greece*, vol. 2, 1999, pp. 1010–1017.

[89] T. Leung and J. K. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[90] O. G. Cula and K. J. Dana, "Compact representation of bidirectional texture functions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, vol. 1, 2001, pp. 1041–1047.

[91] O. G. Cula and K. J. Dana, "3d texture recognition using bidirectional feature histograms," *International Journal of Computer Vision*, vol. 59, no. 1, pp. 33–60, 2004.

[92] F. Schalitzky and A. Zisserman, "Viewpoint invariant texture matching and wide baseline stereo," in *Proceedings of the International Conference on Computer Vision, Vancouver, Canada*, vol. 2, 2001, pp. 636–643.

[93] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence," in *Proceedings of the European Conference on Computer Vision, Copenhagen, Denmark*, vol. 3, 2002, pp. 255–271.

[94] E. Hayman, B. Caputo, M. Fritz, and J. Eklundh, "On the significance of real-world conditions for material classification," in *Proceedings of the European Conference on Computer Vision, Prague, Czech Republic*, vol. 4, 2004, pp. 253–266.

[95] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America,*, vol. 2, no. 3, pp. 1160–1169, 1985.

[96] A. Mojsilovic, M. V. Popovic, and D. M. Rackov, "On the selection of an optimal wavelet basis for texture characterization," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2043–2050, 2000.

[97] F. Ade, "Characterization of texture by eigen filters," *Signal Processing*, vol. 5, pp. 451–457, 1983.

[98] K. Messer, D. de Ridder, and J. Kittler, "Adaptive texture representation methods for automatic target recognition," in *Proceedings of the British Machine Vision Conference, Nottingham, UK*, vol. 2, 1999, pp. 443–452.

[99] T. Randen, "Filter and filter bank design for image texture recognition," Ph.D. dissertation, Norwegian University of Sciene and Technology., 1997.

[100] A. Li and Q. Zaidi, "Information limitations in perception of shape from texture," *Vision Research*, vol. 41, no. 22, pp. 2927–2942, 2001.

[101] S. C. Zhu, Y. N. Wu, and D. B. Mumford, "Filters, random-fields and maximum-entropy (frame): Towards a unified theory for texture modeling," *International Journal of Computer Vision*, vol. 27, no. 02, pp. 107–126, 1998.

[102] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localised spatial filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 55–73, 1990.

[103] D. Dunn and W. E. Higgins, "Optimal gabor filters for texture segmentation," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 947–964, 1995.

[104] T. P. Weldon and W. E. Higgins, "Integrated approach to texture segmentation using multiple gabor filters," 1996, pp. 955–958.

[105] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* John Wiley and Sons, second edition.

[106] A. Mahalanobis and H. Singh, "Application of correlation filters for texture recognition," *Applied Optics*, vol. 33, no. 11, pp. 2173–2179, 1994.

[107] A. K. Jain and K. Karu, "Learning texture discrimination masks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, pp. 195–205, 1996.

[108] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of IEEE Transaction*, vol. 86, no. 11, pp. 2278–2324, 1998.

[109] A. Efrosi and T. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the International Conference on Computer Vision, Corfu, Greece*, 1999, pp. 1039–1046.

[110] A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics, Los Angeles, California*, 2001, pp. 341–346.

[111] K. J. Danai and S. K. Nayar, "Histogram model for 3d textures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, California*, 1998, pp. 618–624.

[112] K. J. Dana and S. K. Nayar, "Correlation model for 3d texture," in *Proceedings of the International Conference on Computer Vision, Corfu, Greece*, vol. 2, 1999, pp. 1061–1067.

[113] M. J. Chantler, M. Schmidt, M. Petrou, and G. Mc-Gunnigle, "The effect of illuminant rotation on texture filters: Lissajous's ellipses.," in *Proceedings of the European Conference on Computer Vision, Copenhagen, Denmark*, vol. 3, 2002, pp. 289–303.

[114] A. Penirschke, M. J. Chantler, and M. Petrou, "Illuminant rotation invariant classification of 3d surface textures using lissajous's ellipses," in *Proceedings of the Second International Workshop on Texture Analysis and Synthesis, Copenhagen, Denmark*, vol. 2, 2002, pp. 103–108.

[115] J. J. Koenderink and S. C. Pont, "Irradiation direction from texture," *Journal of the Optical Society of America*, vol. 20, no. 10, pp. 1875–1882, 2003.

[116] X. D. He, K. E. Torrance, F. X. Sillion, and D. P. Greenberg, "A comprehensive physical model for light reflection," *Computer Graphics*, vol. 25, no. 4, pp. 175–186, 1991.

[117] M. Ashikhmin, S. Premoze, and P. Shirley, "A microfacet based brdf generator," in *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics, New Orleans, Louisiana*, 2000, pp. 65–74.

[118] D. P. Sidlauskas, "3d hand profile identification apparatus," *US Patent No. 4736203*, 1988.

[119] K.-A. Toh, W. Xiong, W.-Y. Yau, and X. Jiang, "Combining fingerprint and hand-geometry verification decisions," in *Proceedings of Audio- and Video-Based Biometric Person Authentication*, vol. 2688, Springer Berlin, 2003, pp. 1059–.

[120] A. Kumar and D. Zhang, "Personal recognition using hand shape and texture," *IEEE Transactions on Image Processing*, vol. 15, pp. 2454–2461, August 2006.

[121] J. Wu and Z. Qiu, "A hierarchical palmprint identification method using hand geometry and grayscale distribution features," in *Proceedings of International Conference on Pattern Recognition*, vol. 4, 2006, pp. 409–412.

[122] A. Kumar and D. Zhang, "Hand-geometry recognition using entropy-based discretization," *IEEE Transactions on Information Forensics and Security*, vol. 2, pp. 181–187, June 2007.

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

[124] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, pp. 195–202.

[125] D. Cofer and R. Hamza, "Method and apparatus for detecting objects using structured light patterns," *US Patent No. 7176440*, 2007.

[126] K. Faulkner, "Apparatus and method for biometric identification," *US Patent No. 5335288*, 1994.

[127] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Filterbank-based fingerprint matching," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 846–859, 2000.

[128] J. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1148–1161, 1993.

[129] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th International Conference on Computer Vision, Kerkyra, Greece*, 1999, pp. 1150–1157.

[130] K. J. Dana and S. K. Nayar. "Columbia-utrecht reflectance and texture database, http://www1.cs.columbia.edu/cave//software/curet/,",", 1999.

[131] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis*, vol. 62, pp. 61–81, April 2005.

[132] A. Kumar, "Imaging and depth estimation in an optimization framework," *MS Thesis*, November 2007.

[133] L. Vandevenne. "http://student.kuleuven.be/ m0216922/cg/fourier.html/#introduction,",".

[134] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson Education Inc., 2002.

[135] J.-K. Kmrinen, "Feature extraction using gabor filters," Ph.D. dissertation, Lappeenranta University of Technology, 2003.