

INFERRING IDENTITY FROM USER BEHAVIOUR

Michael J. Carey[†], Graham D. Tattersall[§], Harvey Lloyd-Thomas*, and Martin J. Russell[‡]

[†]University of Bristol, [§]University of East Anglia, *Imagination Technologies, [‡]University of Birmingham,

In: Carey M.J, Tattersall G.D et al, *Inferring identity from user behaviour*, in IEE Proc. –Vis, Image

Signal Process., Vol. 150, No.6, 2003

© IEE 2003

ABSTRACT

Biometrics using inherited characteristics are frequently used in security systems. An alternative metric is user behaviour. Examples of user behaviour are the set of web pages accessed over a number of Internet sessions, or, the set of television programmes viewed by an individual over a number of days. In this paper we develop a mathematical framework which enables us to present simple and robust identification algorithms based on this kind of user behaviour.

We present experimental results based on a database of thirty three users' television viewing habits. The use of our algorithms in this domain is not intended as a real application, but rather as an illustration of the power of the techniques. Practical applications would include user authentication for fraud prevention and preference prediction on the Internet.

The user discrimination performance using the TV viewing database is modest, an equal error rate of about 18%. However, it is reasonable to suppose that the discrimination given by this technique would be orthogonal to those of other biometrics and so could provide a useful improvement in performance in a system combining the two.

1. INTRODUCTION

Biometrics [1] can take one of three forms, the intrinsic biometric which is determined by the user's genetic make-up, the extrinsic form which is based on the user's learnt behaviour and the hybrid form which is a combination of the first two. Examples of intrinsic biometrics would be DNA samples, fingerprints, or iris or retinal patterns. Intrinsic biometrics cannot be faked. Extrinsic biometrics include signatures and footfalls. Examples of hybrid biometrics would be voice characteristics used in speaker verification which comprise inherited features such as the size of the larynx and the length of the vocal tract together with extrinsic features such as speaking style language or accent.

While intrinsic biometrics are usually more accurate than the other forms it is not always possible or appropriate to use them. In this paper we examine the possibility of using an extrinsic biometric, the pattern of the user's behaviour in a particular domain, for authenticating the identity of the user and also for predicting the future pattern of that behaviour. We have previously suggested [2] the use of a user's calling patterns on a mobile phone as an extrinsic biometric to contribute towards the reduction in mobile phone fraud or theft. Other patterns of user behaviour might be patterns of information access on a computer system or credit card use. We do not suggest that these biometrics would offer a high level of security alone but that they provide information orthogonal to that of other biometrics and so could be combined with these to improve the overall performance.

Similarly when the user identity is known it is possible to make predictions about the user's future actions. This can be useful since it allows the system to anticipate the user's requirements and present the user with the required information while minimising user effort. An example of this would be the prediction of which new Web pages might interest a particular user. We plan to present this aspect of our work in a future publication.

The work reported in this paper uses data concerning television viewing preferences of a group of users. The users were asked to make daily reports of programmes they had watched the previous evening on any of the five UK terrestrial television channels. This choice of domain has several advantages, the choice presented to the users is consistent even though they were geographically diverse. Each user watched between one to two programmes per night, this led to significantly more data than could be acquired from other domains. This gave a total of over twelve thousand programmes watched by thirty-three users over a period of one year. It was relatively easy to collect the data using an email based form.

Using this data we examine the possibility of determining the identity of the user from the pattern of television programmes watched and predicting programmes of interest to the user from the next day's schedules.

We follow two approaches in our characterisation of user behaviour by the set of viewed programmes. In the first, we assume that the user behaviour is related to the content of the programmes and that the content can be characterised by a mixture of one or more topic classes which reflects the interest of a particular user. These topic classes were an augmented set based on the first version of the Metadata categories proposed by the TV Anytime Forum [3].

In the second approach, we again assume that the user behaviour is related to the content of the programmes, but we do not use on explicit topic classes. Instead, we use a word frequency description of the published programme listing as a characterisation of the content of the programme.

Both approaches would be applicable to characterisation of user behaviour on the Internet; the first would be used in situations where a Web page contains metadata relating to its topic. The second approach would be used where the no Web page metadata is available.

