# TITLE: IDENTIFICATION OF GASTROENTERIC VIRUSES BY ELECTRON MICROSCOPY USING HIGHER ORDER SPECTRAL FEATURES

Names of authors: Hannah Ong and Vinod Chandran

Postal address: Speech, Audio, Image and Video Technology Research Program
Queensland University of Technology, Brisbane, Qld 4001, AUSTRALIA

Telephone: (07) 3864 2124

Fax Number: 3864 1516

Email address: cl.ong@student.qut.edu.au, v.chandran@qut.edu.au

Name of corresponding author: Vinod Chandran

## ABSTRACT

Background: Many paediatric illnesses are caused by viral agents. For example, acute gastroenteritis. Electron microscopy can provide images of viral particles and can be used to identify the agents. Objectives: The use of electron microscopy as a diagnostic tool is limited by the need for high level of expertise in interpreting these images and the time required. A semi-automated method is proposed in this paper. Study design: The method is based on bispectal features that capture contour and texture information while providing robustness to shift, rotation, changes in size and noise. The magnification or true size of the viral particles need not be known precisely, but if available can be used additionally for improved classification. Viral particles from one or more images are segmented and analyzed to verify whether they belong to a particular class (such as Adenovirus, Rotavirus etc) or not. Two experiments were conducted - depending on the populations from which virus particle images were collected for training and testing, respectively. In the first, disjoint subsets from a pooled population of virus particles obtained from several images were used. In the second, separate populations from separate images were used. The performance of the method on viruses of similar size was separately evaluated using Astrovirus, HAV and Poliovirus. A Gaussian Mixture Model was used for the probability density of the features. A threshold on the log-likelihood is varied to study false alarm and false rejection trade-off. Features from many particles and/or likelihoods from independent tests are averaged to yield better performance. Results: An equal error rate (EER) of 2% is obtained for verification of Rotavirus (tested against 3 other viruses) when features from 15 viral particle images are averaged. It drops further to less than 0.2% when scores from 2 tests are averaged to make a decision. For verification of Astrovirus (tested against 2 others of the same size) the EER was less than 2% when 20 particles and 2 tests were used. Conclusion: Bispectral features and Gaussian mixture modelling of their probability density are shown to be effective in identifying viruses from electron microscope images. With the use of digital imaging in electron microscopes, this method can be fully automated.

Keywords: electron micrograph, gastroenteric viruses, higher-order spectra, bispectrum, invariant features, identification.

## 1.0 INTRODUCTION

Acute gastroenteritis is one of the most common diseases that affect humans. It continues to be a significant cause of morbidity and mortality worldwide (Glass et al., 2001). The mortality associated with gastroenteritis has been estimated to be 3-5 million cases per year. The majority of these occur in developing countries (Bern and Glass, 1994) and children under 5 years of age (Glass et al., 2001). Viruses, apart from bacteria and other parasites, have been known to be important causes of gastroenteritis. Four major categories of viruses are now recognized as clinically important to this problem. These are Rotavirus, Astrovirus, Adenovirus and Calicivirus (Kapikian, 1996; Glass et al., 2001). Many studies have been conducted to determine which gastroenteric viruses are more prevalent with respect to geography, sex, seasonal pattern and age distribution (Diamanti et al., 1996; Ueda et al., 1996; Haiping et al., 1999; Bereciartu et al., 2002; Subekti et al., 2002; Oh et al., 2003). Research shows that Rotavirus is responsible for the majority of these deaths and 20-52% of all acute gastroenteric episodes (Cunliffe et al., 2002a; Hart, 2003; Rivest et al., 2004). Accurate understanding of the relative prevalence of these agents would help design strategies to control the disease.

The diagnosis of viral gastroenteritis can be carried out by non-molecular techniques such as electron microscopy (EM), enzyme-immunoassay and latex agglutination tests, and by various molecular techniques. The advantage of using EM for direct visualization of virus particles in specimens is that it detects any potentially responsible viral agents present. By contrast, in immunologic tests, reagents may not currently exist that would permit complete immunologic testing (Noel et al., 1997; Green et al., 2002). Molecular genetic techniques also have similar limitations and they are only capable of identifying the presence of genomic material for previously identified agents.

Virus identification by human visual examination of EM images requires highly trained and experienced medical specialists. It is not suitable for screening large numbers of specimens. The virus verification method presented in this paper attempts to provide an important step in overcoming this issue by having a semi-automated verification process. The system can recognise a population of virus particles obtained from one or more EM images. This can make it possible to screen large numbers of images from various parts of the world, compare and check for mutations and confirm viral strains.

Virus cells appearing in EM images can vary in orientation, position and size. The images are also noisy owing to the high magnification and the low dose of electrons used in the microscope. It is desirable to have features that are invariant to orientation, position and size. EM images of different viruses exhibit fine differences in texture that arise from differences in their 3D surfaces and internal structures (as shown in Figures 1 and 2). Symmetries are commonly observed in biological reproduction processes. Symmetries are also evident in reconstructed 3D forms of viruses. Textures on viral particle images tend to exhibit different rotational symmetries as well. However, the variation in texture and any differences in symmetry are difficult to visualize in any single specimen image because of noise and poor resolution. They can be extracted from an ensemble of specimens provided the features are invariant to translation, size and rotation and also robust to noise. Such a set of features are higher order spectral invariants (Chandran and Elgar, 1993). These features capture texture variations as well as differences in contours.

## 2.0 HIGHER ORDER SPECTRAL FEATURES

Higher order spectra (Brillinger and Rosenblatt, 1967) are spectral representations of higher order moments or cumulants of a random process. They have been used for detecting deviation from Gaussianity and identifying non-linear systems (Nikias and Petropulu, 1993). Higher order spectra based on cumulants are zero for Gaussian processes. Higher order spectral theory has been extended to apply to deterministic signals and used for recognition of shapes. For deterministic signals they can be expressed as products of Fourier coefficients (Nikias and Raghuveer, 1987; Chandran and Elgar, 1993). For example the bispectrum $B(f_1, f_2)$ of a one-dimensional, deterministic, discrete-time signal, $x(n)$, is defined as

$$B(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2) \qquad (1)$$

where $X(f)$ is the discrete-time Fourier transform of $x(n)$ and $f$ is frequency normalized by one half of the sampling frequency.

The bispectrum is a function of two frequencies and in contrast to the power spectrum this function is complex-valued in general and thus retains some of the phase information in the Fourier transform. Especially for asymmetric sequences the phase is non-linear and higher order spectra retain the nonlinear phase information. They are also unaffected by a translation of the input. These unique properties of higher order spectra are useful in pattern recognition. If an input is even (or odd) symmetric, the phase of the Fourier transform is zero (or pi) or a linear function of frequency if the input is shifted. In either case, the phase of the bispectrum will be zero. This is expected because all the information resides in the Fourier magnitude for such inputs. The magnitude of the DFT of the input for positive frequencies may then be used to compute HOS invariants (now referred to as indirect HOS invariants). Because virus images can exhibit symmetry in projections, the indirect method is used in this work. An added advantage of using the indirect method is that the DFT magnitude sequence is band-limited and scale invariance is better satisfied for these indirect features.

### *2.1 1-D Bispectral Invariants*

Parameters, $P(a)$, that are translation, average-value, magnitude and scaling invariant are defined (Chandran and Elgar, 1993) from the bispectrum of $x(n)$ as follows:

$$P(a) = arctan(I_i(a)/I_r(a)) \tag{2}$$

where

$$I(a) = I_r(a) + jI_i(a) = \int_{f_1=0+}^{1/(1+a)} B(f_1, af_1)df_1 \tag{3}$$

for 0< $a \leq 1$, and $j$ = √-1. Note that the bispectral values are integrated along straight lines with slope *a* passing through the origin in the bifrequency space (as shown in Figure 3). Refer to (Chandran and Elgar, 1993) for the discrete-time version of $I(a)$.

In practice, the fast Fourier transform (FFT) is used to obtain $X(K)$ where $f = K/N$, $K = 0,1,\ldots,N/2-1$ and the integral in (3) is computed as a summation.

Invariance properties of the bispectral features have been proved in (Chandran and Elgar, 1993; Chandran et al., 1997). For the benefit of researchers in microbiology who may not be aware of this work on bispectrum features and pattern recognition, the proofs are reproduced in Appendix 1.

### *2.2 Radial Spectra of 1-D bispectral invariants*

The 1-D bispectral invariant features can be applied to 2-D images by taking Radon transform projections (Figure 4) and computing features from these projections (Chandran et al., 1997). Rotation invariance is achieved by taking radial spectra of these features.

Figures 4 and 5 show the flow chart of computation of these invariant parameters. The radial spectra of the bispectral invariants, $P(a,\omega_\theta)$ where $\omega_\theta$ is a frequency in cycles per 180 degrees is a set of features that are invariant to rotation, translation and scaling.

This method is illustrated using synthetic images of n-fold symmetries as shown in Figures 6, 7 and 8. Figures 6(c) and 6(d) show the Radon transform projections at 45 degree angle of images with a 5 fold symmetry and 7 fold symmetry binary object, respectively. Figure 7 shows the real and imaginary parts of the bispectrum of these projections. Figure 8 show the plots of the radial spectrum of the bispectral features, $P(1/2)$. Note that the 5 fold symmetry image produces a peak showing a dominant symmetry at 5 cycles per 180 degree whereas the 7 fold symmetry image produces a peak at 7 cycles per 180 degrees.

