Mutual Information Based Gesture Recognition

Peter Harding, Michael Topsom, and Nicholas Costen

School of Computing, Mathematics and Digital Media, Manchester Metropolitan University, UK
p.harding@mmu.ac.uk, m.topsom@mmu.ac.uk, n.costen@mmu.ac.uk

Abstract. Proliferation of gestural interfaces necessitates the creation of robust gesture recognition systems. A novel technique using Mutual Information to classify gestures in a recognition system is presented. As this technique is based on well-known information theory metrics the underlying operation is not as complex as many other techniques which allows for this technique to be easily implemented. A high recognition rate was achieved, 98.55% with recognition occuring in under 10ms.

Keywords: Mutual Information, Pattern Recognition, Classification, User Interface.

1 Introduction

The recent proliferation of touch screen, accelerometer based, haptic, and other gestural interfaces necessitates the creation of robust gesture recognition systems to ensure their fast and reliable operation. The inclusion of these interfaces in modern electronic devices (e.g. mobile phones, hand-held touch devices [15], and computer game consoles [2]), which often have access to limited processing power, requires these recognition systems to be computationally efficient to allow the classification of input at a near real-time speed which is considered acceptable to users [10]. This paper presents a lightweight, simple to implement recognition system, based on information theory techniques, which fulfils these criteria, and details the results of testing as to illustrate the effectiveness of this system.

2 Recognition Problem

The recognition problem addressed is that of the correct classification of two-dimensional glyphs [12], of the type routinely used as control input for touch screen, stylus or wand driven devices ([8] gives an example control interface). A set of sixteen gestural input glyphs is employed, as seen in Figure 1, which has previously been used (in whole or in part) during the testing of this type of system [7, 16], and is believed to offer a reasonable cross section of the possible gestures that would be found in modern user interfaces.

The recognition of these glyphs must be shown to be robust, as even a single user system may have to deal with "noisy" input, for various reasons [17]. The

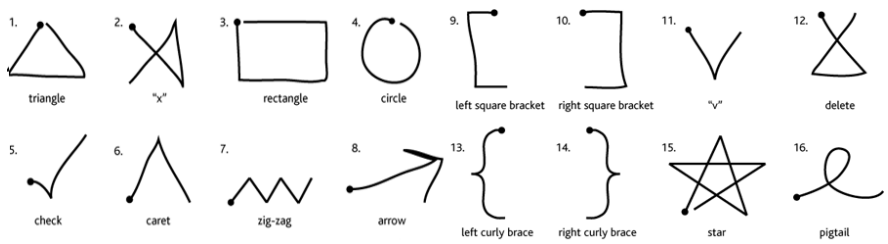

Fig. 1. Unistroke gestures.

recognition provided must also be shown to be computationally efficient, allowing recognition at a rate that may be considered to be in real time, from the user's perspective.

3 Recognition Using Mutual Information

The proposed technique has two basic sections, the first of these concerns the processing of raw data to transform it to a more usable form. The data is then passed to the classification system, which performs the comparisons against a template bank.

The data capture system used is touch screen based, and reads user input as a set of N Cartesian coordinates sampled from a single continuous motion. This method was chosen as it is analogous to myriad accelerometer, wand and mouse based interfaces to modern electronic devices.

3.1 Pre-processing

The initial processing steps performed are not uncommon in classification systems, and consist of the re-sampling, rotation and scaling of the raw data.

User input was found to be of varying size (i.e. the number of points), due to factors such as the speed with which the gesture was made and the data capture technique. The first pre-processing step normalizes the number of points. The raw data are re-sampled, interpolating to ensure that all points are of a fixed size and equidistantly spaced, leaving a vector of points, N'. Figure 3.1 shows a re-sampling of a single input of size N to create a processed set of points at size of N'. Testing shows that user input data is rarely oriented correctly. The second pre-processing stage rotates the gesture based on the angle between the first recorded point of the input, and the centroid of the input, see Figure 3.1, so the angle is uniformly 0^c. This mitigates any error caused by poorly orientated input, and is required to ensure the robustness of the recognition technique. Finally the gestures are scaled to have a fixed bounding box. Each (x, y) value is transformed to lie within the range $\pm\kappa$ (an arbitrarily chosen scaling constant), where

$$v' = 2\kappa \left(\frac{v - min(V)}{max(V)} \right) - \kappa, \quad v \in V. \tag{1}$$

This is applied separately to the (x, y) values (v is an arbitrary symbol) and $\kappa = 1$.

