A Case Study in the use of ROC Curves for Algorithm Design

P. A. Bromiley, P. Courtney and N.A. Thacker,
Imaging Science and Biomedical Engineering,
Stopford Building, University of Manchester,
Oxford Road, Manchester, M13 9PT
paul.bromiley@man.ac.uk

Abstract

We describe the development of a vision system to detect natural events in a low-resolution image stream. The work involves the assessment of algorithmic design decisions to maximise detection reliability. This assessment is carried out by comparing measures and estimates made by the system under development with measures obtained independently. We show that even when these independent measures are themselves noisy, their independence can serve to guide design decisions and allow performance estimates to be made. Although presented here for one particular system design, we believe that such an approach will be applicable to other situations when an image-based system is to be used in the analysis of natural scenes in situations where a precise ground truth is not available.

1 Introduction

Performance evalution is esssential for providing a solid scientific basis for machine vision, and yet its importance is often understated. Current work in this area [1, 3, 5] has tended to emphasise the importance of an objective ground truth (see for example work in medical image registration [8], face recognition [7], photogrammetry [4] and graphics recognition [6]). We present a case study of the evaluation of a machine vision system, which exhibits two examples of performance characterisation in the absence of a suitable ground truth. We believe that our approach will be applicable to other situations when an image-based system is to be used in the analysis of natural scenes, operating in the absence of a precise ground truth.

The system described here is a human fall detector based on a novel infrared sensor with limited spatial resolution. It was developed by an industrial/academic collaboration between IRISYS Ltd., British Telecom Plc., the University of Liverpool, and Manchester University. It observes a natural scene of a person in their home, typically an elderly person living alone. When left undetected falls amongst the elderly often lead to aggravated injuries, admission to hospital and not infrequently death. The purpose of the fall detector was to monitor the image stream from the thermal detector for characteristic signals associated with falls in human subjects, and to issue a fall detection warning when such a motion was observed. Fall events are rather poorly defined and occur infrequently under normal circumstances, making the system design problematic. In order for the system

BMVC 2001 doi:10.5244/C.15.21

to be practically useful it had to reliably detect falls whilst issuing a minimum of false detections, making performance evaluation a vital part of this work.

The fall detector system was based on measuring the vertical velocity of a human subject. Measurements extracted from colour video data were used as a gold standard to demonstrate a correlation between the infrared data velocity estimates and the physical velocities present in the scene. ROC curves were used to perform this correlation analysis in a generic fashion, avoiding the drawbacks of more common techniques such as linear correlation analysis. Despite the fact that the velocity estimates extracted from the colour data did not represent a genuine ground truth, we show that this form of performance evaluation can be used to guide algorithmic development in an objective manner.

The fall detector itself consisted of an MLP neural network trained to detect the patterns of vertical velocities which signified falls in human subjects. An actress was employed to perform a wide variety of falls, together with non-fall events (e.g. sitting) that generated significant vertical velocities. Infrared velocity estimates from these simulations were used for both training and testing the network. A wide variety of networks were trained in order to find the optimum architecture, and the best-performing network was identified using ROC curves. A nearest neighnour classifier and a classification system based on single velocity measurements were used to generate upper and lower bounds respectively on the performance of the neural networks. We demonstrate that, using sensible definitions of true and false detections, it is possible to evaluate the performance of a system for identifying events in a temporal data stream in this manner.

2 Data collection

Data was collected to provide realistic video data on both fall and non-fall scenarios that could then be used in the construction of a fall detector. A list of types of fall (e.g. slips, trips etc.), together with non-fall scenarios that might generate significant vertical velocities (e.g. sitting), was prepared. An actress was employed to perform the scenarios, and sequences of images were captured from both the infrared sensor and a colour CCD video camera. Each scenario was performed in six orientations of the subject and the cameras, giving a total of 84 fall and 26 non-fall sequences for each camera.

3 Velocity Measurement Evaluation

Algorithms were developed to calculate the velocity of the actress from both the infrared image sequences and the colour video. The approach taken to extract velocity information from the colour video relied on the use of colour segmentation. The image segmentation algorithm is described in more detail elsewhere [2]. During the simulations, the actress wore a shirt of a different colour to any other object in the scene. A colour segmentation algorithm was used to extract the shirt region from the colour video images, allowing the centroid of the actresses upper body, and thus its velocity, to be calculated.