To illustrate how robustness can be achieved by averaging these features, the 7 fold symmetry image is used and white Gaussian noise is added to the image with SNR equal to 0dB. With only one image used, the plot of the radial spectrum of the bispectral features, as shown in Figure 9(a) does not reveal a dominant symmetry at 7 cycles per 180 degrees. In Figure 9(b), the individual spectra are accumulated over 75 of the 7 fold symmetry images with similar signal to noise ratio (SNR = 0dB).

As we take an ensemble of images, the spectrum (see top line in Figure 9(b)) eventually converges to a form revealing a visible peak at 7 cycles per 180 degrees. This is applicable to EM images where due to the low signal to noise ratio, individual viral particles are difficult to discern visually and present a challenge to feature extraction. Averaging of these features improves the classification performances.

Next, a 3-D reconstructed Adenovirus displayed on greyscale image in 2-D is compared with a 5 fold symmetry image. In Figure 10, plots of $P(a,\omega_\theta)$ are compared, where $a = 1/2$ and $\omega_\theta = 1, 2\ldots\ldots 16$ cycles per 180 degrees. $P(a,\omega_\theta)$ which is invariant to translation, rotation and scaling captures information from the contour and texture properties of the virus image that is useful for verification.

3-D reconstructions of virus particles are normally taken from a large set of electron cryomicroscopy (cryo-EM) images. Comparing with conventional EM, where the specimens are metal stained and dried for observation, in cryo-EM, the unstained particles are preserved in a flash-frozen aqueous environment. The drying process in conventional EM (example negative staining) tends to flatten the structure onto the support plane, causing distortions to the 3D structure. In other words, what is seen in the electron micrographs might not be a faithful representation of the virus. Due to this reason, the 3D reconstruction image is not used in this work to build the reference feature vectors for classification. A large set of negative stained EM images were used, which will produce a more accurate classification result when the testing set is of similar preparation method.

## 3.0 IMAGE ANALYSIS

Gastroenteric viruses are normally shed in high concentrations, often reaching particle concentrations of $10^{11}\text{ml}^{-1}$ which makes it suitable to diagnose these viruses using EM. The difficulty arises in distinguishing these viruses with one another since they are all nearly circular in shape with very little visual differences and produces the same pathological symptoms in the patients. This can be overcome by having a semi-automated verification system. We considered the problem of detecting Rotavirus against the other gastroenteric viruses such as Calicivirus, Adenovirus and Astrovirus.

Rotavirus was chosen as the target virus due to prevalence of this virus in causing acute gastroenteritis. Although size alone can distinguish these viruses, this study is conducted under the assumption that the magnification level and true particle size is unknown. Classification of these viruses are based upon contour and texture. Another experiment is conducted with viruses of similar size. The viruses chosen are Astrovirus, Hepatitis A virus (HAV) and Poliovirus.

The following steps comprise the image analysis:

**Segmentation**: Segmentation is performed to separate individual viral particles from the image. Viral particles in the images are segmented out into subimages (64 by 64 pixel). The images are aligned using the centroid of each subimage. This alignment need not be perfect because the features extracted are translation invariant. Then a circular mask is applied to each particle to eliminate the peripheral region that may contain neighbouring virus cells. A circular mask will not corrupt the type of periodicity of the bispectral features as exhibited by the virus image within it. The features extract asymmetry information from the virus image and a perfectly circular mask will not introduce any asymmetry as a result of masking. Examples of segmented image of each type of virus are shown in Figure 2.

**Extraction of cell features:** Each subimage (Figure 2) is subjected to the steps shown in Figure 4 and 5. Radon transform projections at 32 angles are computed from each subimage to yields one-dimensional functions. 10 bispectral features, $P(a,\theta)$ (where $a = 1/10, 2/10, \ldots\ldots, 1$ and $\theta(rad) = \pi/32, 2\pi/32, \ldots\ldots, \pi)$ are computed from each such projection. A total of 320 features can thus be extracted from each subimage. These features are not rotation invariant. A discrete Fourier transform (DFT) is then computed on the features considered as sequences with the angle of rotation as the index, to yield a new set of features, $P(a,\omega_\theta)$, where $\omega_\theta$ is a frequency in cycles per 180 degrees. The resulting features are invariant to rotation, translation and scaling and are therefore robust to small changes in the sizes of viral particles and their orientation. They are sensitive to asymmetries in the shape of the virus, and because of their robustness to orientation and size they can be combined for a population of virus particles, to pick up useful shape differences that are not visually evident from electron micrograph images of single particles.

**Feature Selection**: In general, the dimension of the feature vector (i.e., the number of features) may be very large at the feature generation stage. Given a number of available features, the main task of feature selection (or reduction) process is to select information rich features providing a large interclass distance and a small intraclass variance in the feature vector space. In this work, the dimensionality is reduced by computing the largest distance separation, $F$ between the target virus and the background virus.

$$F = (\mu_1 - \mu_2)^2 / (\sigma_1^2 + \sigma_2^2) \tag{4}$$

where $\mu_1$ and $\mu_2$ are the mean of the target virus and the background virus and $\sigma_1$ and $\sigma_2$ are the standard deviation of the target virus and background virus computed over some subset of the training set. Using the largest distance separation, the 320 features were reduced to 10 features that were used for verification. These 10 features are the ones that produced the largest distance separation between the target and background virus.

## 4.0 GMM MODELLING

The virus particles were trained using Gaussian Mixture Model (GMM) (Bilmes, 1998). GMM was chosen to model the target virus and background virus because the various modes or clusters that exist due to images of different scale, background and contrast that was used in training in each virus type. Figure 11 shows the cluster plot of 2 randomly chosen bispectral features of three different set of images of Rotavirus. They differ in scale, background and contrast. Each feature is an average from a subpopulation of 10 viral particles. As can be seen, the plot which represents probability distribution of the features shows quite compact and isolated clusters in feature space.

Each type of virus is represented by a GMM describing its features, with its mean vectors, covariance matrix and the mixture weights as parameters. A diagonal covariance form of the GMM was used with a covariance matrix for each component. The number of mixture components was determined by computing the mixture components (from 1,2…10) and the one which produces the smallest equal error rate (EER) was chosen. The k-means algorithm was used to pre-processes the input data which performs an unsupervised learning in order to find centres of clusters which reflect the distribution of the data, then followed by iterations of the Expectation Maximization (EM) algorithm (Bilmes, 1998). The iteration was stopped when the change in log likelihood of the error function at the solution between two steps of the EM algorithm of the target and background virus was below a preset threshold and considered insignificant.

## 5.0 GMM VERIFICATION

The test set is scored against both the target model (Rotavirus) and the background model (Calicivirus, Astrovirus and Adenovirus). In the experiment of viruses of similar size, the target model is Astrovirus and the background model are HAV and Poliovirus. In the first verification test, the decision score $S( X ,\lambda_N )$ is a log likelihood ratio of a subpopulation of $N$ particles, where $N$ is the number of particles used to compute a feature vector by averaging.

$$T = S( X ,\lambda_N ) = log\ f'( X_i | \lambda_N ) - log\ f'( X_i | \beta_N ) \tag{5}$$

In the second verification test, each decision score, $S'( X ,\lambda_N )$ is an average of $M$ sets where each set consist of a subpopulation of $N$ particles, as can be shown in the formula below;

$$T = S'( X ,\lambda_N ) = \sum_{i=1}^{M}\ log\ f'( X_i | \lambda_N ) - log\ f'( X_i | \beta_N ) \tag{6}$$

Where $\lambda_N$ is the target model, $X_i$ is the test set which comprises of a subpopulation of $N$ particles, $f'( X_i | \lambda_N )$ is the likelihood of the target virus and $f'( X_i | \beta_N )$ is the average likelihood of the three background virus, each of the background virus is weighted equally.

$$\text{If } T \begin{cases} > 0\text{, then the virus is identified as the target virus} \\ < 0\text{, then the virus is identified as non-target virus} \end{cases}$$

## 6.0 EXPERIMENTS

Two types of experiments were conducted. In the first experiment, the training and testing were carried out on a population of virus images pooled from all the EM images of that type obtained from various sources (referred to as a pooled population). A pooled population is useful when a single image does not provide enough viral particle subimages for statistically reliable training or testing. Although the population is pooled, viral particle subimages used for training are different from those used for testing.

In the second experiment, the testing and training is done on populations derived from separate images. The images used in the training set were different from those in the test set. All viral particles in a given population are selected from the same image (referred to as a single image population). The feature vector in each case is obtained by averaging features from a set of $N$ number of particles as shown in Figure 9. The first verification test was performed on both the single image population and the pooled population, while the second verification test was only performed on the pooled population. Figure 12 illustrates the selection of viral particles from EM images used for testing and training in the pooled population and the single image population cases.

### *6.1 Experiment on a pooled population*

Different sets of digitised electron microscope images obtained from various sources (see acknowledgement) of these gastroenteric viruses were used for testing and training. 10 images of Rotavirus of different scale, background, contrast and appearance of contours were chosen. A total of 12 images of Adenovirus, Astrovirus and Calicivirus were used as the background virus. For each virus type, the training set consisted of 75 particles. The remaining particles were used to select the test set. 20 particles of each of the three background viruses were chosen randomly to form 60 particles for the test set of the background virus.

A subpopulation of $N$ particles was chosen from the test sets (a pooled population) of the target virus and the background virus. In the first verification test of the pooled population, each decision score is produced by a subpopulation of $N$ particles, where $N$ is the number of particles used to compute a feature vector by averaging (referred to as a subpopulation test). In the second verification test, likelihood scores of $M$ sets where each set consists of a subpopulation of $N$ particles, were averaged to produce a decision score (referred to as the averaged score subpopulation test).

