World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:12, 2008

Similarity Measure Functions for Strategy-Based Biometrics

Roman V. Yampolskiy, and Venu Govindaraju

***Abstract*—**Functioning of a biometric system in large part depends on the performance of the similarity measure function. Frequently a generalized similarity distance measure function such as Euclidian distance or Mahalanobis distance is applied to the task of matching biometric feature vectors. However, often accuracy of a biometric system can be greatly improved by designing a customized matching algorithm optimized for a particular biometric application. In this paper we propose a tailored similarity measure function for behavioral biometric systems based on the expert knowledge of the feature level data in the domain. We compare performance of a proposed matching algorithm to that of other well known similarity distance functions and demonstrate its superiority with respect to the chosen domain.

***Keywords*—**Behavioral Biometrics, Euclidian Distance, Matching, Similarity Measure.

I. Introduction

BIOMETRIC systems are becoming a standard methodology for the enforcement of security of computer systems, networks and work spaces. Biometric recognition is a subset of the general pattern recognition problem, and follows a similar algorithm. First the data collection takes place, followed by the feature extraction step. Next, a similarity measure function is applied to determine the closest pattern in the database to the one just collected; finally a decision is made as to the similarity of the two profiles being compared [4].

Functioning of a biometric system in large part depends on the performance of the similarity measure function. Frequently a generalized similarity distance measure function such as Euclidian distance or Mahalanobis distance is applied to the task of matching biometric feature vectors [7]. However, often accuracy of a biometric system can be greatly improved by designing a customized matching algorithm optimized for a particular biometric application such as fingerprint recognition [17, 10, 3, 12], signature verification [8] or speaker recognition [9, 6]. Design of a customized well performing matching algorithm is a complicated task which involves taking into account noise attributes of the collected data, expert knowledge about the features and their statistical distributions as well as time-efficiency of a proposed algorithm on large scale databases [5].

In this paper we propose a novel similarity measure function for strategy-based behavioral biometric systems. We compare performance of a proposed matching algorithm to that of other well known similarity distance functions with respect to strategy based behavioral biometrics to demonstrate its superiority with respect to the chosen domain [16]. We begin with an overview of strategy based behavioral biometrics. This is followed by a survey of the most popular similarity measure functions used in biometric applications. Finally, we present our similarity measure functions and describe experiments we performed in order to establish the best performing similarity distance function.

Manuscript received September 24, 2006. This work was supported in part by the by National Science Foundation Grant No. DGE 0333417 "Integrative Geographic Information Science Traineeship Program", awarded to the University at Buffalo.

R. V. Yampolskiy is with the Center for Unified Biometric and Censors and IGERT in GIS, Buffalo NY, USA (e-mail: rvy@buffalo.edu).

Venu Govindaraju is with the Center for Unified Biometric and Censors and department of Computer Science and Engineering University at Buffalo, Buffalo NY, USA (e-mail: govind@buffalo.edu).

II. Strategy-Based Biometrics

Strategy-based biometrics is a sub-type of behavioral biometrics. Behavioral biometrics provides a number of advantages over traditional biometric technologies. They can be collected non-obtrusively or even without the knowledge of the user. Collection of behavioral data usually does not require any special hardware and is so very cost effective. While behavioral biometrics is not unique enough to provide reliable human identification they have been shown to provide high accuracy identity verification.

Yampolskiy et al. [15, 16] proposed a system for verification of online poker players based on a behavioral profile which represents a statistical model of player's strategy. The profile consists of frequency measures indicating range of cards acted on by the player. It also measures how aggressive the player is via such variables as percentages of re-raised hands. The profile is actually human readable meaning that a poker expert can analyze and understand strategy employed by the player from observing his or her behavioral profile [11]. For example just by knowing the percentage of hands a particular player chooses to play it is possible to determine which cards are being played with high degree of accuracy. Table I demonstrates a sample profile for a player named Bob.

TABLE I
BASIC STRATEGY PROFILE

Player Name:	Bob
Action	**Frequency**
Folded	77%
Checked	55%
Called	33%
Raised	6%
Check-Raised	4%
Re-Raised	2%
All-In	37%

A combination of such statistical variables taken together produces a feature vector which is used by a pattern recognition algorithm to determine if a current profile is consistent with that previously seen one from this particular player or if a possible intruder has taken the control of the account. In the Table I we see a 7 dimensional feature vector, explanation for the meaning of the variables in our feature vector follows [16].

