BIOMETRIC AUTHENTICATION USING MOUSE DYNAMICS

GROUP 6 TEAM MEMBERS:

Avinab Saha (15EC10071)
Atul (15EC10067)
Kanakvi Aggarwal (15EC10068)
Joyjit Paul (15EE10059)
Sohom Chakraborty (15EE10060)

INTRODUCTION: Need & Motivation

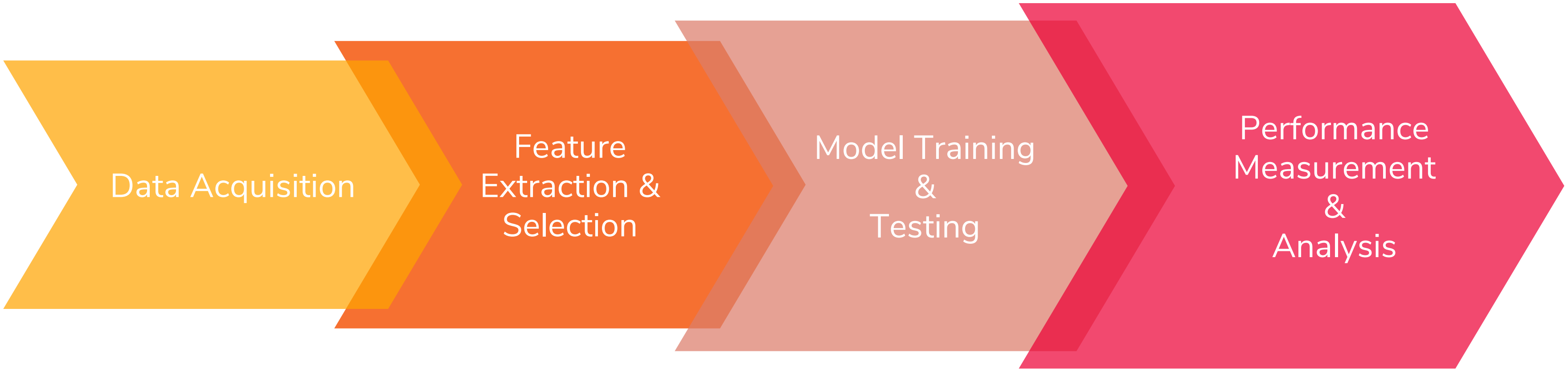
- Personal identification systems have gained interest for security and personal reasons
- Keys, magnetic cards, chip cards have been in use for a long time
- A more reliable system has been the need for a long time
- Such a system would be immune to authorization theft or loss
- Biometric authentications have surfaced due to this need
- Physiological and behavioural traits can prove to be more reliable for identification and imposter detection
- Mouse dynamics is one of the behavioural biometric technique which is based on human-computer interaction

PROBLEM STATEMENT

To authenticate the user by training and testing the assigned classifier by the neutral, happy and sad mood data

- **Acquire data of mouse dynamics.** The data collection process involves typing sentences using a virtual keyboard before and after watching a video to capture the different moods of the user and track the mouse movements in these moods
- **Recognise the user using mouse dynamics.** One Class Support Vector Machine is implemented to differentiate between and authenticate the user
- **Apply five-fold validation to report results.** The accuracy of correct authentication of the user by classifier is reported.

SYSTEM ARCHITECTURE



THE RAW DATA

- Mouse strokes are analysed for detecting patterns in the biometric authentication system, that is, the classifier
- A mouse stroke is the set of points between two mouse clicks
- The mouse contains the following information:
  - Mouse move
  - Mouse press
  - Mouse release
  - Mouse drag
  - X,Y coordinates of the screen
- Processing is done on strokes to extract spatial and temporal information
- The statistics of the above information is analysed to create features that are inputs to the classifier

THE RAW DATA

```
***********************************
LOGGING TIME: 20181020_140619
CLIENT IP: 192.168.137.1
USERNAME: Kanakvi Aggarwal
OS: Windows 10
********************Neutral database*************
MM, 29, 0, 1541893259210
MM, 10, 1, 3576
MM, 40, 1, 12
MM, 22, 1, 12
MM, 68, 40, 48
MM, 42, 6, 11
MM, 62, 1, 79
MM, 41, 0, 211
MM, 18, 3, 11
MM, 2, 6, 13
MM, 3, 23, 313
MM, 9, 21, 11
MM, 16, 17, 12
MM, 23, 16, 12
MM, 27, 16, 12
MM, 30, 15, 13
MM, 31, 14, 11
MM, 33, 14, 12
MM, 34, 14, 49
MC, 1, 404
MM, 53, 13, 2371
MM, 87, 11, 12
MM, 134, 4, 11
MM, 175, 4, 12
```