In this paper there are six more sections. In Section Two we develop a mathematical framework which enables us to present simple and robust identification algorithms for the situation in which the user behaviour is characterised by an interest in topics of explicitly defined category. Section Three describes a framework for the alternative case in which the topic categories are not known explicitly. The database collection exercise and the resultant data is described more fully in Section Four, and Sections Five and Six present our experimental results. We discuss these results in Section Seven.

2. INFERRING USER IDENTITY FROM EVENTS OF KNOWN CLASS

We start our theoretical treatment of inferring the identity of a user from a behavioural pattern by considering the case in which there a set of events, $E_j = e_1 \dots e_j \dots e_j$ is observed. In the context of the television viewing domain E_j is the set of programmes watched by a particular user, u_i over a period of time. In this section we will develop a theory in which it is assumed that each event is statistically associated with one or more of a set of K topic classes, c_k .

We wish to know the identity of the user. Using Bayes Theorem:

$$p(u_i | E_J) = \frac{p(u_i)p(E_J | u_i)}{p(E_J)} \quad \dots 1$$

If the events were independent then,

$$p(E_J | u_i) = \prod_j p(e_j | u_i) \quad \dots 2$$

In some applications the nature of the set of events initiated by a user may not be completely independent. For example, the Web pages accessed in a single session by an individual might have similar content. In this case, the evidence provided about user identity by each event (Web page accessed) would be similar. In the extreme, where each event in the set is fully correlated, the probability of the set of events given the user would be equal to the probability of any individual event given the user. However, in our analysis we will assume that the events are independent. This is arguably the case in applications such as TV viewing and Web page access that takes place in separate sessions spread over hours or days. In this case we have,

$$p(u_i | E_J) = \frac{p(u_i) \cdot \prod_j p(e_j | u_i)}{\prod_j p(e_j)} \quad \dots 3$$

These events have been initiated by one of an equiprobable set of I users $u_1 \dots u_i \dots u_I$

$$p(u_i | E_J) = \frac{1}{p(u_i)^{J-1}} \prod_j \left(\frac{p(e_j | u_i)p(u_i)}{p(e_j)} \right) \quad \dots 4$$

hence

$$p(u_i | E_J) = \frac{1}{p(u_i)^{J-1}} \prod_j p(u_i | e_j) \quad \dots 5$$

Equation 5 suggests that to fulfill our goal of identifying the most likely user given a set of observed events, E_J , we require estimates of $p(u_i | e_j)$. However, there is no practical method of estimating this conditional probability because each event is unique. Our solution is to assume that an event is categorized by one or more topic classes, c_k , with probability $p(c_k | e_j)$. This type of categorisation is often already attached to events. Analysis of the frequency with

which a particular user is linked to events of each class can then be used to estimate the probability of the user given each class, $p(c_k | u_i)$. Then by Bayes theorem,

$$p(u_i | c_k) = \frac{p(c_k | u_i)p(u_i)}{p(c_k)} \quad \dots 6$$

The probability of the user given the event can then be expressed as the sum over all classes of the product of the probability of the class given the event and the user given the class.

$$p(u_i | e_j) = \sum_K p(u_i | c_k)p(c_k | e_j) \quad \dots 7$$

Again assuming independence of the events within the event set, we have,

$$p(u_i | E_J) = \frac{1}{p(u_i)^{J-1}} \prod_J \sum_K p(u_i | c_k)p(c_k | e_j) \quad \dots 8$$

Assuming that the training data provides an estimate of $p(c_k | u_i)$, then by Bayes theorem,

$$p(u_i | c_k) = \frac{p(c_k | u_i)p(u_i)}{p(c_k)} \quad \dots 9$$

therefore,

$$p(u_i | E_J) = p(u_i) \prod_J \sum_K \frac{p(c_k | u_i)}{p(c_k)} p(c_k | e_j) \quad \dots 10$$

The index of the most probable user is then given by

$$\hat{i} = \arg \max_i p(u_i | E_J) = \arg \max_i \left(p(u_i) \prod_J \sum_K \frac{p(c_k | u_i)}{p(c_k)} p(c_k | e_j) \right) \quad \dots 11$$

If the event can be uniquely assigned to a single class that is $p(c_l | e_j) = 1$ for $l = k$ and $p(c_l | e_j) = 0$ otherwise, then

$p(u_i | e_j) = p(u_i | c_k)$ and defining $\alpha(j)$ to be the index of the class to which event j is assigned.