- **folded** Percentage of times this particular player has decided to give up his claims to the pot
- **checked** Percentage of times this particular player has decided to check
- **called** Percentage of times this particular player has paid an amount equivalent to the raise by some other player ahead in position
- **raised** Percentage of times this particular player has chosen to raise
- **check-raised** percentage of times a player has checked allowing another player to put some money into the pot, just to come over the top and raise the pot after the action gets back to him
- **re-raised** Percentage of times this particular player has chosen to re-raise somebody-else's raise. This would include a re-re-raise and re-re-re-raise and so on
- **all-in** Percentage of times this particular player has chosen to invest all his money in the hand

The complete system for player verification works as follows: First a player profile is generated either by data mining an existing database of poker hands or by observing a live game of poker. Next a similarity measure is obtain between the feature vector generated based on the recently collected player data and the data for the same player obtained in previous sessions. A score is generated indicating how similar the current style of play is to the historically shown style of play for a particular player. If a score is above a certain threshold, it might indicate that a different user from the one who has originally registered is using the account and so the administrator of the casino needs to be alerted to that fact. If the score is below some threshold, the system continues collecting and analyzing the player data [16].

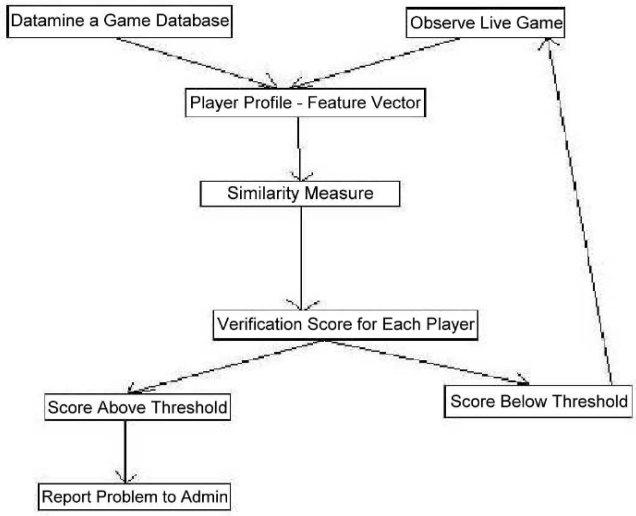

Fig. 1 The architecture of the strategy based behavioral biometric system [16]

Descriptive accuracy of a behavioral profile can be greatly increased if additional information is included. In their published reports Yampolskiy et al. [15, 16] utilize a profile structure which separates player's actions into the four stages of the hand, making temporal information available and as a result description of player's strategy more meaningful. Table III is an example of such temporal profile.

TABLE II
TEMPORAL STRATEGY PROFILE [16]

Player Name: Bob				Hands Dealt: 224
	Pre-Flop	Flop	Turn	River
# of Hands Played	224	68	46	33
Folded	67%	28%	24%	18%
Checked	7%	54%	52%	52%
Called	21%	32%	28%	33%
Raised	4%	1%	4%	6%
Check-Raised	0%	4%	0%	0%
Re-Raised	0%	1%	0%	0%
All-In	1%	3%	4%	39%

Profiles can be further enhanced with the inclusion of spatial information, essentially making a separate profile for each of the ten positions a player can have around the table. Such profiles clearly demonstrate dependence of player's strategy on position. Table III demonstrates such a multi-dimensional profile based on relative position of players.

TABLE III
SPATIAL STRATEGY PROFILE

Action	Small Blind	Big Blind	Under the Gun	4th Seat	5th Seat	6th Seat	7th Seat	8th Seat	9th Seat	Dealer
Folded	77%	73%	71%	69%	67%	64%	61%	59%	57%	51%
Checked	55%	53%	50%	49%	48%	44%	41%	39%	37%	34%
Called	14%	16%	19%	22%	26%	29%	33%	37%	43%	53%
Raised	2%	3%	4%	6%	8%	11%	13%	15%	17%	20%
Check-Raised	31%	28%	23%	19%	17%	15%	12%	9%	6%	4%
Re-Raised	0%	1%	2%	4%	6%	10%	14%	18%	25%	30%
All-In	37%	39%	41%	43%	47%	51%	55%	59%	62%	65%

Finally with the addition of contextual information about the cards revealed at the flop divided into 7 flop types described in the poker literature [1] we have a 3D information space, which for every stage of the game, every position and every flop provides frequency counts of player's actions. Dimensionality of such a profile could be extremely high, compared to the basic profiles.