Fig 1: The format of data saved by the software

Notation	Meaning
MC, n, t:	*Mouse Clicked, Click count, Relative time*
MP, n, t:	*Mouse Pressed, Button ID, Relative time*
MR, n, t:	*Mouse Released, Button ID, Relative time*
MM, x, y, t:	*Mouse Moved, x-coordinate, y-coordinate, Relative time*
MD, x, y, t:	*Mouse Dragged, x-coordinate, y-coordinate, Relative time*
MWM, x, y, w, a, s, t:	*Mouse Wheel Moved, x-coordinate, y-coordinate, Wheel rotation sense, Amount of scrolling, Scroll type, Relative time*

Table 1: Mouse data logging format

SPATIAL INFORMATION

- Horizontal coordinates
- Vertical coordinates
- Path distance from the origin
- Angle of the path with respect to X axis
- Curvature of the path
- Derivative of the curvature of the path

TEMPORAL INFORMATION

- Input x values
- Input y values
- Input t values
- Horizontal velocity
- Vertical velocity
- Tangential velocity
- Tangential acceleration
- Tangential jerk
- Angular velocity

FEATURES EXTRACTED

- Out of the above mentioned information vectors, all except input x, y, t and path distance from the origin vectors were not used, as they were used to derive other vectors
- The features extracted for these vectors were:
  - Statistical features:
    - Mean
    - Standard deviation
    - Maximum
    - Minimum
    - Range
  - Straightness of the path
  - Jitter
  - High curvature points (also called critical points, can be multiple for the same vector)
  - Number of pauses, paused time and paused time ratio

DIMENSIONALITY REDUCTION

Random Forest Classifier was used to select the most relevant features

- **The number of features generated initially was very high.** Therefore we performed dimensionality reduction.
- **Random Forest Classifier was implemented.** The features with information gain > 0.005 were retained. This gave <10 features to use as inputs.
- **Why?** The number of data points was comparable to the number of features.

```
Using Data From user-15 to train our model
Ranking of features: [(0.5672, 'pause_count'), (0.0872, 'total_pause_time'), (0.0558, 't'), (0.0469, 'pause_time_rat
io'), (0.0317, 'x_mean'), (0.024, 'l'), (0.009, 'critical_count'), (0.0089, 'theta_std'), (0.0077, 'theta_mean'), (0
.0076, 'click_time'), (0.0075, 'c_mean'), (0.0066, 'vx_mean'), (0.006, 'w_min'), (0.0059, 'w_diff'), (0.0058, 'vdot_
mean'), (0.0056, 'w_max'), (0.0052, 'v_mean'), (0.0051, 'vdd_mean'), (0.0049, 'y_std'), (0.0049, 'delta_c_mean'), (0
.0048, 'x_std'), (0.0047, 'vdd_std'), (0.0045, 'y_mean'), (0.0042, 'vdd_diff'), (0.004, 'vdd_min'), (0.0038, 'vdot_s
td'), (0.0037, 'x_max'), (0.0036, 'w_mean'), (0.0036, 'vdot_max'), (0.0035, 'y_diff'), (0.0035, 'vdot_min'), (0.0034
, 'v_std'), (0.0033, 'theta_diff'), (0.0032, 'vy_std'), (0.0032, 'vy_mean'), (0.003, 'vdd_max'), (0.0029, 'y_max'),
(0.0026, 'delta_c_std'), (0.0025, 'c_max'), (0.0025, 'c_diff'), (0.0024, 'delta_c_diff'), (0.002, 'c_std'), (0.0019,
'w_std'), (0.0019, 'c_min'), (0.0017, 'delta_c_max'), (0.0016, 'x_diff'), (0.0015, 'vx_diff'), (0.0014, 'vy_max'), (
0.0014, 'vx_std'), (0.0014, 'vdot_diff'), (0.0013, 'vx_min'), (0.0013, 'theta_min'), (0.0012, 'x_min'), (0.0011, 'de
lta_c_min'), (0.0009, 'y_min'), (0.0008, 'vx_max'), (0.0008, 'v_max'), (0.0006, 'vy_min'), (0.0003, 'vy_diff'), (0.0
003, 'v_diff'), (0.0001, 'theta_max'), (0.0, 'v_min'), (0.0, 'label')]
Selected Features: x_mean vx_mean v_mean vdot_mean vdd_mean theta_mean theta_std c_mean w_min w_max w_diff t l criti
cal_count click_time pause_count total_pause_time pause_time_ratio ()
Index of selected features: [0, 10, 20, 25, 30, 35, 36, 40, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61]
Number of selected features: 18
```

Fig 2: Information gain of the features

CLASSIFIER: One Class Support Vector Machine (SVM)

- One-class SVM is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set.
- **Kernel used:** RBF
- **Optimal value of v found to be:** 0.2