The approach taken to extracting velocity information from the infrared images was quite different, and exploited the basic physics of the detector itself. The infrared sensor used was a differential sensor i.e. it registered changes in temperature. Stationary objects in the scene were therefore ignored. Any moving object warmer than the background created two significant regions in the image. The first was a region of positive values

covering the pixels that that the object was moving into, which were becoming warmer. This was trailed by a region of negative values covering the pixels that the object was moving out of, which were becoming colder. If the object was colder than the background, the signs of the two regions were reversed. In either case, a zero-crossing existed between the two regions that followed the trailing edge of the moving object. The approach taken was to track this zero-crossing to give measurements of the position of the actress in the infrared images, which could in turn be used to calculate the velocity. Fig. 1 shows an example of this process for a sequence of IR images of one of the simulated falls. Since only the vertical velocity estimates were required, the infrared images were summed across rasters to produce column vectors, eliminating any horizontal velocity component. This also led to a compact method for data storage, called the "crushed" image, which contained a temporal sequence of these column vectors.

The use of two independent estimators of motion was a key part of this work. Although both used radiation from the scene, they were independent in that they sensed different kinds of information, and processed it in fundamentally different ways. The colour segmentation algorithm was based on the interaction of visible light and reflectance on clothing in spatial regions, whereas the algorithm for the infrared sensor was based on signal zero-crossings on thermal radiation using a completely different lens system. Furthermore, it is probable that the correlation between the velocities calculated from the infrared images and the physical velocities present in the scene was better than the correlation between the velocities calculated from the infrared and colour video. The two velocity extraction techniques measured slightly different quantities. In the case of the colour video, the actresses upper body was segmented from the scene and its centroid calculated. The limbs and head were ignored, and so the calculated velocities corresponded closely to the movement of the centre of gravity. In contrast, the infrared images measured the movements of all body parts. As an example, if the actress waved her arms whilst otherwise standing still this movement was detected in the infrared images but not in the colour video. We would therefore expect the information in the two estimators to suffer from noise, bias and distortion independently.

In order to examine the extent to which the infrared velocity estimates were correlated with the physical velocities present in the scene, a correlation analysis was performed on the colour and infrared velocity estimates from a sub-set of 26 of the fall scenario data sets. Initially the correlation was significant but poor. This was expected given the exacerbation of noise in the position data by the differentation used to calculate the velocity. Therefore methods of combining data within the temporal stream were studied.

A variety of smoothing techniques were tested, including a five-point moving window average, a five point median filter, and a combination of the two, termed a median rolling average (MRA) filter, which took a five-point window and averaged the median three points. Each technique had advantages. The median filter tended to be more robust to outliers, whereas the moving window average was very susceptible to outliers, but had a stronger smoothing effect. The MRA filter combined the advantages of each, providing stronger smoothing whilst retaining resistance to outliers. The key design question was to determine the best method and the most appropriate metric for their comparison.

Fig. 3 shows the result of plotting one of the post-processed estimates from the thermal sensor against the velocity estimates derived from the colour processing. The smoothed data produced tighter distributions than the unsmoothed data, but none of the three smoothing methods had an obvious advantage from simple visual inspection of their

plots. The plots for the three smoothing methods had the same typical shape: the velocity estimates from the infrared images were directly proportional to those from the colour video up to velocities of approximately 2 pixels per frame on the x-axis of the plots (corresponding to the colour video velocity estimates), and then they flattened out and appeared to saturate. The infrared velocity at which this occured was around 0.19 pixels per frame, approximately the value that would be expected from the ratio of pixel sizes in the two image types. Above this point, the infrared velocity estimates no longer increased with colour video velocity estimate. This was due to the basic physics of the detector rather than the velocity extraction methods. The thermal sensor took several seconds to saturate, and this placed an upper limit on the rate of temperature change it could detect[1].

Of the standard simple techniques used for correlation analysis, none were found suitable for this data. Linear correlation coefficients assume both linearity and a Gaussian noise distribution, neither of which were valid, and are sensitive to outliers. Gaussian fitting to the saturated region of the data (after outlier rejection), where the only variation in the data was due to noise, was also attempted, but again this made the assumption of a Gaussian distribution. The inliers and outliers of the outlier rejection were counted, but this was found to be uninformative. Each technique indicated that the smoothed data showed a stronger correlation than the unsmoothed data, but none was able to distinguish between the various smoothing techniques.