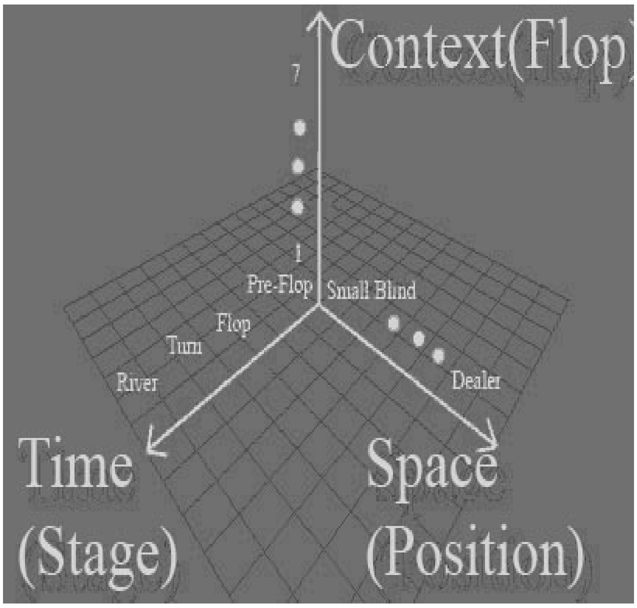

Fig. 2 3D profile structure with temporal, spatial and context axis

Table IV summarizes different possible profile types which can be used with strategy based behavioral biometrics along with the information they include and lists the profile's dimensionality. Ideally any similarity measure function we propose to utilize should be flexible enough to handle any of the presented profile types.

TABLE III
PROFILE TYPES BY INFORMATION INCLUDED AND VECTOR DIMENSIONALITY

Profile Type	Information Included	Profile Dimensionality
Basic	Frequency counts for actions	7
Temporal	Frequency counts for actions at different stages of the game	7 x 4 = 28
Contextual	Frequency counts for actions with respect to the flop type	7 x 7 = 49
Spatial	Frequency counts for actions at different positions around the table	7 x 10 = 70
Temporal-Spatial	Frequency counts for actions with respect to the stage of the game and relative position around the table	7 x 10 x 4 = 280
Temporal-Contextual-Spatial	Frequency counts for actions with respect to the stage of the game and relative position around the table and the flop type (post flop action only)	7 x 10 x 4 + 3 x 7 x 7 =427

As the amount of contextual information increases so does the dimensionality of the behavioral profile. This results in what is known as the "curse of dimensionality". The matching algorithm needs a large number of feature measurements to account for all the different possibilities of potential situations. The complexity of a high-dimensional space increases exponentially with the number of features. This large collection of features forms a high-dimensional space, in which it is very difficult to find the best decision boundary [2].

III. SIMILARITY MEASURE FUNCTIONS

Then a new biometric data sample is presented to a security system it is necessary to measure how closely it resembles template data. A good similarity measure takes into account statistical characteristics of the data distribution assuming enough data is available to determine such properties [7]. Alternatively expert knowledge about the data can be used to optimize a similarity measure function, for example a weighted Euclidian distance function can be developed if it is known that certain features are more valuable then others. The distance score has to be very small for two feature vectors belonging to the same individual and therefore representing a similar strategy. At the same time it needs to be as large as possible for feature vectors coming from different individuals, as it should represent two distinct playing strategies [16].