$$p(u_j | E_J) = p(u_i) \prod_J \frac{p(c_{\alpha(j)} | u_i)}{p(c_{\alpha(j)})} \quad \dots 12$$

In the log domain we have,

$$\log(u_j | E_J) = \log p(u_i) + \sum_J (\log p(c_{\alpha(j)} | u_i) - \log p(c_{\alpha(j)})) \quad \dots 13$$

Since the users are equiprobable,

$$\hat{i} = \arg \max_i \sum_j (\log p(c_{\alpha(j)} | u_i) - \log p(c_{\alpha(j)})) \quad \dots 14$$

We see that if $p(c_{\alpha(j)} | u_i) > p(c_{\alpha(j)})$, the event is positive evidence of the user, while negative evidence is given by $p(c_{\alpha(j)} | u_i) < p(c_{\alpha(j)})$. If $p(c_{\alpha(j)} | u_i) = p(c_{\alpha(j)})$ there is no evidence for or against the user since the sequence could be generated by randomly selecting events from the classes.

To evaluate the above we need estimates for $p(c_{\alpha(j)})$ and $p(c_{\alpha(j)} | u_i)$. In the TV viewing case these correspond to the classes of programmes and the probability that a user watched a programme of a particular class during the training period,

$$\hat{p}(c_{\alpha(j)}) = \frac{n_k}{\sum_K n_k} \quad \dots 15$$

and

$$\hat{p}(c_{\alpha(j)} | u_i) = \frac{n_{k,i}}{\sum_K n_{k,i}} \quad \dots 16$$

where n_k is the number of programmes in the schedules assigned to class k and $n_{k,i}$ is the number of programmes in the schedules assigned to class k that user i watched. An alternative evaluation of $\hat{p}(c_{\alpha(j)})$ that takes into account the viewing habits of all the users, not just the schedules is given by,

$$\hat{p}(c_{\alpha(j)}) = \frac{\sum_I n_{k,i}}{\sum_K \sum_I n_{k,i}} \quad \dots 17$$

Yet another approach would be to use a neural network algorithms or similar method to generate estimates of multiple class assignment probabilities $\hat{p}(c_k | e_j)$ and Equation (11) can then be used.

3. INFERRING USER IDENTITY FROM A SET OF EXAMINED TEXTS

In this section, we extend the analysis to show how the identity of a user might be inferred from a set of events,

$E_j = e_1 \dots e_j \dots e_J$, that are not associated with any explicit topic classes, c_k . Instead, we use the set of associated words as a characterisation of the event. In the context of the TV viewing experiments, these are the words in the listing for each of the TV programmes watched by a user. In an Internet application, the words would be taken from a Web page viewed by the user.

Let the event, e_j , be characterised by a word frequency vector, $W = w_1 \dots w_N$ where N is the size of the system vocabulary and w_k is a normalised count of the number of occurrences of the k^{th} vocabulary word in the text associated with an event.

The set of events, E_j , can be written in terms of the set of word frequency vectors $E_j = W_1 \dots W_j \dots W_J$, and developing the analysis in a similar manner to Equations 1 and 3 in Section Two, we have,

$$p(u_i | E_j) = \frac{p(u_i) \prod_j p(W_j | u_i)}{\prod_j p(W_j)} \quad \dots 18$$

Assuming equiprobable users:

$$p(u_i | E_j) = \frac{1}{p(u_i)^{J-1}} \prod_j \left(\frac{p(W_j | u_i) p(u_i)}{p(W_j)} \right) \quad \dots 19$$

To evaluate Equation 19 we would need to estimate the distributions $p(W | u_i)$ and $p(W)$. Unlike the categorical distributions of Equations 1 and 3, these distributions are continuous because the word frequency vector, W , is a random variable. However, the word frequency characterisation of the event is likely to be very noisy given the shortness of typical texts. As a consequence, the estimated distributions would be poor. A possible remedy is to use a method such as Latent Semantic Analysis (LSA) [7], to reduce the dimensionality of the word frequency representation of the event, and thereby remove some of the noise caused by sparseness of the text. However, LSA relies on the use of a linear transform method for dimensionality reduction and on strong assumptions such as a Gaussian model for the distributions. We therefore propose a more practical approach: directly use a linear transform, T_U , in a linear classifier to map the set of word frequency vectors, $E_j = W_1 \dots W_j \dots W_J$ to set of scores for each of the system users. This approach can provide the inherent noise removal

of LSA, and if trained under a minimum mean square error criterion, generates scores that are proportional to the required user probabilities $p(u_i | E_j)$ as shown by Gish [8].