The verification tests were performed on subpopulations of $N$ viral particles where $N = 5, 8, 10, 13$ *and* $15$. For each subpopulation size, 100 tests (scores) with randomly selected subpopulations from the test set were conducted and the results are presented in Detection Error Tradeoff (DET) curves. In the second verification test, $M$ was set to $2$.

Results of the subpopulation test are presented in Figures 13 and 14. From Figure 13, it can be seen that as we increase the test ensemble for feature averaging from 5 particles to 15 particles, the equal error rate (EER) drops quite significantly. Figure 14 shows clearly that the EER drops from 15% to 2.5%. The test was stopped at 15 particles because there was no improvement in the EER as we increase the subpopulation size for feature averaging, from 13 particles to 15 particles.

The EER drops even lower when the number of sets of subpopulations used for comparing probability densities modelled by Gaussian Mixture Models (GMMs) increases. Results of an averaged score subpopulation test with $M=2$ and $N=15$ particles is presented in Figure 17. 5000 tests (scores) with randomly selected subpopulations from the test set were conducted. From the DET curve, the EER drops from 2% to less than 0.2% as $M$ goes from 1 to 2.

If two test populations are thrown at the GMM and log-likelihood scores are obtained, they could be used in different ways. In output fusion, decisions are combined. Each score could be used to obtain a decision and the decisions may be combined. For example, a decision to accept may be made only if both individual decisions are to accept. In this case, false acceptance rate will be the product of individual false acceptance rates; however, the false rejection rate will be the sum of the individual ones. The false acceptance will go from 2% to 0.04% but at the expense of false rejection which goes from 2% to 4% at the same threshold (assuming we had an equal error rate of 2% at that threshold).

Alternately, a decision to reject may be made only if both individual decisions are to reject. In this case, the false rejection rate will go to 0.04% but false acceptance to 4% at that threshold. This is shown in Figure 17.

By contrast in input fusion, features or scores are combined. Individual scores may be weighted and combined depending upon the confidence one has in each score. There will exist some optimal weighting of the scores for which the performance is best. When there is no prior knowledge of confidence or the two tests are equal in all respects, the scores may simply be averaged (50% weight for each of the two scores). This is done here. It turns out that averaged scores yield better performance than output fusion in this case.

### *6.2 Experiment on single image population*

In the second experiment, only images that have more than 20 viral particles per image were chosen since the testing and training were conducted on subpopulations of particles that are drawn from the same image. 9 images of Rotavirus of different scale, background, contrast and appearance of contours were chosen and 5 of these were used for training and the rest for testing. For the background virus, a total of 11 images were used; 5 images of Adenovirus, 4 images of Astrovirus and 3 images of Calicivirus. Out of these images, 3 Adenovirus images were used for training and two each of Astrovirus and Calicivirus; and the rest were used for testing.

The verification test was performed on subpopulations of viral particles of sizes *N = 5, 8, 10, 13, 15 and 18.* 100 tests (scores) were performed for each subpopulation size. Results of the subpopulation test are presented in Figures 15 and 16. The figures show that as we increase the test ensemble for feature averaging from 5 particles to 18 particles, the equal error rate (EER) drops from 20% to 2%. The test was stopped at 18 particles (Figure 16) because there was no improvement in EER as we increased the subpopulation size from 15 to 18 particles.

### *7.0 Experiment on viruses of similar true size*

Two types of viruses with similar size; Astrovirus of diameter 28-30nm and viruses of Parvoviridae family, Hepatitis A virus (HAV) and Poliovirus of diameter 22-30nm were chosen. This experiment was conducted to determine the ability of this method to distinguish virus particle of similar true size. 6 images of Astrovirus and 7 images of HAV and Poliovirus were used for training and testing. Figure 18 shows the electron micrograph of Astrovirus and HAV of the same magnification level. Astrovirus was used as the target virus and HAV and Poliovirus were used as the background virus.

The verification test of a *pooled population* was performed on subpopulations of viral particles, where *N= 5, 10, 15, 20*. Figure 19 presented the results in subpopulations of 500 tests. Figure shows that as we increased the subpopulations to 20 viral particles, the EER drops to 5%. The EER drops further as we increased *M from 1 to 2* of a subpopulation of 20 particles to less than 2%.

## 8.0 CONCLUSION

The paper presents a new semi automated identification method for viruses from digitised electron micrograph images based on higher-order spectral features that are invariant to rotation, scaling and translation. The system can be made fully automated by automatically segmenting the individual virus particles. Verification tests have been conducted on 4 major types of viruses that causes gastroenteritis; Adenovirus, Astrovirus, Rotavirus and Calicivirus, where Rotavirus was chosen as the target virus and the rest as the background virus. Results are presented for tests with various subpopulation sizes, *N*, used for averaging feature values and varying number of subpopulations, *M*, used for averaging likelihoods. EER of around 2% is achieved for *N=15*, *M=1*. EER drops to less than 0.2% for *M=2*. Tests were also conducted on viruses of similar true size. Astrovirus was scored against HAV and Poliovirus. Results shows that the EER drops to less than 2% for *N=20, M=2*. This work could form the basis of a reliable automated virus identification system that can be used as a research or diagnostic tool.

## ACKNOWLEDGEMENT

The authors are grateful to the Centre of Disease Control, USA and The Universal Virus Database of the International Committee on Taxonomy of Viruses for their public domain virus images and also to Prof Hans Gelderblom for providing some of the electron micrograph.

## REFERENCES

Bereciartu A, Bok K, Gomez J. Identification of viral agents causing gastroenteritis among children in Buenos Aires, Argentina. Journal of Clinical Virology 2002; 25(2), 197-203.

Bern C, Glass RI. Impact of diarrhoeal disease worldwide. Marcel Dekker, New York, 1994;1-26.

Bilmes J. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models. International Computer Science Institute 1998.

Brillinger D, Rosenblatt M. Computation and Interpretation of k-th Order Spectra. Wiley, 1967;907-938

Chandran V, Carswell B, Boashash B, Elgar S. Pattern recognition using invariants defined from higher order spectra: 2-D image inputs. IEEE Transactions on Image Processing 1997; 6(5), 703-711.

Chandran V, Elgar SL. Pattern recognition using invariants defined from higher order spectra one-dimensional inputs. IEEE Transactions on Signal Processing 1993; 41(1), 205-212.

Cunliffe N, Bresee J, Gentsch J, Glass R, Hart C. The expanding diversity of rotaviruses. Lancet 2002a; 359, 640-642.

Diamanti E, Superti F, Tinari A, Marziano ML, Giovannangeli S, Tafaj F, Xhelili L, Gani D, Donelli G. An epidemiological study on viral infantile diarrhoea in Tirana. New Microbiol. 1996; 19(1), 9-14.

Glass RI, Bresee J, Jiang B, Gentsch G, Ando T, Fankhauser R, Noel J, Parashar U, Rosen B, Monroe S. Gastroenteritis viruses: an overview. Novartis Found Symp 2001; 238, 5-19.

Green KY, Belliot G, Taylor JL, Valdesuso J, Lew JF, Kapikian AZ, Lin FY. A predominant role for Norwalk-like viruses as agents of epidemic gastroenteritis in Maryland nursing homes for the elderly. Journal of Infect Dis. 2002; 185(2), 133-146.

Haiping O, Nilsson M, Abreu ER, Hedlund KO, Johansen K, Zaori G, Svensson L. Viral diarrhea in children in Beijing, China. Journal of Medical Virology 1999; 57(4), 390-396.

Hart C. Rotavirus:Antigenic variation. Academic Press, London, 2003;84-101.

Kapikian AZ. Overview of viral gastroenteritis. Arch Virol Suppl. 1996; 12, 7-19.

Nikias CL, Petropulu AP. Higher-order spectra analysis: a nonlinear signal processing framework. PTR Prentice Hall, Englewood Cliffs, N.J., 1993; 537.

Nikias CL, Raghuveer M. Bispectrum estimation: A digital signal processing framework. Proc.IEEE 1987; 75, 869-889.

Noel JS, Ando T, Leite JP, Green KY, Dingle KE, Estes MK, Seto Y, Monroe SS, Glass RI. Correlation of patient immune responses with genetically characterized small round-structured viruses involved in outbreaks of nonbacterial acute gastroenterities in the United States, 1990 to 1995. Journal of Medical Virology 1997; 53(4), 372-383.

Oh DY, Gaedicke G, Schreier E. Viral agents of acute gastroenteritis in German children:prevalence and molecular diversity. J Med Virol 2003; 71(1),82-93.

Rivest P, Proulx M, Lonergan G, Lebel MH, Bedard L. Hospitalisations for gastroenteritis: the role of rotavirus*1. Vaccine 2004; 22(15-16), 2013-2017.

Subekti D, Lesmana M, Tjaniadi P, Safari N, Frazier E, Simanjuntak C, Komalarini S, Taslim J, Campbell JR, Oyofo BA. Incidence of Norwalk-like viruses, rotavirus and adenovirus infection in patients with acute gastroenteritis in Jakarta, Indonesia. FEMS Immunology and Medical Microbiology 2002; 33(1), 27-33.

Ueda Y, Nakaya S, Takagi M, Ushijima H. Diagnosis and clinical manifestations of diarrheal virus infections in Maizuru area from 1991 to 1994- especially focused on small round structured viruses. Kansenshogaku Zasshi 1996; 70(10), 1092-1097.