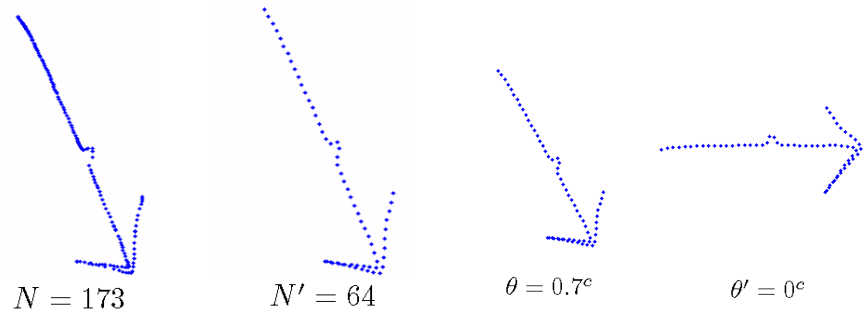

Fig. 2. Resampling and subsequent rotation of an input glyph.

3.2 Mutual Information Analysis

The actual recognition method for these glyphs is based on Mutual Information (MI), a probabilistic method for quantifying the interdependence of two signals. It has previously been employed as an analytical technique in many areas [3, 5, 4], including classification tasks [1, 14], but appears not to have been applied to the problem of gesture recognition.

The weighted mutual information of two *discrete* time-series variables, T and U, is defined as

$$I(T;U) = \sum_{i,j} w(t_i, u_j) P(t_i, u_j) \log_n \frac{P(t_i, u_j)}{P(t_i)P(u_j)} \tag{2}$$

where $P(t_i)$, $P(u_j)$, and $P(t_i, u_j)$ are the individual and joint probability distributions of T and U respectively. In general terms, the MI of two signals quantifies their interdependence; therefore if T and U are entirely independent, then $I(T;U) = 0$, but in *all* other cases $I(T;U) > 0$. The use of weights $w(t_i, u_j)$ in MI can increase recognition accuracy [11, 6]. This scales $I(t_i, u_j)$ either upwards if $t_i \approx u_j$, or downwards if $t_i \neq u_j$. This creates a reward structure for correct values, whilst penalising any pairs of values that are not correctly identified.

The weighting function employed in this paper is a Gaussian distributed function of the absolute difference of the two input values, scaled by σ. Initial experimentation showed that the application of the weighting matrix improved results considerably, but only for very small variances, so $\sigma^2 = 10^{-2}$, while $\kappa = 1$.

User input is read in the form of a vector of Cartesian coordinates, U, which is then re-sampled, rotated and scaled as described previously. These coordinates are then separated into their x and y components, and discretised into R equally

sized bins $R \in \{3, \ldots, 9\}$, leaving two discrete vectors U_x and U_y. The each set of template data, T, is processed in exactly the same manner. The mutual information, I, is calculated as $I = I(T_x; U_x) \times I(T_y; U_y)$.

4 Experimentation and Results

Experimentation was carried out in three stages; ideal data recognition, user data (including comparison with an existing system) and additional noise. These testing stages were designed to test the limits of the system under different circumstances. Idealised data testing shows performance with known inputs and parameterized variations. The user data testing tests the ability of the method to classify gestures in a real world context. The addition of noise to data tests the ability of the system to recognise and correctly classify distorted data, which is an important test for any system that may not be deployed in an ideal scenario.

4.1 Ideal Data Recognition

A set of *perfect patterns* (precisely defined, uniform inputs) were created, which consisted of five points joined by four straight lines. The position of the final point of the pattern was then repositioned to a total 121 different locations, which were uniformly distributed inside the pattern, to create a test set. An example of one of these patterns may be seen in Figure 3.

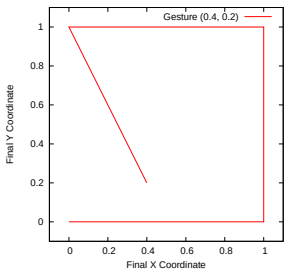

Fig. 3. The "ideal data" pattern; the outer lines enforce orientation. The variable point is at (0.4,0.2).

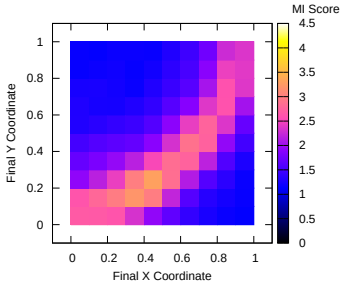

Fig. 4. Variation in recognition performance with location of the point. The probe point is at (0.4,0.2).

To ensure that the system could recognise data reliably *all* of the 121 data items were both used as a test set and a template set for these experiments. The system was presented with each of the 121 data items in turn, and then logged both the classification given and the MI score returned at $R = 7$. The system achieved a 100% accuracy in classification during these tests, i.e. each input presented was identified as its corresponding data item from the template

bank. This shows that the system is able to accurately distinguish between large sets of relatively similar gestures and retain a good degree of accuracy.