We have examined two variants of this approach. In the first, each word frequency vector, W_j , associated with an event is taken separately from the set, $E_j = W_1 \dots W_j \dots W_J$, and applied to the classifier. The score generated for the i^{th} user is y_{ij} .

$$y_{ij} = T_U * W_j \quad \dots 20$$

An overall score, Y_i , for the i^{th} user given the complete set, E_j , is then found as the product of the individual scores. This is based on the assumption that each score is proportional to the user probability given the word frequency vector, and that each of the vectors is independent.

$$Y_i = \prod_J y_{ij} \quad \dots 21$$

Our second approach is to classify the mean word frequency vector, \bar{W} of the set and generate a single set of user scores,.

$$\bar{W} = \frac{1}{J} \sum_J W_j \quad \dots 22$$

$$Y_i = T_U * \bar{W} \quad \dots 23$$

In practical experiments we have found the mean vector approach to be far more reliable than the former approach and our experimental results will be restricted to this technique.

Equation 23 defines the overall process of identifying a user from the texts of the listings of watched TV programmes. However, there are detailed variations to the form and way in which the classification transform is generated. All the transforms are linear and are learnt using either gradient descent minimisation of a mean square error performance measure, or by singular value decomposition of a correlation matrix. The types of transform are as follows.

Single layer transform: A transform learnt by gradient descent minimisation of mean square error between actual and desired classifier outputs over the training set examples. The transfer function of the classifier is:

$$\begin{bmatrix} y_1 \\ \cdot \\ y_M \end{bmatrix} = \begin{bmatrix} t_{11} & \cdot & t_{1N} \\ \cdot & \cdot & \cdot \\ t_{M1} & \cdot & t_{MN} \end{bmatrix} * \begin{bmatrix} \bar{w}_1 \\ \cdot \\ \bar{w}_N \end{bmatrix} \quad \dots 24$$

Where the ‘y’ vector is the set of scores for each of the M users, the ‘w’ vector is a word frequency vector formed from the set of watched TV programs, and the matrix ‘t’ is the classification transform.

SVD Derived Transform: The training set matrix is processed to form a matrix of the same size but containing true word-user correlation coefficients. The correlation matrix is decomposed into a pair of transforms, alpha and beta, by singular value decomposition (SVD). A m-dimensional bottleneck is imposed on these transforms such that:

$$\begin{bmatrix} y_1 \\ \cdot \\ y_M \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{1m} \\ \alpha_{M1} & \alpha_{Mm} \end{bmatrix} * \begin{bmatrix} \beta_{11} & \beta_{1N} \\ \beta_{m1} & \beta_{mN} \end{bmatrix} * \begin{bmatrix} \bar{w}_1 \\ \cdot \\ \bar{w}_N \end{bmatrix} \quad \dots 25$$

In practice, alpha and beta are then recombined into a single transform that is algebraically equivalent to Equation 25.

4. DATABASE

The database used in the experiments was collected from the programmes watched on the UK terrestrial television channels (BBC1, BBC2, ITV, C4 and Channel 5) over the period Sep 2000 to Aug 2001. This data was partitioned into a 9 month training set and a 3 month test set..

Each of the programme listings in the schedules was annotated with three levels of class label. These classes were an augmented set based on the first version of the Metadata categories proposed by the TV Anytime Forum[3].

Level I	Level II	Level III
INFORMATION	Humanities	History Literature
DRAMA	Docudrama	Science
ENTERTAINMENT	Shows	Game Show
SPORT	Water Sports	Sub-Aqua
LEISURE	Holidays	Walking
ARTS & MEDIA	Media	Press
DOCUMENTARY	Science	Engineering
MUSIC	Jazz	Traditional
ENRICHMENT	Education	Languages
MOVIES & ANIMATIONS	Animations	Comic cartoons

Table 1 Classes by Level

Each of the levels has differing degrees of detail between the categories. Examples are shown in Table 1. Entries for Levels II and Levels III are examples of classes contained in Levels I and II respectively, e.g. Literature is one of the sub-classes of Humanities which is one of the sub-classes of Information. In addition we added a fourth level of classification which was defined as the title of the programme. The distribution of these classifications is shown in Table 2. During this period the thirty-three users who participated reported watching a total of 12,904 programmes, typically a single programme per night. In this they were atypical of the population as a whole who watch an average of 3h 41m television per day.