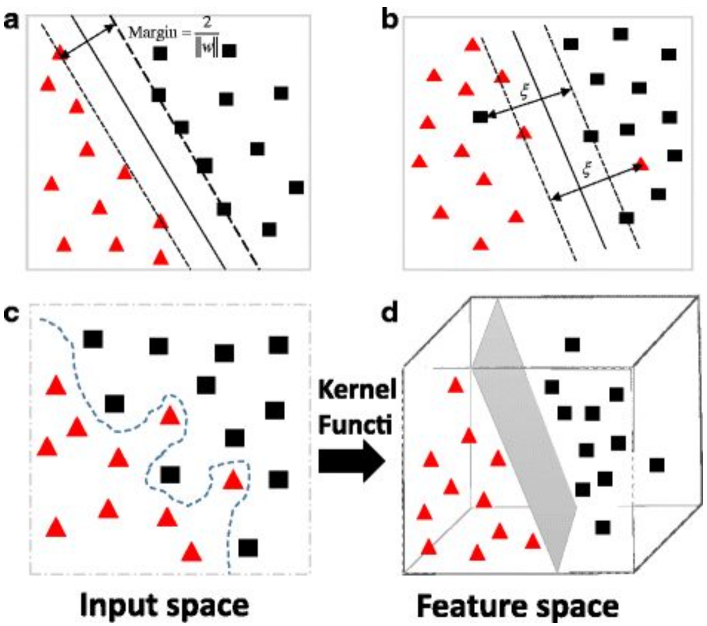


Fig 3: One Class SVM Hyperplane

The parameter v is the upper bound on the fraction of margin of errors and a lower bound on the fraction of support vectors relative to the total number of training examples. Small value of v causes overfitting and a high value causes underfitting

FACTORS AFFECTING PERFORMANCE

- Environmental conditions
  - Height of chair
  - Distance between mouse and body
  - Touchpad vs conventional mouse
- User conditions
  - Mood
  - Knowledge & practice of application
  - Typing errors
- GUI/mouse setting
  - Screen resolution
  - Pointer speed
- Noise
  - Hardware error
  - Software error

FUTURE IMPROVEMENTS

- Mood analysis
  - A separate label for mood could constitute a useful feature
- Data collection
  - More amount of data
  - Standard environment
  - Standard computer settings
- Data Preprocessing
  - Noise removal, smoothening and error nullification
- Training
  - State-of-the-art classification algorithms

RESULTS AND DISCUSSIONS

- ‘Rbf’ kernel was used and **v** was varied to obtain a maximum training accuracy of 94.64% and test set accuracy of 86.17%
- Precision of 0.1-1.2 and recall of 0.8-0.9 was achieved which implies that more true negative samples were correctly classified as compared to true positive samples which could be due to a myriad of factors involving the unequal number of data samples for every user and factors already discussed.

User-13

1. Training accuracy: 94.9475065617
   Test accuracy: 86.9121706986
   Precision: 0.1098546042
   Recall: 0.890052356021
   F1 Score: 0.195570894449
2. Training accuracy: 94.8196721311
   Test accuracy: 87.6930276088
   Precision: 0.112495845796
   Recall: 0.887287024902
   F1 Score: 0.199675564076
3. Training accuracy: 94.3606557377
   Test accuracy: 85.2316331306
   Precision: 0.107771802172
   Recall: 0.893356643357
   F1 Score: 0.192340265362
4. Training accuracy: 94.4262295082
   Test accuracy: 85.8072063641
   Precision: 0.106895208414
   Recall: 0.899672131148
   F1 Score: 0.191086350975
5. Training accuracy: 94.6885245902
   Test accuracy: 85.2363125877
   Precision: 0.104859020824
   Recall: 0.895592864638
   F1 Score: 0.187737145999

User-1

1. Training accuracy: 78.1124497992
   Test accuracy: 41.4629479022
   Precision: 0.0310827007591
   Recall: 0.779559118236
   F1 Score: 0.0597817734747
2. Training accuracy: 78.030807827
   Test accuracy: 41.8763754665
   Precision: 0.0309642226857
   Recall: 0.774322968907
   F1 Score: 0.0595472251147
3. Training accuracy: 77.8223783241
   Test accuracy: 40.9673715434
   Precision: 0.0312175648703
   Recall: 0.784615384615
   F1 Score: 0.0600460711543
4. Training accuracy: 78.273958856
   Test accuracy: 42.7279686154
   Precision: 0.031311468829
   Recall: 0.78273958856
   F1 Score: 0.0602142236804
5. Training accuracy: 78.4746613146
   Test accuracy: 42.632283992
   Precision: 0.0312525171164
   Recall: 0.778803693296
   F1 Score: 0.0600935476876

THANK YOU

Reference: Gamboa, H. and Fred, A.L., 2003, April. An Identity Authentication System Based On Human Computer Interaction Behaviour. In PRIS (pp. 46-55)