In order to overcome these drawbacks, a superior method based on ROC curves was used. Such curves are usually applied to classification systems, and so a simple classification decision was made by placing thresholds on the colour and infrared velocity estimates to divide the data into “fast” and “slow” velocities, as shown in Fig. 2a. The colour threshold was set at 2 pixels per frame, the point at which the infrared velocity estimates saturated, and the threshold which gave the maximum ability to differentiate high velocities from noise. Then a second, varying threshold was applied to the infrared velocity estimates, and was used to make a fast/slow decision based on the infrared data. The points above this second threshold therefore fell into two groups. Firstly, those defined as fast using the colour data, i.e. those with x values higher than 2 pixels per frame, represented correct decisions based on the infrared data. Secondly, those defined as slow using the colour data represented incorrect decisions. The number of data points in each of these categories was counted, and a percentage was calculated by dividing by the total number of points either above or below the threshold applied to the colour data. The two probabilities calculated in this way produced a point on the ROC curve, and by varying the infrared velocity threshold the whole ROC curve was plotted for each smoothing option.

The threshold applied to the colour data was placed at the point which would give the maximum ability to distinguish high velocities in the infrared data. At the most basic level this is precisely the task that the fall detector system had to perform. It is therefore justifiable to state that the smoothing filter which produced the strongest correlation according to the ROC curves was the best filter to apply in this problem. Fig. 4 shows the ROC curves, and it is clear that the MRA filter gave the stongest correlation. It should be noted that the aim of this analysis was to rank the filters, rather than calculate an absolute value for the correlation. The method described provides a generic technique for this type of analysis that avoids assumptions of either a linear correlation or a specific noise distribution, which are often invalid. Furthermore, it is clear that an independent measurement

[1] Following discussions with the sensor manufacturer, it seems possible that this arose due to pre-processing in the sensor, which has since been modified.

can be used in the absence of a suitable ground truth to guide algorithmic development.

4 Neural Network Evaluation

To produce the fall detector itself, an MLP neural network was trained to take temporal windows of velocity measurements from the infrared sensor and make a fall/non-fall decision. The training process allowed the neural network to automatically identify the regions in the input pattern space that contained the fall data points i.e. the patterns of velocities characteristic of a fall. The key design decision was which network architecture to use i.e. how close the network performance was to optimal. The input data for the network represented high-dimensional vectors of velocity measurements for temporal windows. Each temporal window of velocity measurements defined a point in the high-dimensional space in which the neural network was operating. The network then attempted to define decision boundaries which encompassed the region of the space containing the points corresponding to falls, thus identifying the characteristic patterns of velocities present during falls. The dimensionality of the input vectors, and thus the number of input nodes in the network, was selected on the basis of the timespan of the falls recorded during the simulations.

Training data for the neural networks were provided by synchronising the infrared and colour images, identifying the positions of falls by visual comparison, and labelling the three highest-velocity points during each fall event as falls. The remaining data points were labelled as non-falls. This provided approximately 50,000 classified data points: 20% were extracted at random for training, leaving 80% for testing.

A total of 120 MLP neural networks were trained, in an attempt to find the optimum architecture for this problem. All available parameters were varied, including: the number of hidden nodes (2-21); the number of hidden layers (1-2); the initialisation factor for the network weights; the training algorithm - both RPROP (resilient back-propagation) and CGM (conjugate gradient minimisation); and the number of iterations (200-2000).

The trained neural networks gave outputs in the form of a probability i.e. ranging from 0 to 1, with higher values indicating that the input data were more likely to represent a fall. Rather than apply a simple threshold to this output, counting all values above the threshold as falls and all values below it as non-falls, a more sophisticated method was applied in order to increase detection reliability. The output from the network was monitored for local peaks in the probability, and then the height of the peak was compared to a threshold. This ensured that, during extended high-velocity events such as falls, the network issued only one detection, rather than a series of detections.

ROC curves were again used to rank the performance of the various neural networks and thus to select the best architecture for this problem. The issue of which quantities to plot on the axes of the ROC curves was problematic in this case, and we believe this to be a common problem in the interpretation of temporal events. Firstly, falls were marked in the input data only at the three points of highest velocity during the fall. The falls were, however, extended events covering more than three frames, and so it was reasonable to expect the network to issue detections shortly before or after the marked three-frame window. It was therefore unreasonable to count as correct only those detections which coincided exactly with the marked positions of falls. Secondly, the network may issue more than one detection during the course of a fall. This provided the choice of whether

to count all of the detections occurring close to a marked fall as correct detections, or whether to reverse the problem and look at how many marked falls had one or more detections in their temporal vicinity.