Class	Train Set	Test Set	Both Sets
Level I	10	10	10
Level II	70	57	56
Level III	206	140	134
Title	2580	922	276
Programmes	10861	3180	14041

Table 2 Database Summary

Table 3 shows how data from the 3 month test set was partitioned for comparative experiments. The same data was divided into 92 trials of duration one day, 13 trials of seven days and 6 trials of fourteen days as shown in the second row of the table. However, the programmes watched by each user are not evenly distributed across the test periods, and some users did not watch any programmes during the shortest periods of one day. As a result, for a trial period of one day and an average user, only 30 out of the possible 92 trials contained any events. Lesser reductions are seen for longer trial periods as shown in the third row of Table 3.

Test Period	1 day	7 days	14 days
Total Trials	92	13	6
Average populated Trials	30	10	5

Table 3 Test Set

5. USER IDENTIFICATION EXPERIMENTS USING EVENTS OF KNOWN CLASS

The frequencies of the classes in the training data were used in Equation (15) to estimate $p(c_{\alpha(j)})$, and $p(c_{\alpha(j)} | u_i)$ was estimated using Equations (16) and (17). The estimate based solely on the schedules (Equation 16) is the *schedule* estimate while the estimate using the sum of the user frequencies (Equation 17) is the *world* estimate. A Turing-Good correction was made for small probabilities [4].

The three test data sets detailed in Table 3 were used in three separate experiments. Each set was used as a true user trial for the target user and as an impostor trial for the other users. Equation (3) was used to score the trials. The scores from Equation (3) for different users were normalised using $znorm$ as proposed by Reynolds, see [5]. That is all other user training sets were scored against each user model. The mean, m_{train} , and standard deviation, d_{train} , of these scores were used to normalise the scores from the test trials using,

$$\hat{s}_{test} = \frac{s_{test} - m_{train}}{d_{train}} \quad \dots 26$$

The results were then used to generate detector error trade-off (DET) plots [6], as illustrated in Figures 1 and 2. Figure 1 shows that the best discrimination between users was given using the most detailed classes, Level III, which performed better than the less detailed classes. However continuing this process to the Title Class causes the performance to degrade. Table 2 shows that while nearly all of the Classes of Levels I, II, and III occurring in the test data have been seen in the training data less than half the Titles in the test data occurred in the training data.

Not surprisingly Figure 2 shows that increasing the number of events in each set improved performance. It also shows that normalising using the world estimate gives better results than normalising by the schedule estimate. This result supports our assertion that the weight given to the evidence should be proportional to the deviation of the user's behaviour from that of other users, as shown in Equation (17).

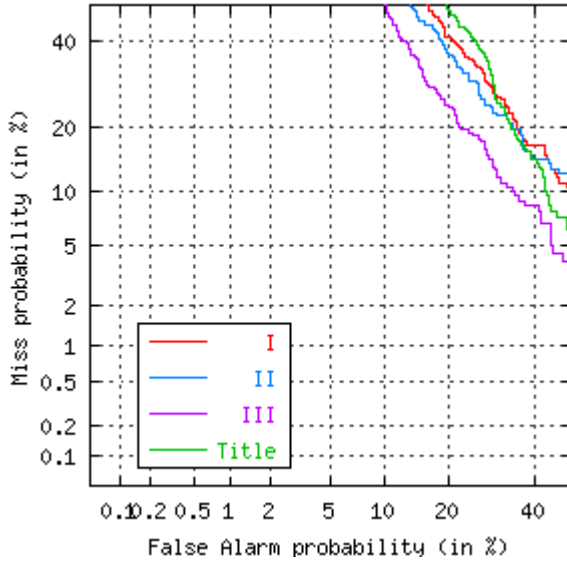


Figure 1. User Recognition Performance for each of the Four Classes, using 14 Day Sets normalised by the World Estimate.