## APPENDIX 1. Proofs of invariance properties.

Claim: *P(a)* are translation invariant

Proof: Translation produces linear phase shifts of sequence *x(n)* that cancel in (1). Integrating the bispectrum along lines passing through the origin in bifrequency space preserves the translation because the integral is translation invariant if the integrand is. The phase of this complex entity *I(a)* must also be translation invariant because its real and imaginary parts are. Thus, *P(a)* are translation invariant.

Claim: *P(a)* are scale invariant

Proof: Scaling the sequence *x(n)* results in an expansion or contraction of the Fourier transform that is identical along the $f_1$ and $f_2$ directions. The real and imaginary parts of the integrated bispectrum along a radial line are multiplied by identical real-valued constants upon scaling and therefore the phase, *P(a)* of the integrated bispectrum is unchanged.

Rotation invariance is achieved by deriving invariants from the Radon transform of the image and using the cyclic-shift invariance property of the discrete Fourier transform magnitude (Chandran and Elgar, 1993; Chandran et al., 1997) (refer to Figure 4).

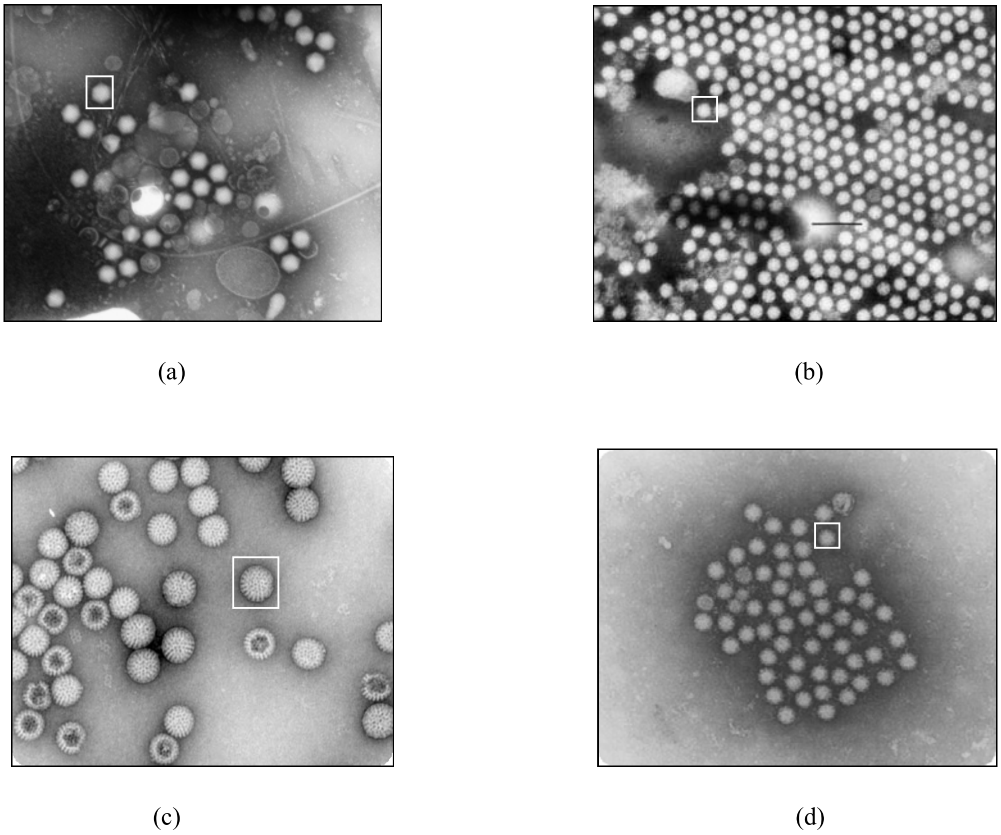


Figure 1: A sample image of each type of virus used for testing. These images are different in magnification and resolution. (a) Adenovirus (b) Astrovirus (c) Rotavirus (d) Calicivirus

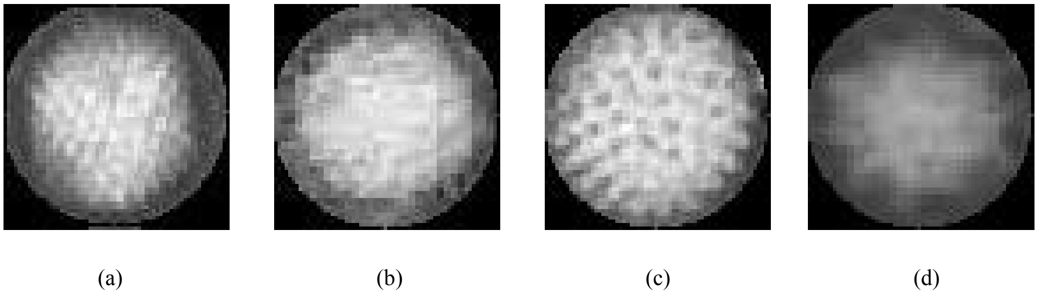


Figure 2: A single virus of each type. (a) Adenovirus (b) Astrovirus (c) Rotavirus (d) Calicivirus. These subimages are extracted from portions shown by square boxes in figure 1 and a circular mask is applied to each. Note that although there are small differences in texture, it is difficult to tell them apart by visual examination. Pseudo colouring could be used to emphasize the differences in texture but the difficulty arises when there is some variation in texture within the same virus type, on images that are obtained from various sources of different background, scale, contrast and noise.

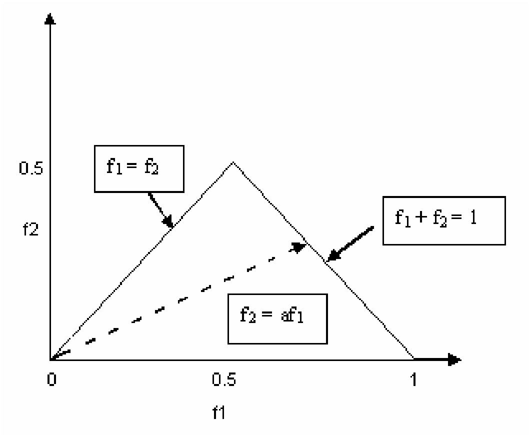


Figure 3: Owing to symmetries of the bispectrum in equation (1), the bispectrum possess redundancy and need only be computed for the triangular region shown above. Features are extracted by integrating the bispectrum along a radial line as shown and taking the phase of the complex-valued integral. $f_1$ and $f_2$ are frequencies normalized by one half of the sampling frequency.

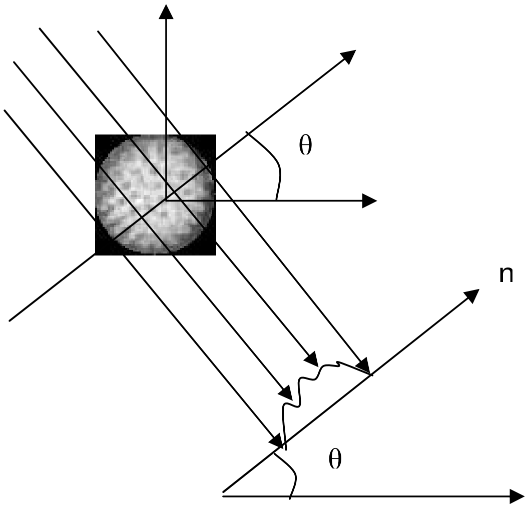


Figure 4: The Radon transform of the virus image yields 1-D parallel beam projections, $x(n)$ at various angles, $\theta$.