To investigate the variance in MI results across increasingly distorted versions of the same input, the MI scores returned when the input shown in Figure 3 was compared with *all* of the 121 templates. Figure 4 shows that higher recognition values lie on the arc where the pattern retains exactly the same length, i.e. the points at which the length of the fourth line in the pattern retains the same length as in the recognition template. When two ideal data patterns are of the same length re-sampling will produce many corresponding points along the first three straight lines increasing the MI score.

4.2 User Data

Sets of test data, consisting of three examples of each of the sixteen Figure 1 glyphs, were collected from 26 test subjects. Four data items were not recorded due to experimenter error, resulting in a total of 1244 unique data items being used for testing. The testing was carried out using a *leave-one-user-out* testing strategy; in turn, each user was supplied probes, and all remaining 25 sets of user data formed the gallery.

Considering the fast and lightweight nature of the MI system, a suitable comparision is the $1 recogniser [16]. This has been shown to operate at a faster speed than both a Rubine Classifier [13] and a Dynamic Time Warping based matcher [9], and has at worst comparable but often more accurate recognition to these systems. The same gesture set and testing strategy were employed. The recognition rate and recognition speed was recorded for various re-sampling values of N with both systems.

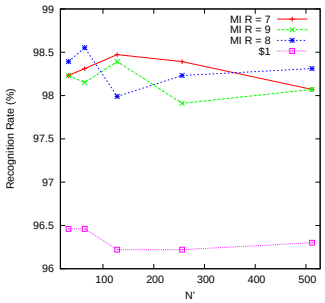

Fig. 5. Comparison of the accuracy of the MI recognition system ($7 \leq R \leq 9$) and $1 recogniser.

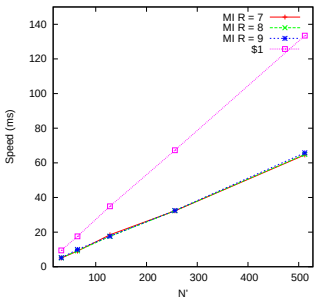

Fig. 6. Comparison of the speed of the MI recognition system ($7 \leq R \leq 9$) and $1 recogniser.

The accuracy of the MI technique reached a maximum recognition rate of 98.55% while the $1 recogniser reached a maximum recognition rate of 96.46%. The confusion matrix in Figure 7 shows the results for the MI classifier. For all

values of $R > 3$ the recognition technique out performed the \$1 recogniser in classification and speed. Figure 5 shows the MI technique's highest recognition rates ($7 \leq R \leq 9$) compared to the absolute highest recognition rate achieved by the \$1 recogniser.

The speed at which each technique recognised and classified an input gesture was recorded, this was calculated by sequentially recognising each of the 1244 data items and averaging the net time taken. Experiments were run on a desktop computer with Intel® Core™2 Quad CPU Q2800 running at 2.33GHz and 4GB of RAM with Java version 1.6.0_11. Figure 6 shows the speeds at which the MI system performed recognitions (binning values $7 \leq R \leq 9$; note these render to the same line). The MI system took approximately half the time to perform a classification that was required by the \$1 recogniser for a give value of N'. The value of R was found to have little effect on the speed of the MI system in comparision with the value of N'.

4.3 Additional Noise

To further investigate the robustness of the recognition technique, a series of experiments were run in which with additional noise applied to the probes before classification. The noise, η, was applied according to a directed, Gaussian distributed function

$$\eta_{n+1} = \eta_n + d\,|(\mathcal{N} : \mu, \sigma)| \quad d \in \{-1, 1\} \tag{3}$$

where d defines the directionality of the noise, and will change with a probability of $P(d_{n+1} = -d_n) = \frac{N}{2}$, yielding one expected change in the directionality of the noise for each probe. Noise is applied independently to both the x and y components of the signal, so at any time each component will have a separate and independent η. The noise is cumulative, which ensures that the signal will not be raised and lowered repeatedly; rather it will increased or decreased in a natural manner over time. This is arguably similar to the atypicalities found in human movements.

The same testing technique was employed across the new data set. The results of these experiments can be seen in Figure 8. The best recognition rates were achieved at low σ and μ values, where the recognition rate peaked at the 98.55% recorded during the user testing experiments, and recognition rates show a steady decrease as both μ and σ is increased. Even at the largest values $\mu = 10$ and $\sigma^2 = 10$ (note: maximum bounds of the glyphs were approximately 350 by 350 pixels before processing) the lowest recognition rate recorded was still 75.8%, which is twelve times greater than the naïve rate for this template set.