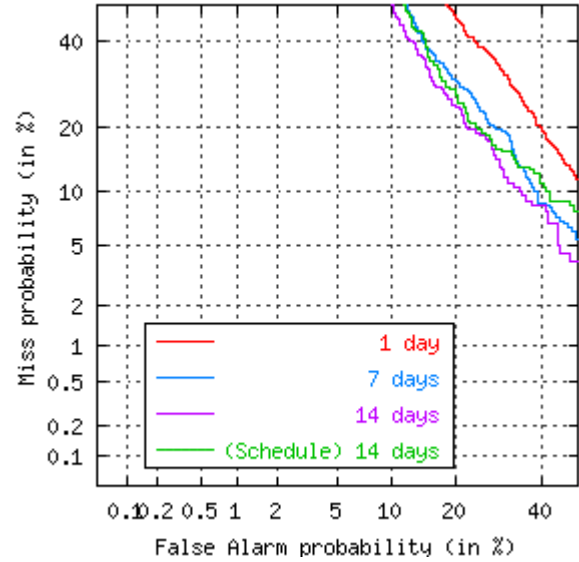


Figure 2. User Recognition Performance using 1, 7, and 14 Day sets and Level III Classes, normalised by the World Estimate. Also shown is the 14 Day Set Performance normalised by the Schedule Estimate.

6. USER IDENTIFICATION EXPERIMENTS USING EXAMINED TEXTS

Two sets of experiments were performed to test how well we could infer user identity from the texts of the listings of watched programmes. These were both based on a linear transform classification of the mean, normalised, word frequency vector of the set of listings, but differed in the way in which the transform was learnt.

The training and test sets used in the experiments on user identification using examined texts, are identical to the data used in the experiments based on events of known class described in Section 5. That is, there are 33 users and the listings of programs watched by each user are compiled as training set covering 9 months, and the test set covers consecutive 14 day periods in a separate three month period immediately following the training set period.

For the purpose of these experiments, the training set is formulated as a matrix in which columns denote a particular user, and rows correspond to frequency of usage of each word in the system vocabulary. The word frequencies are *averaged* over the set of programs watched by the user over a training period. The test set consists of mean word frequency vectors for each user. Each mean word frequency vector is derived from the listings watched by the user over each of the 14 day periods.

Our first experiments used a linear, single layer transform derived by gradient descent minimisation of the squared error between the actual user score vector and the desired score vector for each training set input. The desired score vector was coded as '1' for the desired user and '0' for all other users. The mean square error converged rapidly to a very low value under gradient descent learning, and provides the test set results in the DET curve shown in Figure 3 with an equal error rate of approximately 30%.

The second set of experiments were done using transforms generated by SVD of the training set user-word frequency correlation coefficients. This is equivalent to a double layer transform with bottleneck. The results for bottleneck dimensions of 10, 20, and 33 are shown in Figure 4. The results show that any form of bottleneck reduces the classification performance below the best performance of 18% equal error rate which is obtained when the "bottleneck" dimension is 33, showing that the SVD process has no effect. This suggests that the same performance could be obtained using the correlation matrix, subject to suitable normalisation, directly as a transform. A comparison of the results obtained using this method, and the unreduced SVD transform is shown in Figure 5.

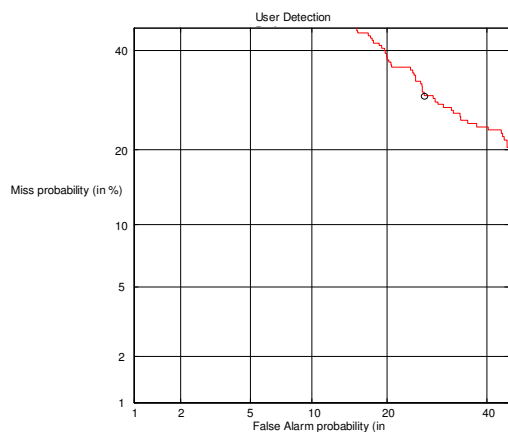


Fig. 3 User Recognition DET Curve Using Single Layer User-Discriminative Transform