A similar issue applied to false detections. If they can be caused by an event such as the subject sitting down, which is not marked as a fall but nevertheless causes a local period of high velocities, then several false fall detections may be issued during the event. This raised the problem of whether to count these as multiple false detections or as a single false detection, which in turn raised the logical problem of how to specify non-events. It should be noted that any choice allowed the network performances to be compared as long as the same procedure was applied to the outputs of all networks.

The approach chosen for plotting the ROC curves for the neural network outputs is shown in Fig. 2b. The outputs from the neural networks were monitored, and the heights of local peaks were compared to some threshold. Peaks higher than the threshold represented fall detections, and the positions of these fall detections in the input data file were recorded. A second loop scanned through the data looking for the specified positions of falls. Any true fall that had one or more fall detections by the neural network within 20 frames in either temporal direction was counted as a correct detection, and any specified fall that did not have such a detection by the network within that time period was counted as a real fall not detected. Therefore the reconstructed signal axis of the ROC curves showed the number of genuine falls detected by the network as a proportion of the total number of genuine falls. This treatment ensured ease of comparison between the different networks, as the maximum possible number of events represented in the recovered signal measurement was limited to the number of labelled falls in the data. It avoided the potential problems inherent in examining the raw number of detections by the network e.g. if a particular network produced multiple detections during a small percentage of the labelled falls, but did not detect the remaining labelled falls, looking at the raw number of detections gave that behaviour an unfair advantage, whereas looking at the number of genuine falls which had one or more detections did not. Finally, the procedure was repeated, varying the threshold, to plot out the whole span of the ROC curve.

The treatment applied to generating data for the error rate axis of the ROC curves was slightly different. The same arguments applied, in that there were underlying events in the data (e.g. sitting down) which may generate one or more false fall detections by the neural network. However, in this case there was the logical problem of how to specify a non-event. Therefore, in the absence of a viable solution to this problem, the error rate was calculated as the number of false detections divided by the total number of data points which did not fall within a forty-point window around one of the falls.

The results were also compared to those obtained from a nearest neighbour classifier, which for unlimited data can be shown to approach the Bayes optimal classification i.e. the classification that would be obtained if the underlying probability distributions which generated the data were known and were used at each point in the input pattern space to make the classification decision. A nearest neighbour classifier operates by searching for the n nearest points to each data point using e.g. a Euclidean distance metric, and then taking an average of the labels assigned to these points i.e. whether the point has been specified as belonging to a fall or not. A threshold was then applied to this score, and points with higher values represented fall detections by the nearest neighbour classifier. As with the neural network, this output threshold was the parameter that was varied to plot the whole range of the ROC curve. The treatment applied to convert these detections

into percentages for ROC curve plotting was kept exactly the same as that used with the neural network ROC curves, including scanning for local peaks in the output, to ensure that the curves could be compared. The number of nearest neighbour points used to make the classification decision was varied between 10 and 50, and the best-performing nearest neighbour classifier was selected using the ROC curves.

In order to produce a measurement of the performance improvement gained through applying a neural network to this problem, a classifier based on single velocity data points was used. The ROC curve for this classification system was produced and compared to those for the neural networks and nearest neighbour classifiers. Fig. 5 shows the ROC curves for the best networks trained using RPROP and CGM, the nearest neighbour classifier, and the single data point classifier. Network 89, which had 18 hidden nodes in one layer and was trained with 2000 iterations of CGM, gave the overall best performance. The performance approached that of the nearest neighbour classifier and was significantly better than the single data point decision, justifying the application of a neural network to this problem. The inability of any of the networks to achieve optimal performance was a reflection of the inherent confusability in the data. Some high velocity non-fall events, such as sitting, can include a period of free-fall, and so the underlying physics of these events and genuine falls is identical, and they could not be distinguished on the basis of vertical velocity measurements alone. However, the false detections generated by this mechanism were limited to the most violent sitting events, and such events are rare in the target group for this system. In-situ studies with the target group to quantify the typical occurrence rate of these events, when combined with the ROC curve for the best-performing network, will be necessary and sufficient to calculate the false detection rate for this system in a practical situation.