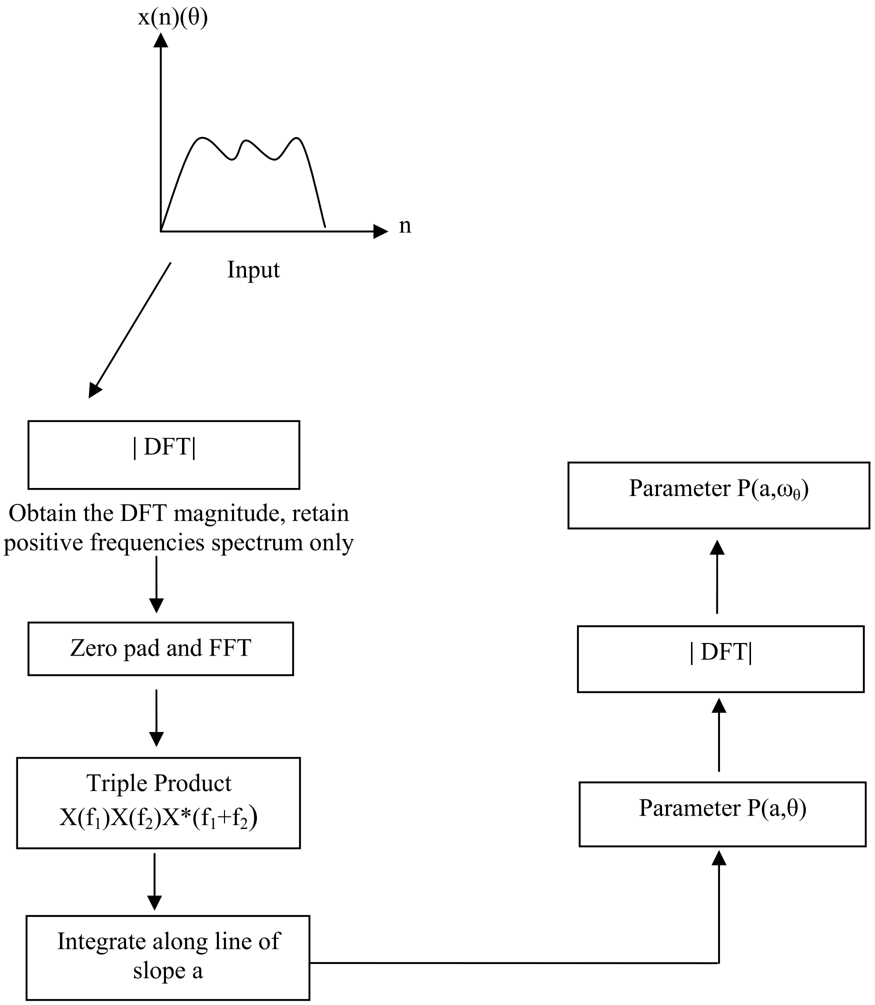
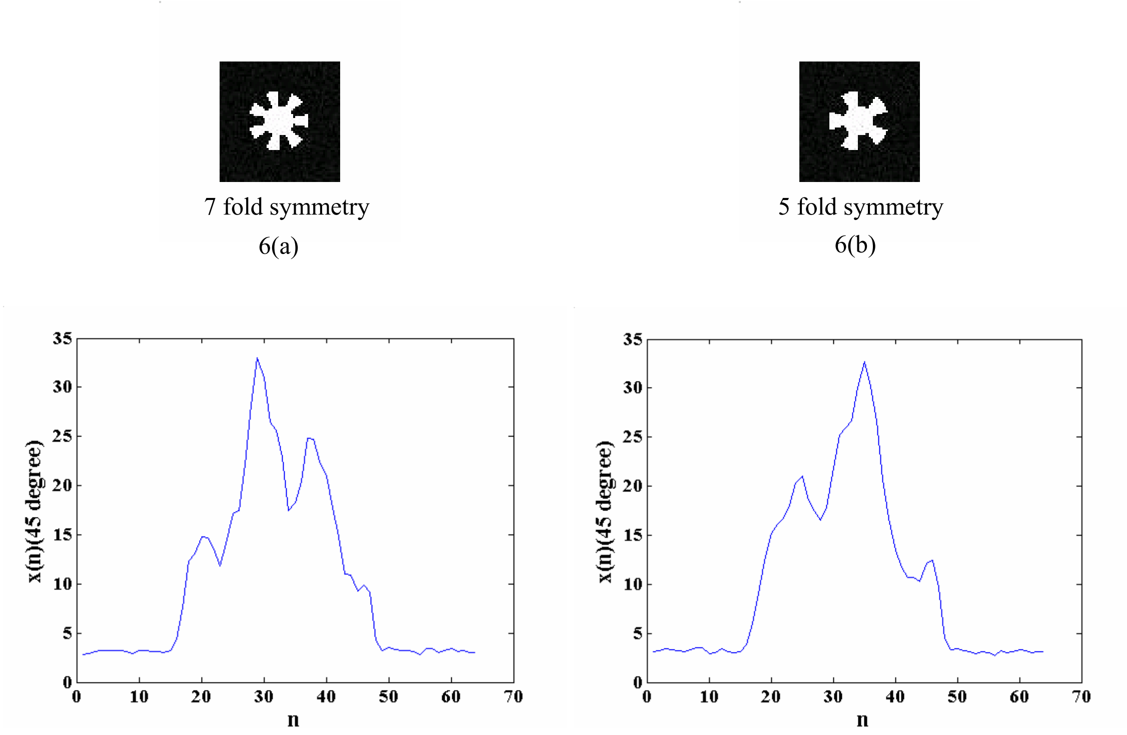


Figure 5: Flow chart of computation of invariant parameters. $P(a,\theta)$ is invariant to scaling and translation and $P(a,\omega_\theta)$ is invariant to scaling, translation and rotation. The algorithm was tested on a 5 fold symmetry and a 7 fold symmetry image and results are presented in Figures 6, 7 and 8.



7 fold symmetry

6(a)

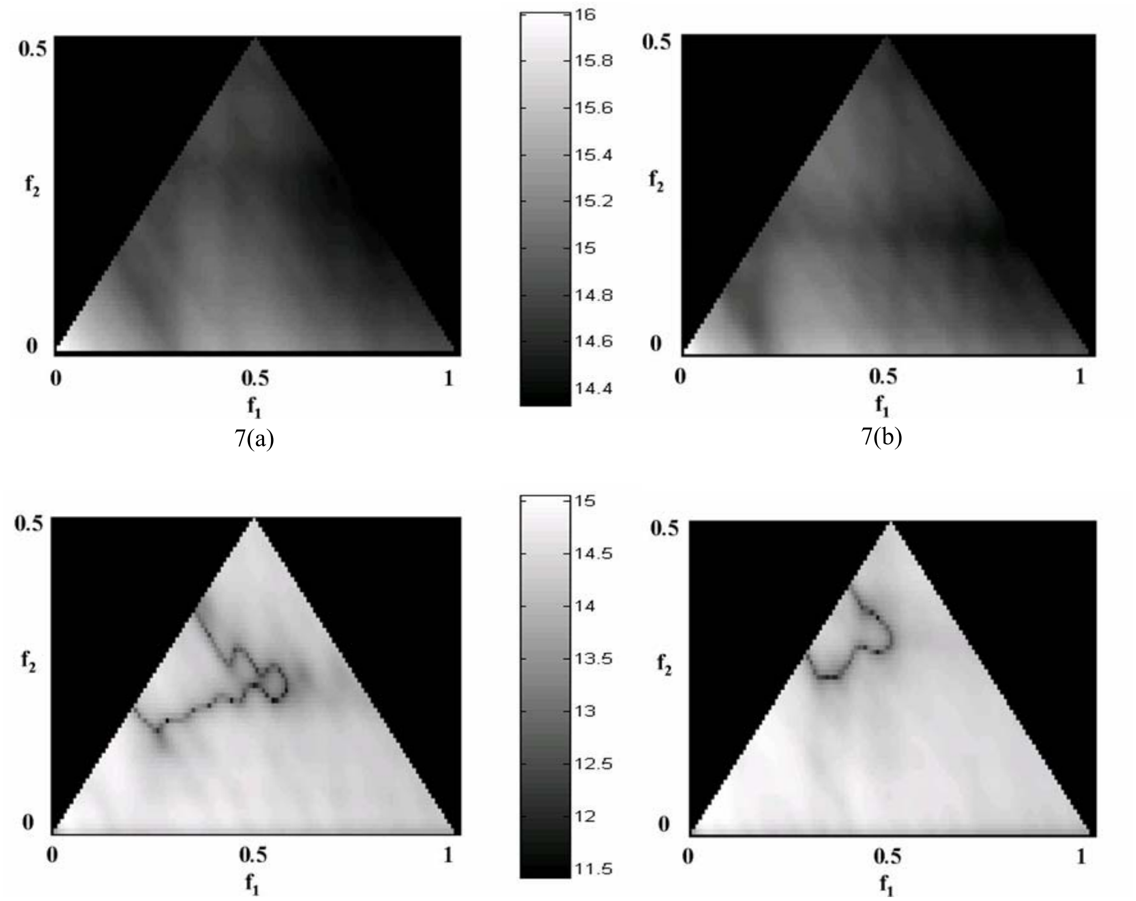
5 fold symmetry

6(b)

6(c)

6(d)

Figure 6: Figure 6(c) and 6(d) show the Radon transform projection at 45 degree angle of the 7 fold symmetry image, Figure 6(a) and the 5 fold symmetry image, Figure 6(b).



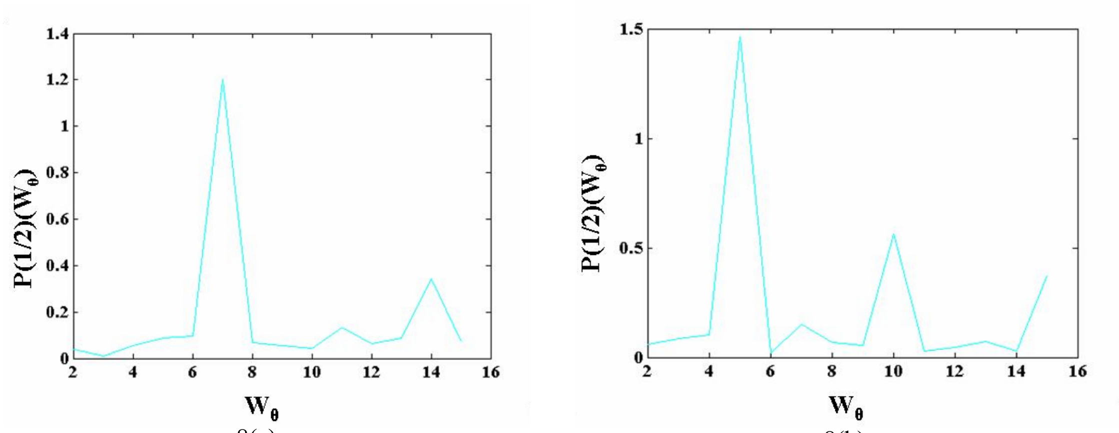
7(a)

7(b)

7(c)

7(d)