Lee et al. [7] describe the following method for making a similarity measure based on the statistical properties of the data: data is represented as a random variable $x=(x_1,\ldots,x_D)$ with dimensionality D. The data set $X=\{x_n|n=1,\ldots,N\}$ can be decomposed into sub-sets $X_k = \{x_{nk}|n_k = 1,\ldots, N_k\}$ $(k=1,\ldots,K)$, where each sub-set X_k is made up of data from the class C_k corresponding to an individual k. For identification the statistical properties of data X_{nk} are usually considered, which can be represented by a probability density function $p_k(x)$. If we have $p_k(x)$ for each k, for given data x, we calculate $f(p_k(x))$, where f is a monotonic function and find a class C_k maximizing $p_k(x)$. The similarity measure between a new data item and the center of mean μ_k of class C_k is given by the Euclidean distance. If we also estimate the covariance matrix Σ_k for $p_k(x)$, then the similarity measure defined as $-\log p_k(x)$ is the Mahalanobis distance [7].

A. Euclidian Distance

One of the most popular similarity distance functions is the Euclidian distance. It is just the sum of the squared distances of two vector values (x_i,y_i) [13]:

$$d_E = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \tag{1}$$

Euclidian distance is variant to both adding and multiplying all elements of a vector by a constant factor. It is also variant to the dimensionality of the vectors, for example if missing values reduce the dimension of certain vectors produced output will change. In general the value of Euclidian similarity measure may fall in the range from zero indicating a perfect match to sqrt(n) (where n-dimensional vector is used) indicating maximum dissimilarity of playing styles. Obviously both of those extreme cases don't occur in real life and represent only theoretical possibilities not related to any viable playing style. In experiments with real life data Euclidian Similarity measure is always in between the two extremes [16].

B. Mahalanobis Distance

Mahalanobis distance is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2}$$

with mean $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_p)$ and covariance matrix Σ for a multivariate vector $x = (x_1, x_2, x_3, \ldots, x_p)$. Mahalanobis distance can also be defined as dissimilarity measure between two random vectors $\vec{x}$ and $\vec{y}$ of the same distribution with the covariance matrix Σ:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T\Sigma^{-1}(\vec{x}-\vec{y})} \tag{3}$$

If the covariance matrix is the identity matrix then it is the same as Euclidean distance. If the covariance matrix is diagonal, then it is called normalized Euclidean distance:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{p}\frac{(x_i-y_i)^2}{\sigma_i^2}} \tag{4}$$

where σ_i is the standard deviation of the x_i over the sample set. Mahalanobis distance is not dependent on the scale of measurements [14].

C. Manhattan Distance

The Manhattan distance between two points, in a Euclidean space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes. In other terms, Manhattan distance is the sum of the absolute differences of the two vector values (x_i,y_i) [13].

$$d_M = \sum_{i=1}^{n}\left|x_i - y_i\right| \tag{5}$$

D. Weighted Euclidean Distance

Performance of the Euclidian similarity measure function can be greatly improved if an expert knowledge about the nature of the data is available. If it is known that some values in the feature vector hold more discriminatory information with respect to others, it is possible to assign proportionally higher weights to such vector components and as a result influence the final outcome of the similarity function.

In the case of the poker domain, it is believed by the experts in the field, that the style of the poker player is particularly evident in the pre-flop card selection. Before the flop cards are revealed the player has relatively little information to analyze and often acts based on a small set of rules, which dictate how hands should be played based on the hand itself, position of the player and betting action so far observed. Application of such rules is relatively long-term consistent by most players and so has higher discrimination value as compared to action at the later rounds in the game. In such later rounds additional information about communal cards and opponent reading skills become more important than pre-established rules and so are more situation dependent.

IV. Experiments

A. Data

Experiments were conducted with a 100 authentic user profiles and a 100 imposter profiles used in each. Three different experiments were conducted in each one a different type of behavioral profile representation was used. Specifically a 28-dimensional temporal profile, a 280-dimensional temporal-spatial profile and a 427-dimensional temporal-spatial-contextual profile were chosen as this allowed us to observe the influence of increasing the amount of environmental information available to the security system on systems performance. We also had an opportunity to observe the effect of the curse of dimensionality with respect to the performance of our similarity measure functions.

For each similarity function a continuously varying threshold curve was generated demonstrating the relationship between False Accept Rate (FAR) and a False Reject Rate (FRR). Changing threshold trades the FAR off against the FRR, so the error rates can be adjusted according to the requirements of the security application [7]. For our experiments the value of the threshold which makes FRR equal to FAR was selected for each similarity measure function and is used as the representative accuracy of the utilized similarity measure function.