5 Discussion and Conclusions

Mutual information has been shown to work well when applied to the recognition of 2D gestures. In this series of experiments the MI system was shown to classify

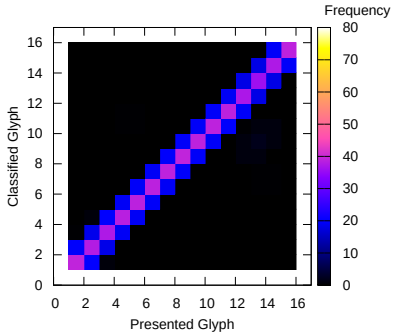

Fig. 7. Confusion matrix showing the probe glyph against the system's classification.

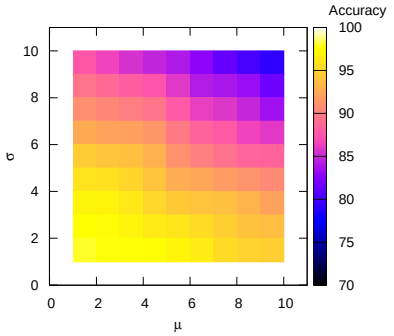

Fig. 8. Recognition accuracy of the MI system with varying values of μ and σ in the added noise.

gestures with a high degree of accuracy; with a 100% recognition rate on artificial gesture data and 98.55% with user gesture data. The addition of noise to the user data lowered the accuracy of recognitions, although a recognition rate of over 75% was achieved in the worst conditions.

The recognition system managed to perform recognitions, on average, in under 70ms in the worst cases (highest R and N' values). The optimum recognition rates (98.55%) were achieved in under 10ms. The limiting factor in terms of speed of recognition was found to be the re-sampling rate N', this is not considered a problem as the optimum value for N' will, in most circumstances, be dictated by the sampling rate of the hardware in question. In the case of the machine used during the testing covered in this paper the maximum number of samples collected for any gesture was less than 500 points and was regularly found to be lower than 200 points. It is safe to assume that systems that have a higher sampling rate are also likely to have more processing power available for the MI recognition technique itself. For a user interface to be seen as responsive by users, it is suggested that the system should respond in under 100ms [10], the MI based system fulfils this requirement amply. It is also significantly more accurate and faster than other algorithms.

6 Further Work

As the x and y components of the signals are analysed seperately this method can be simply extended to a third dimension, allowing for input from a 3-axis accelerometer based device. As the technique has been found to be so fast, a large template bank was used during these experiments; template reduction techniques may be adapted to further increase recognition speed, which could allow a whole

new area of low computational power micro-devices to incorporate gesture based control techniques into their software.

References

1. L. Bahl, P. Brown, P. De Souza, and R. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52, 1986.
2. A. H. Cummings. The evolution of game controllers and control schemes and their effect on their games. In *The 17th Annual University of Southampton Multimedia Systems Conference*, 2007.
3. A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Physics Review A*, 33(2):1134–1140, 1986.
4. P. J. Harding, M. Amos, and S. Gwynne. Mutual information for the detection of crush. In *Proc. 4th Intl. Conference on Pedestrian and Evacuation Dynamics*, 2010.
5. J. Jeong, J. Gore, B. Peterson, et al. Mutual information analysis of the EEG in patients with Alzheimer's disease. *Clinical Neurophysiology*, 112(5):827–835, 2001.
6. L. Junli, C. Rijuan, J. Linpeng, and W. Ping. A medical image registration method based on weighted mutual information. In *Bioinformatics and Biomedical Engineering*, pages 2549–2552, 2008.
7. A. C. Long, Jr., J. A. Landay, L. A. Rowe, and J. Michiels. Visual similarity of pen gestures. In *CHI '00*, pages 360–367.
8. M. Moyle and A. Cockburn. The design and evaluation of a flick gesture for 'back' and 'forward' in web browsers. In *AUIC*, volume 18 of *CRPIT*, pages 39–46. Australian Computer Society, 2003.
9. C. Myers and L. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389–1409, 1981.
10. J. Nielsen. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995.
11. J. Novovičová, P. Somol, M. Haindl, and P. Pudil. Conditional mutual information based feature selection for classification task. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 417–426. Springer-Verlag, 2007.
12. J. Rubin. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. Wiley, 1994.
13. D. Rubine. Specifying gestures by example. In *SIGGRAPH '91*, pages 329–337.
14. C. Shan, S. Gong, and P. McOwan. Conditional mutual information based boosting for facial expression recognition. In *BMVC*, volume 1, pages 399–408, 2005.
15. A. D. Wilson. Touchlight: an imaging touch screen and display for gesture-based interaction. In *ICMI*, pages 69–76. ACM, 2004.
16. J. O. Wobbrock, A. D. Wilson, and Y. Li. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In *UIST*, pages 159–168. ACM, 2007.
17. W. Yee. Potential limitations of multi-touch gesture vocabulary: Differentiation, adoption, fatigue. In *HCI (2)*, LNCS 5611:291-300, 2009.