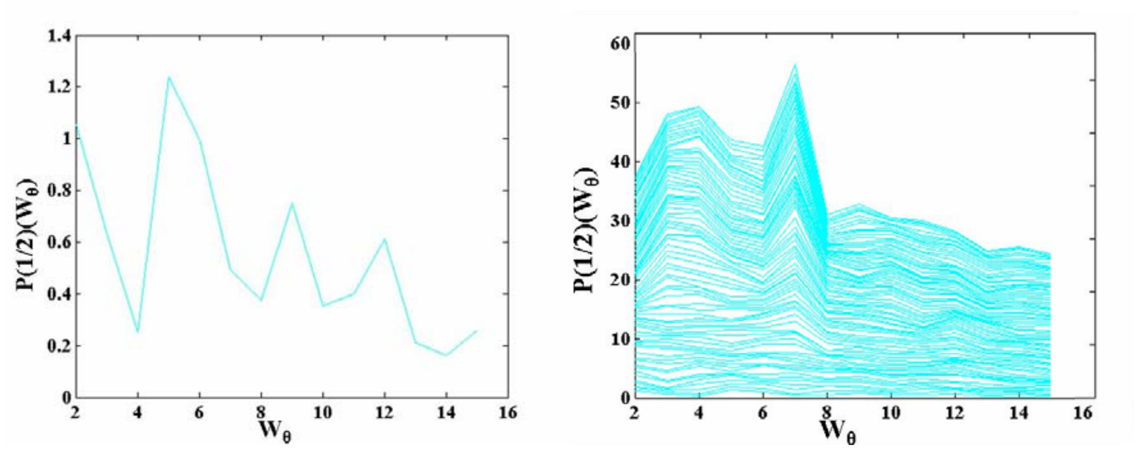
Figure 7: Figure 7(a) and Figure 7(c) show the real and imaginary parts of the bispectrum of the Radon transform projection at 45 degree angle of Figure 6(a). Figure 7(b) and Figure 7(d) show the real and imaginary parts of the bispectrum of Figure 6(b). The bispectrum is a triple product of Fourier coefficients and is a complex valued function of two frequencies, $f_1$ and $f_2$, where $f_1$ and $f_2$ are frequencies normalized by one half of the sampling frequency. Different shaped projections result in different bispectra. Invariant features are extracted by integrating along radial lines and taking the phase. The scale shown at the colorbar above is the log of the absolute value of the real and imaginary parts of the bispectrum. The above plots show that features, $P(a)$ close to $a = 1/2$ (the 45º line) may capture differences as well.



8(a)

8(b)

Figure 8: Figure 8(a) and 8(b) show plot of $P(1/2)(\omega_\theta)$ as a function of $\omega_\theta$, (where $\omega_\theta$ is a frequency in cycles per 180 degrees) of Figure 6(a) and Figure 6(b). $P(1/2)(\omega_\theta)$ is invariant to scaling, translation and rotation. Note that Figure 8(a) shows a dominant symmetry at 7 cycles per 180 degrees whereas Figure 8(b) shows a dominant symmetry at 5 cycles per 180 degrees.



9(a) 9(b)

Figure 9: Figure 9(a) shows the plot of the radial spectrum of the bispectral features, $P(1/2)(\omega_\theta)$ as a function of $\omega_\theta$, (where $\omega_\theta$ is a frequency in cycles per 180 degrees) of a 7 fold symmetry. White Gaussian noise has been added and SNR = 0dB to the image. In Figure 9(b), the individual spectra are accumulated over 75 such images. Note that Figure 9(a) does not demonstrate a dominant symmetry at 7 cycles per 180 degrees due to the low signal to noise ratio. As an ensemble of images is taken and the spectra are accumulated, it eventually converges to a shape. A peak at 7 cycles per 180 degrees can be seen in Figure 9(b). Robustness to noise can thus be achieved by averaging these features.

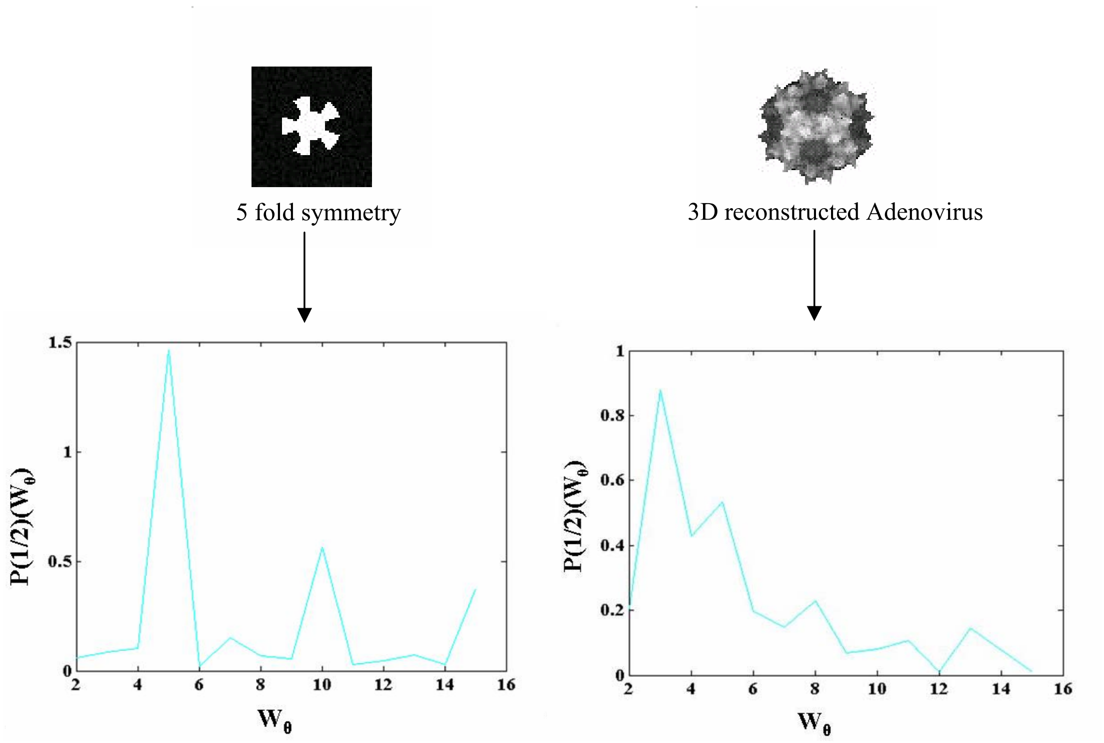


Figure 10: Comparison between a 5 fold symmetry image and a 3D reconstructed virus image using plots of radial spectra of the bispectral features, *P(1/2).* These features which are invariant to translation, rotation and scaling contain information from the contour and texture that are useful for verification.

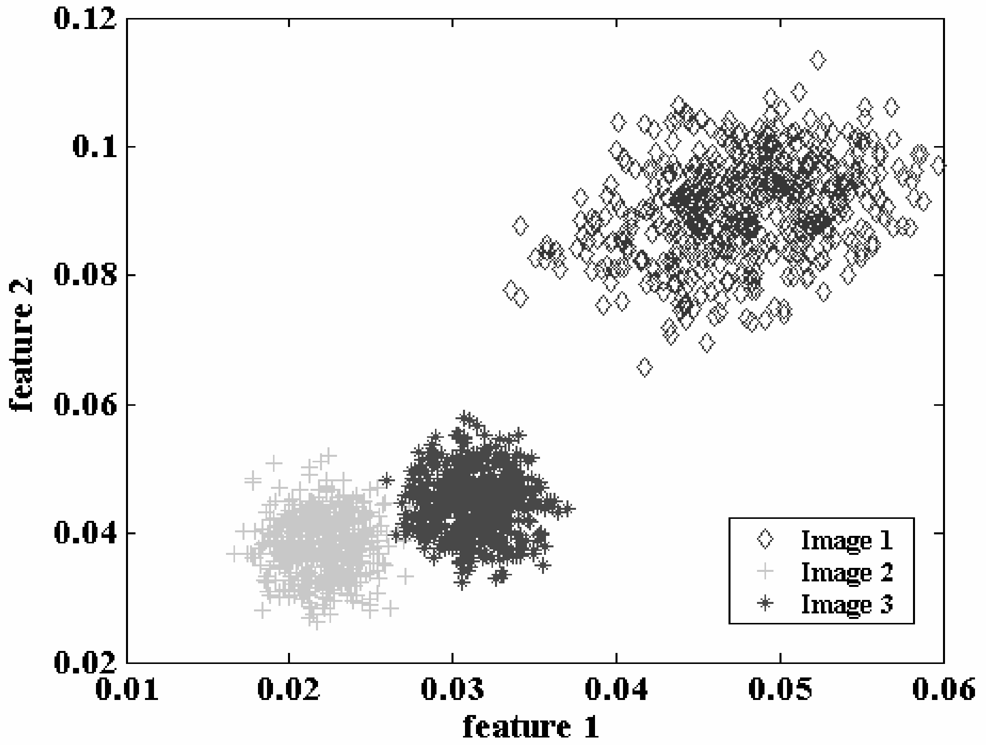


Figure 11: Cluster plot of features from three different sets of Rotavirus images with different backgrounds, contrast and scale. Each point is an average feature from a subpopulation of 10 viral particles. The plot shows quite compact and isolated clusters or modes in feature space.