B. User Verification

In this paper we compared three general similarity measure functions (Euclidian, Mahalanobis, Manhattan) with a domain specific function developed by us (Weighted Euclidian). The Weighted Euclidian distance measure we have utilized in our experiments assigns a weight of 3 to all pre-flop features of the vector and weight of 1 to all other features. The weight of 3 has been experimentally established by trial and error of different weights in the range from 1 to 10. The weight is incorporated into the formula by dividing the difference between corresponding values in the two feature vectors by the selected weight.

As can be seen from Table V general similarity measure functions (Euclidian, Mahalanobis and Manhattan) showed a very similar performance, with Mahalanobis distance being slightly inferior to Euclidian and Manhattan distances which showed identical performance of 12% Equal Error Rate (EER). Best performance was shown by a task specific Weighted Euclidian distance which had a 10% EER.

TABLE V
VERIFICATION RESULTS USING TEMPORAL-SPATIAL PROFILES

Similarity Measure	Equal Error Rate
Euclidian Distance	12%
Mahalanobis Distance	13%
Manhattan Distance	12%
Weighted Euclidean Distance	10%

A great improvement in performance of the strategy based behavioral biometric system was observed with the inclusion of spatial information into the profiles as demonstrated in the Table 6. Once again the Weighted Euclidian distance function was the best matching algorithm obtaining 7% EER with general similarity measure functions performing in the range of 9-10% EER. Improvement in the performance of most similarity measure functions can be explained by a more refined capture of the player's strategy associated with inclusion of information about the spatial location of the player.

TABLE VI
VERIFICATION RESULTS USING TEMPORAL-SPATIAL PROFILES

Similarity Measure	Equal Error Rate
Euclidian Distance	9%
Mahalanobis Distance	10%
Manhattan Distance	9%
Weighted Euclidean Distance	7%

With the inclusion of the contextual information the dimensionality of behavioral profile has ballooned to 427D and the influence of the "curse of dimensionality" became apparent. Performance of all similarity measures has significantly decreased. With such a high-dimensionality-behavioral-profile the number of zero-value variables becomes overwhelming as the amount of time needed to collect sufficient data is unreasonable for any real-life security system.

TABLE VII
VERIFICATION RESULTS USING TEMPORAL-SPATIAL-CONTEXTUAL PROFILES

Similarity Measure	Equal Error Rate
Euclidian Distance	33%
Mahalanobis Distance	36%
Manhattan Distance	33%
Weighted Euclidean Distance	29%

V. CONCLUSION

A number of conclusions can be drawn from the results of our experiments. Examined general similarity measure functions showed an acceptable profile verification performance with Euclidian and Manhattan distances being indistinguishable from each other in terms of their accuracy. Mahalanobis distance function performed slightly worse possibly as a result of the normalization procedure which took into account variance of the data in each profile. Since the degree of variance in each user profile is different it is possibly that normalization was not evenly distributed and so produced a slight decrease in the performance of this general similarity measure function.

Customized Weighted Euclidian measure function specifically designed for the domain of poker-based behavioral profiles showed the best performance on all types of data representation. Heavier consideration for pre-flop player's actions allowed this similarity measure function to pick out the fundamental tendencies of the player's strategy and as a result improve algorithms verification accuracy to as low as the 7% EER for the behavioral profiles enhanced with temporal and spatial information.

In this paper we have compared performance of well established similarity measure functions to that obtained from customized field-specific approach in the domain of strategy-based behavioral biometrics. While all similarity measure functions showed a relatively high accuracy levels during user verification, Weighted Euclidian similarity measures has slightly outperformed general approaches such as Manhattan distance or Mahalanobis distance. This is probably cased by the fact that customized functions take advantage of the expert knowledge about the nature of the feature level data and give more weight to values with higher discriminatory ability.

Matching algorithms are a fundamentally important component of any biometric system. While general similarity measure functions are valuable for quickly developing prototype systems, only customized functions can provide the desired level of accuracy demanded by the modern security systems. In the future we would like to investigate optimal ways to combine output from the developed similarity measure functions for multiple behavioral profiles, such as those used in multimodal biometric systems. Such systems decrease the influence of the noise in the data and as a result make accurate individual verification more likely.