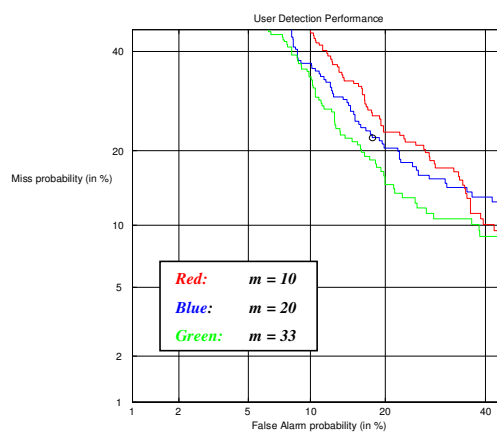


Fig. 4 User Recognition using SVD derived transforms

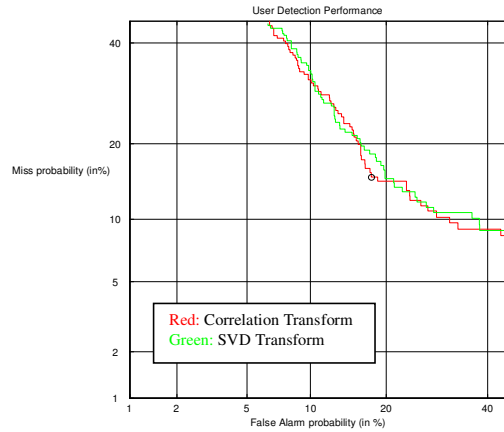


Fig. 5 Comparison of user recognition performance using correlation transform and SVD transform with no dimensionality reduction

6. DISCUSSION

In this paper we have shown that the initiation of a small set of events by a user can help discriminate between that user and other users. Two approaches have been explored. In the first, we assume that the user behaviour is related to the content of the programmes and that the content can be characterised by a mixture of one or more topic classes which reflects the interest of a particular user. In the second approach, we use a word frequency description of the text that is associated with the event as a characterisation.

We observe equal error rates of about 20% when using events with explicit class, and about 18% when using the associated texts without explicit class information. The better result obtained using the text associated with an event, rather than the explicit class description of the event is rather surprising. It is may because the explicit classes do not fully describe the range of programme content and because the class annotation of the training set was done by a single person, who was forced to make subjective judgements of the nature of each of the TV programmes in the database.

A single layer transform using gradient descent learning gives very poor performance on the test set (equal error rate=30%), presumably because the classifier is “over-trained”. By contrast, much better performance is obtained using non-discriminative learning. That is, SVD or the correlation matrix transform. The results obtained using the word-user correlation matrix transform (equal error rate =18%) is almost the same as using SVD with subspace size of 33. Reducing

the subspace dimension degrades performance, indicating, as expected, that users cannot be separated in a space with fewer dimensions than users.

All three sets of performance figures are rather poor when compared with intrinsic biometrics [1]. However, it is reasonable to suppose that the scores given by this technique would be orthogonal to those of other biometrics and so could provide a useful improvement in performance in a system combining the two.

11. REFERENCES

- [1] Wayman, J, 'Theory and Characterisation and Testing of General Biometric Technologies', Proc. Speaker Recognition Workshop, Crete 2001.
- [2] Carey, M, and Auckenthaler, R, 'User Validation for Mobile Telephones', Proc. ICASSP2000, Istanbul, Vol II, pp1093-1096.
- [3] Mornington, A Van Smirren, D and Montie, E 'On: Content Referencing Requirements', Requirements Series: R-4 , TV-Anytime Forum, 2002, <http://www.tv-anytime.org>.
- [4] Nadas, A, 1985. 'On Turing's formula for word probabilities', IEEE Transactions on Acoustics, Speech, and Signal Processing' ASSP-33 (6), pp. 1414--1416.
- [5] Auckenthaler, R., Carey, M., and Lloyd-Thomas, H, 'Score normalization for text-independent speaker verification systems', .Digital Signal Processing, 10: pp. 42-54
- [6] Martin, A, Doddington, G, Kamm, T, Ordowski, M, and Przybocki.,M, 'The det curve in assessment of detection task performance', Proc. of Eurospeech 97, Vol 4, pp. 1895--1898, 1997.
- [7] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). *Introduction to Latent Semantic Analysis. Discourse Processes*, **25**, 259-284.
- [8] Gish H, 'A probabilistic approach to the understanding and training of neural network classifiers', Proc. ICASSP 1990, pp1361 -1364