5 Conclusions

The study of performance evaluation is vital in placing machine vision on a solid scientific basis. We have described a case study: the development of a vision system to detect natural events in a low-resolution image stream. The work has involved two distinct examples of the assessment of algorithmic design decisions to maximise detection reliability. In the first example this assessment was carried out by comparing measures and estimates made by the system under development with measures obtained independently, in the absence of genuine ground truth data. We have shown that even when these independent measures are themselves noisy, their independence can serve to guide rational design decisions and allow performance estimates to be made. In the second example we have shown that the temporal identification of events can be subjected to a similar performance analysis, and that upper and lower bounds placed on the data by independent classifiers can guide algorithmic design in a different way, providing an estimate of the proximity of the system to optimal performance. In both cases the analyses were performed using ROC curves, showing that, with suitable consideration of the definitions of true and false detection rates, such curves can provide a unified, generic approach to performance evaluation in a wide range of machine vision problems. We therefore believe that, although presented here for one specific system design, such an approach will be applicable to other situations when an image-based system is to be used in the analysis of natural scenes in the absence of a precise ground truth.

Acknowledgments

The authors would like to acknowledge the support of the MEDLINK programme, grant no. P169, in funding part of this work. Data was collected in collaboration with Dr. Andrew Sixsmith of the University of Liverpool. The support of the IST programme of the European Commission is also gratefully acknowledged under the PCCV project (Performance Characterisation of Computer Vision Techniques) IST-1999-14159. All software is freely available from the TINA website www.niac.man.ac.uk/Tina.

References

[1] K.W. Bowyer and P.J. Phillips, *Empirical Evaluation Techniques in Computer Vision*, IEEE Computer Press, 1998.

[2] P.A. Bromiley, N.A.Thacker and P. Courtney, *Colour Image Segmentation by Non-Parametric Density Estimation in Colour Space*, in Proc. BMVC 2001, BMVA, 2001.

[3] H. I. Christensen and W. Foerstner, *Machine Vision Applications: Special issue on Performance Characteristics of Vision Algorithms*, vol. 9 (5/6), 1997, pp.215-218.

[4] E. Guelch, *Results of Tests on Image Matching of ISPRS III/4*, Intl. Archives of Photogrammetry and Remote Sensing, 27(III), 1988, pp.254-271.

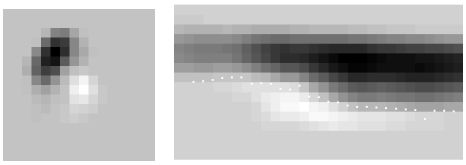
[5] R. Klette, H.H. Stiehl, M.A. Viergever and K.L. Vincken, *Performance Characterization in Computer Vision*, Kluwer series on Computational Imaging and Vision, 2000.

[6] I.T. Phillips and A.K. Chhabra, *Empirical Performance Evaluation of Graphics Recognition Systems*, IEEE Trans PAMI, 21(9), 1999, pp.849-870.

[7] P.J. Phillips, H. Moon, S.A. Rizvi and P.J. Rauss, *The FERET Evaluation Methodology for Face-Recognition Algorithms*, IEEE Trans PAMI, 2000.

[8] J. West, J.M. Fitzpatrick, *et al.*, *Comparison and Evaluation of Retrospective Intermodality Brain Image Registration Techniques*, J. Comput. Assist. Tomography, 21, 1997, pp.554-566.

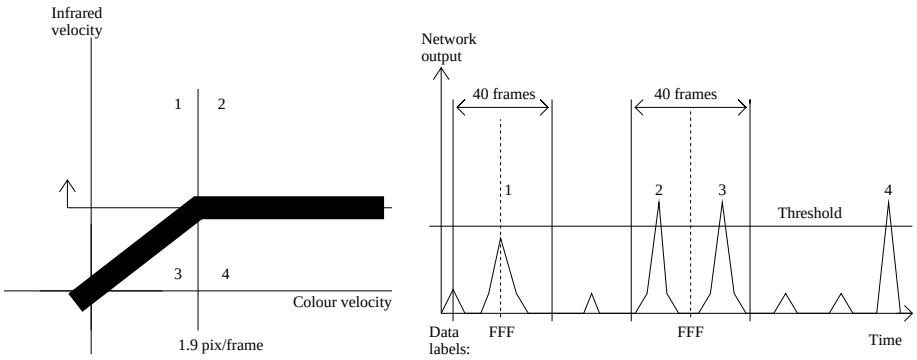
6 Figures



(a) Single infrared frame.

(b) The “crushed” image.

Figure 1: A frame from the infrared sensor (a) taken during a fall, showing the positive (white) and negative (black) regions, and the “crushed” image (b) for 30 frames during this fall, where the white points are the zero-crossings.