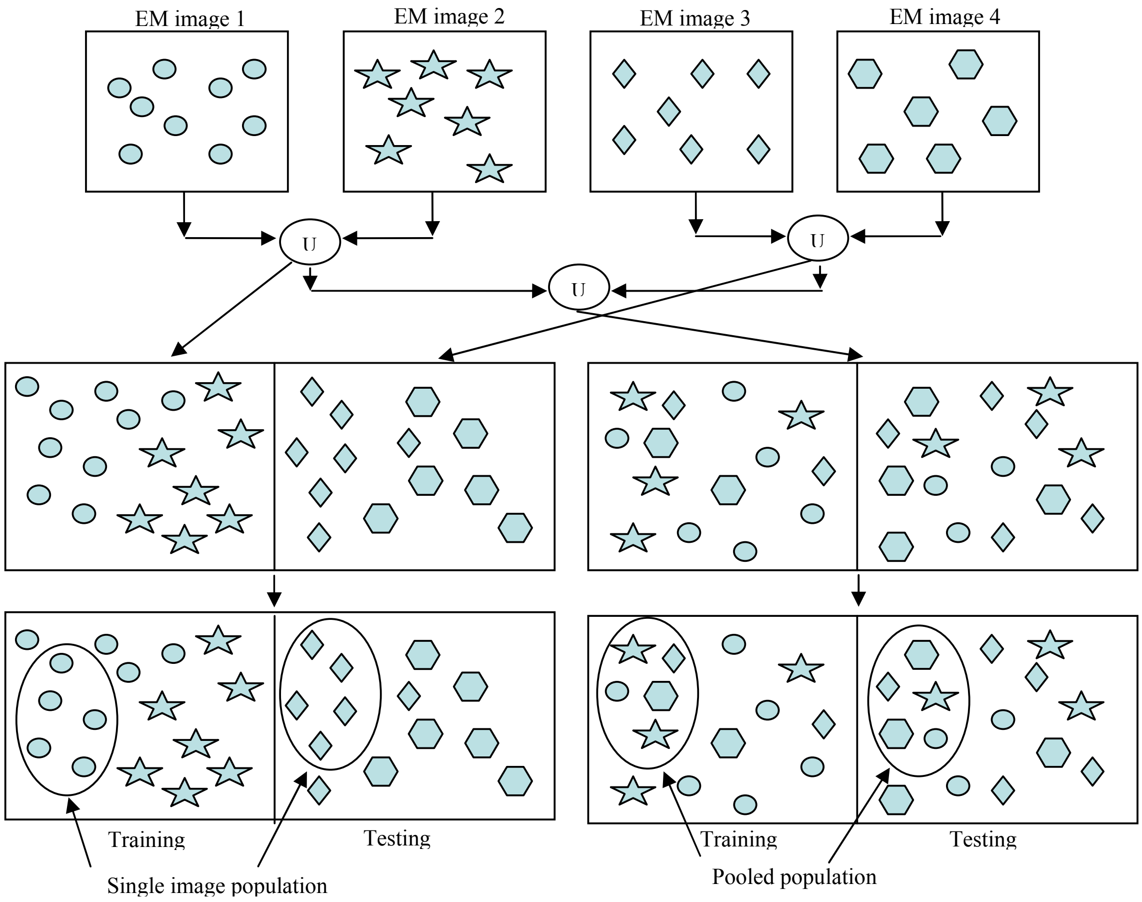


Figure 12: Illustration of selection of viral particles from different sets of EM images used for testing and training in pooled population and single image population. In the single image population, the testing and training is done on populations derived from separate images. In a pooled population, the virus images pooled from all the EM images of that type obtained from various sources.

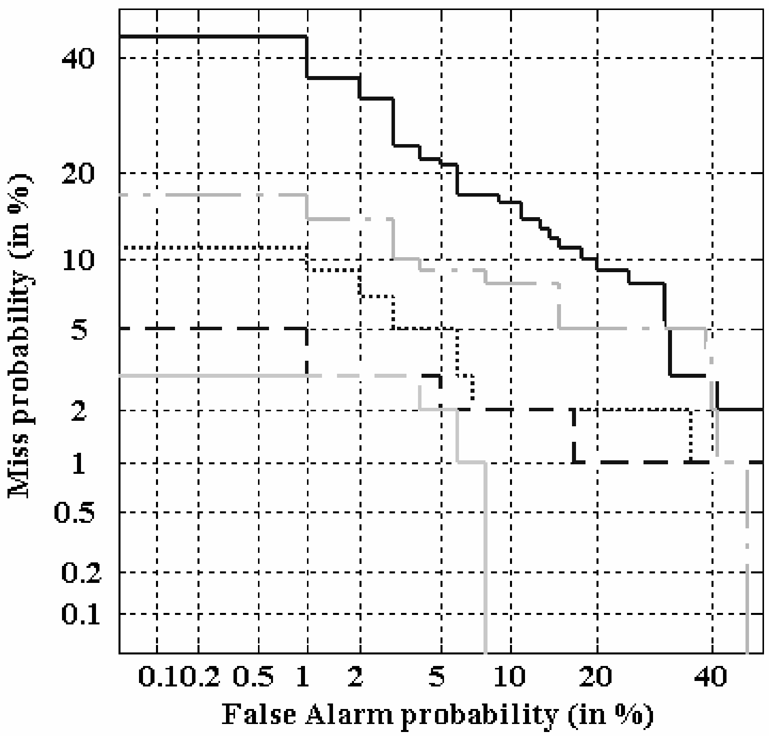


Figure 13: DET curve using the bispectral features from subpopulations of 5, 8, 10, 13 and 15 viral particles on a *pooled population*. As we increase the test ensemble size for feature averaging, the EER drops. EER is the point on the DET curve where the false alarm probability is equal to the miss probability. Refer to Figure 14 for EER values of each subpopulation size.

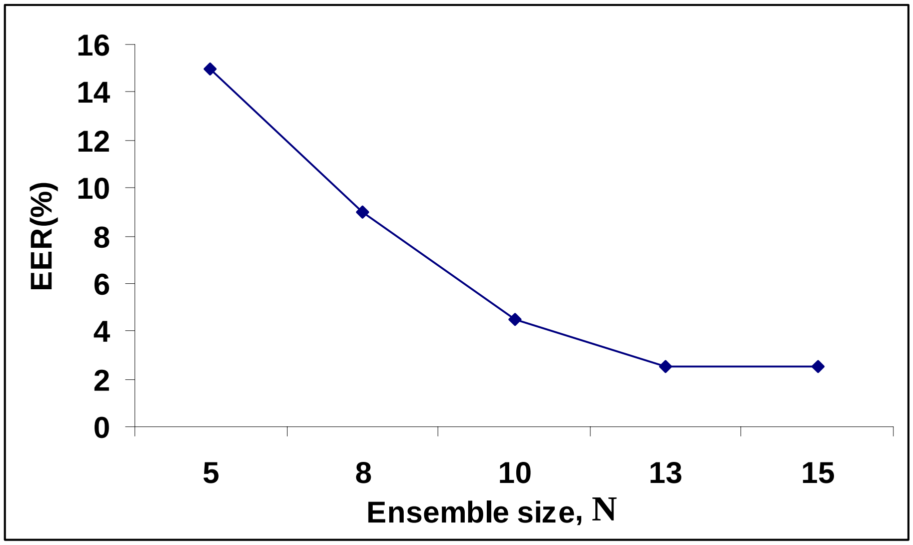


Figure 14: Plot of EER versus ensemble size shows that as we increase the test ensemble for feature averaging, the EER drops for N= 15 to 2.5%.

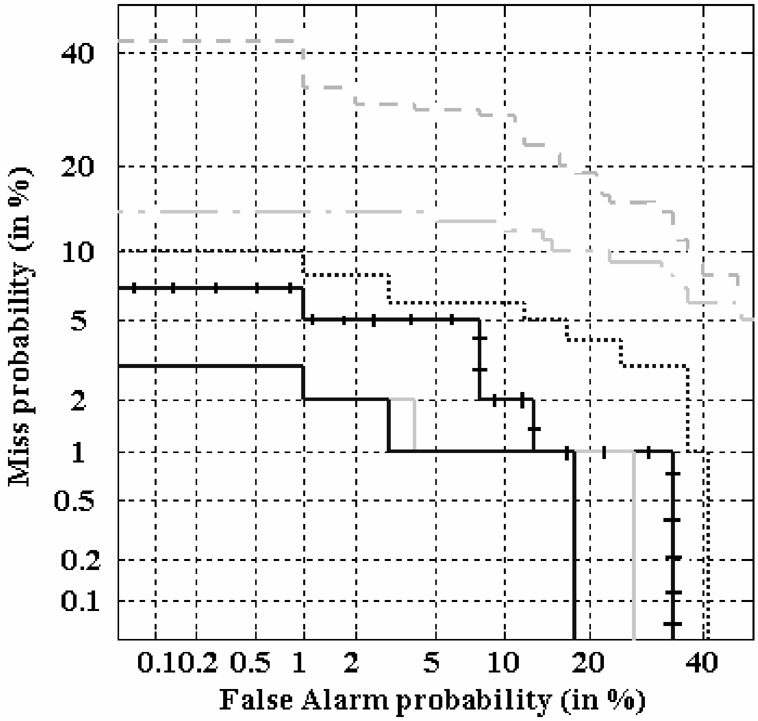


Figure 15: DET curve using the bispectral features from subpopulations of 5, 8, 10, 13, 15 and 18 viral particles on a single image population. The EER drops as we increase the feature averaging of the subpopulation size. Refer to Figure 16 for EER values of each subpopulation size.