(a) (b)

Figure 2: ROC curve generation mechanisms. In the velocity correlation analysis (a) the two thresholds divided the data (the dark band) into four quadrants. Data points in quadrant 2 were classified "fast" by both IR and colour measurements, and those in quadrant 1 were classified "slow" by the IR but "fast" by the colour measurement. Thus the true acceptance rate was $N_2/(N_2 + N_4)$ and the false acceptance rate was $N_1/(N_1 + N_3)$, where N is the number of points in the quadrant. The method applied to the neural network output is represented in (b): points above the threshold represented network detections. Event 1 is a genuine fall not detected; 2 and 3 are both detections of a genuine fall, but lie in the 40-point window around a fall label and so are counted as a single detection; 4 is a false detection since it lies outside a 40-point window around a fall label in the data.

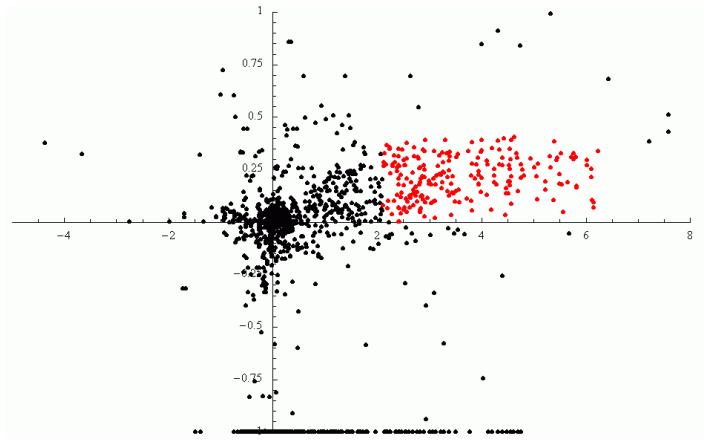


Figure 3: The velocity data after 5 point median/ 3 point moving window average, showing infrared velocity against colour video velocity.

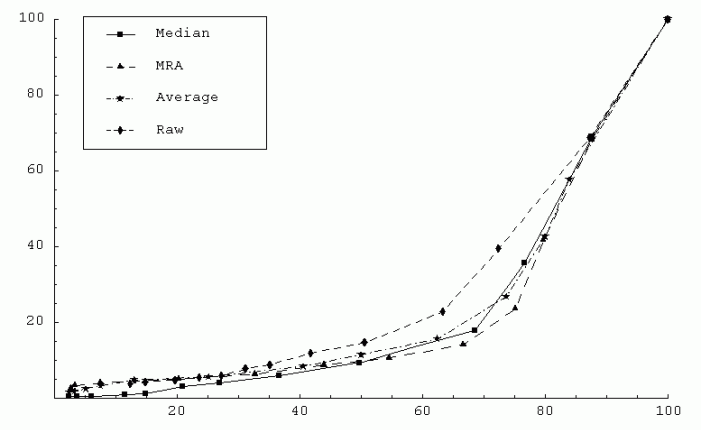


Figure 4: ROC curves (showing percentage false acceptance rate plotted against percentage true acceptance rate) for the four velocity calculation methods.

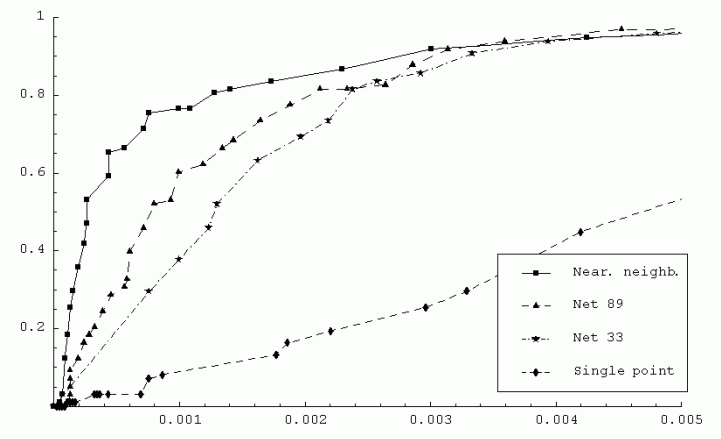


Figure 5: ROC curves for (reading across the graph from left to right) the nearest neighbour classifier using 30 neighbours, net 89, net 33, and for classifications based on single velocity measurements. The x-axis shows the error rate as a proportion of the total number of data points and the y-axis shows the proportion of true falls detected.