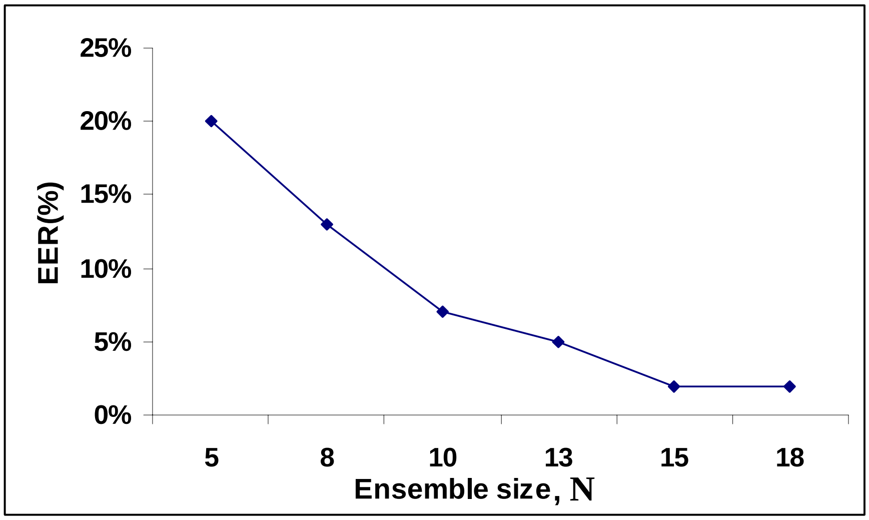


Figure 16: Plot of EER versus ensemble size shows that as we increase the test ensemble for feature averaging, the EER drops for 18 particles to 2%.

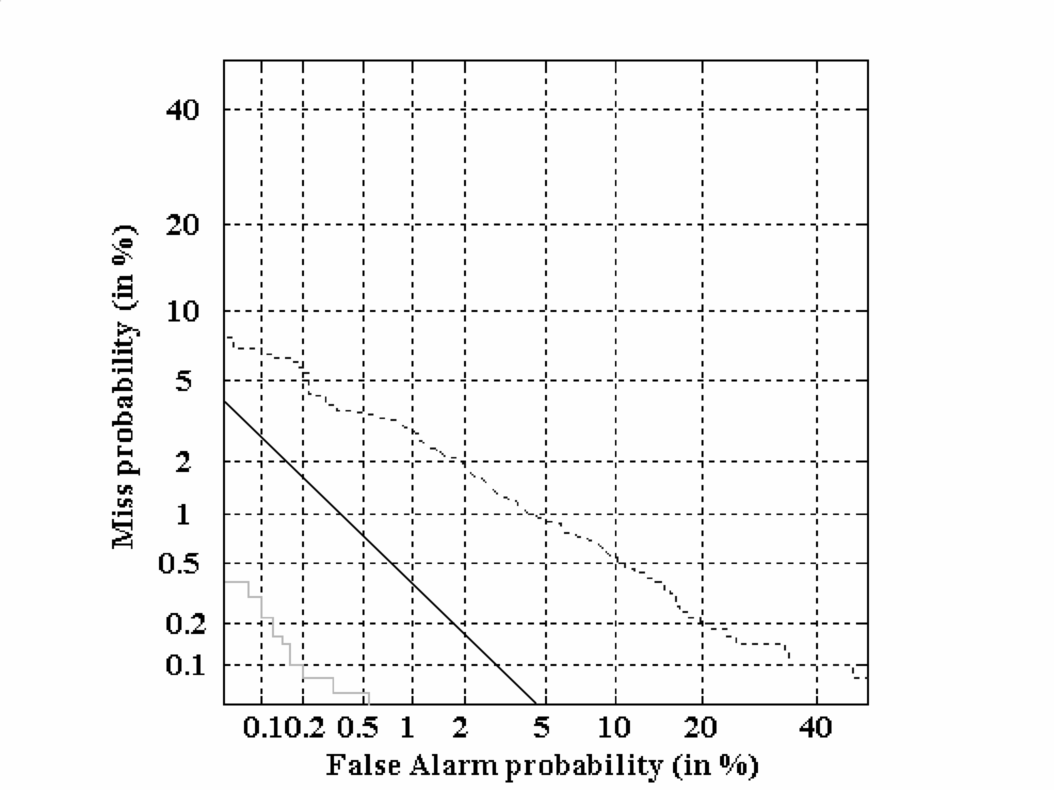
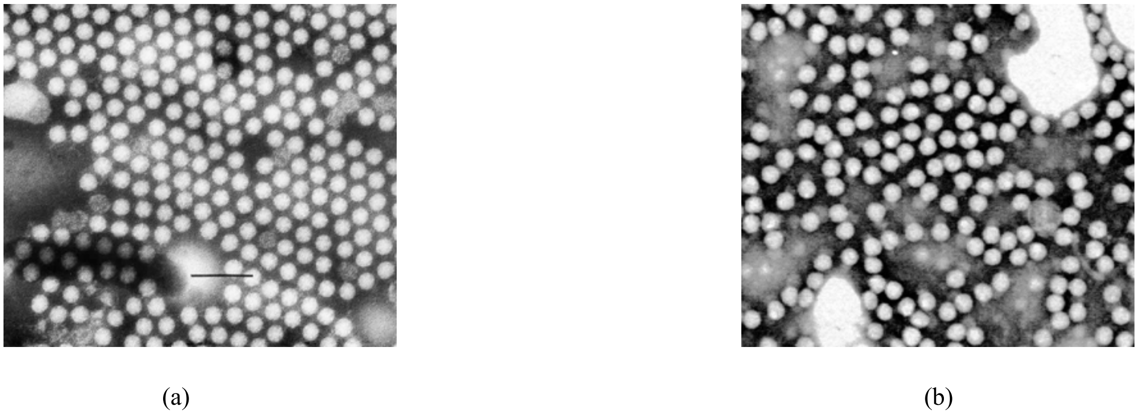


Figure 17: DET curve using the bispectral features from subpopulation of $N=15$ viral particles on a pooled population. The solid line shows an average of 2 sets of subpopulation of 15 viral particles while the dotted line shows the case for $M = 1$ ($M$, $N$, refer to equations 5, 6). The EER drops to less

than 0.2%. The darker solid line shows an output fusion (a decision to accept or reject may be made only if both individual decisions suggest so) of two test populations at the threshold of 2% EER. This shows that the averaged scores yield better performance than output fusion in this case.



(a) (b)

Figure 18: A sample image of (a) Astrovirus and (b) Hepatitis A virus. The magnification of these images are the same.

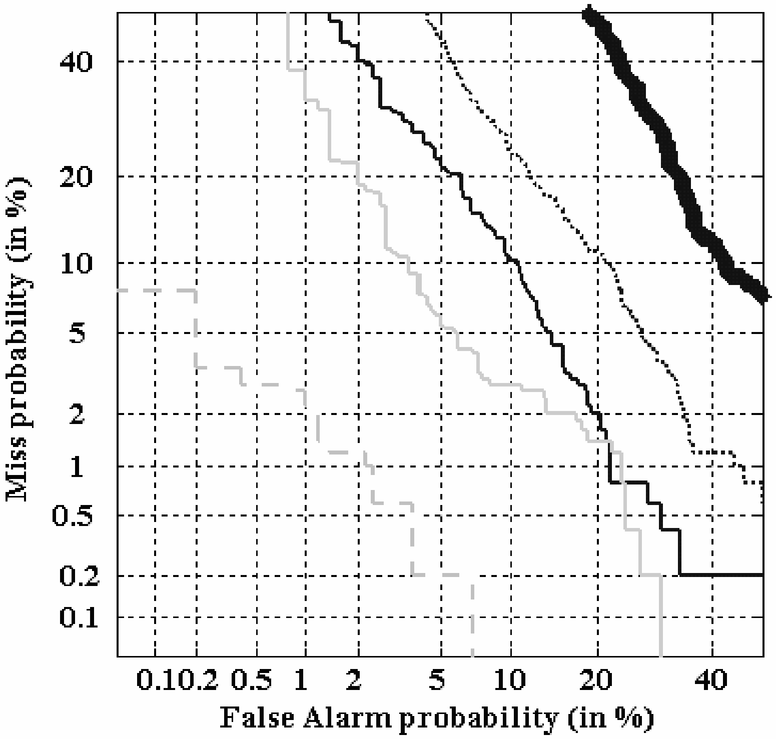
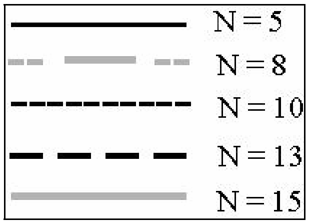
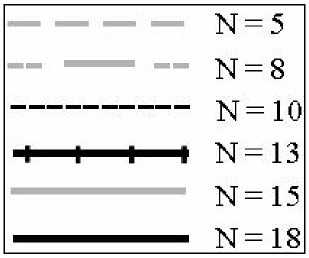


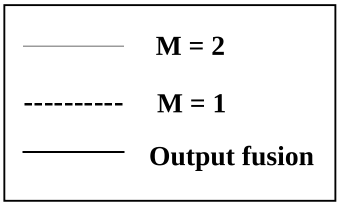
Figure 19: DET curve using the bispectral features from subpopulation of *N = 5, 10, 15, 20* particles, *M =1* and subpopulation of *N = 20, M=2* on a pooled population. The EER drops to less than 2% when features of 2 sets of subpopulation of 20 particles were averaged.



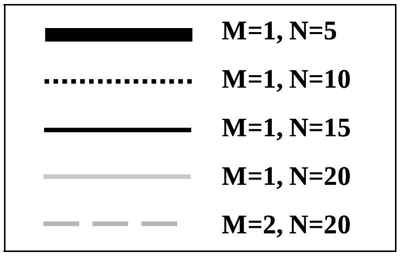
Legend for Figure 13



Legend for Figure 15



Legend for Figure 17



Legend for Figure 19