REFERENCES

[1] M. Badizadegan, *Texas Hold'em Flop Types*, Goldstar Books, Los Angeles, California, 1999.

[2] P. M. Baggenstoss, *Class-specific classifier: avoiding the curse of dimensionality*, *Aerospace and Electronic Systems Magazine*, Jan 2004, pp. 37- 52.

[3] C. Barral, J. Coron and D. Naccache, *Externalized Fingerprint Matching*, *Cryptology ePrint Archive*, 2004.

[4] H. Eidenberger, *Evaluation and Analysis of Similarity Measures for Content-based Visual Information Retrieval*, *ACM Multimedia Systems Journal*, 2006.

[5] A. K. Jain, R. Bolle and S. Pankanti, *BIOMETRICS: Personal Identification in Networked Society*, Kluwer Academic Publishers, 1999.

[6] T. Kinnunen and I. ainen, *Class-discriminative weighted distortion measure for VQ-based speaker identification*, *In Proc. Joint IAPR International Workshop on Statistical Pattern Recognition*, Windsor, Canada, August 6-9, 2002, pp. 681-688.

[7] K. Lee and H. Park, *A New Similarity Measure Based on Intraclass Statistics for Biometric Systems*, *ETRI Journal*, Oct. 2003, pp. 401-406.

[8] H. Lei, S. Palla and V. Govindaraju, *ER2: An Intuitive Similarity Measure for On-Line Signature Verification*, *IWFHR '04: Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR'04)*, IEEE Computer Society, 2004, pp. 191--195.

[9] O. Mut and M. Göktürk, *Improved Weighted Matching for Speaker Recognition*, *The Third World Enformatika Conference, WEC'05*, Istanbul, Turkey, April 27-29, 2005, pp. 229-231.

[10] M. Neuhaus and H. Bunke, *An error-tolerant approximate matching algorithm for attributed planar graphs and its application to fingerprint classification*, *Proc. Joint IAPR Int. Workshops Structural, Syntactic, and Statistical Pattern Recognition*, 2004, pp. 180 -189.

[11] Poker-edge.com, *Stats and Analysis*, Available at: http://www.poker-edge.com/stats.php, Retrieved June 7, 2006.

[12] A. Schwaighofer, *Sorting it out: Machine learning and fingerprints*, *Telematik*, 2002, pp. 18-20.

[13] A. Sturn, *Cluster Analysis for Large Scale Gene Expression Studies*, *Masters Thesis. The Institute for Genomic Research,* Rockville, Maryland, USA, December 20, 2000.

[14] Wikipedia, *Mahalanobis Distance*, Available at: http://en.wikipedia.org/wiki/Mahalanobis_distance, Retrieved August 22, 2006.

[15] R. V. Yampolskiy, *Behavior Based Identification of Network Intruders*, *19th Annual CSE Graduate Conference (Grad-Conf2006)*, Buffalo, NY, February 24, 2006.

[16] R. V. Yampolskiy and V. Govindaraju, *Use of Behavioral Biometrics in Intrusion Detection and Online Gaming*, *Biometric Technology for Human Identification III. SPIE Defense and Security Symposium*, Orlando, Florida, 17-22 April 2006.

[17] S. Yang and I. Verbauwhede, *A Secure Fingerprint Matching Technique*, *In Proc. ACM Workshop on Biometrics Methods and Applications*, 2003, pp. 89-94.

Roman V. Yampolskiy Roman V. Yampolskiy holds an MS in Computer Science degree from Rochester Institute of Technology (2004) and is a PhD candidate in the department of Computer Science and Engineering at the University at Buffalo. His studies are supported by the National Science Foundation IGERT fellowship. Roman’s main areas of interest are artificial intelligence, behavioral biometrics and intrusion detection. Roman has a number of publications describing his research in neural networks, genetic algorithms, pattern recognition and behavioral profiling.

Venugopal Govindaraju is a professor of Computer Science and Engineering at State University of New York at Buffalo, and Associate Director of CEDAR, the Center of Excellence for Document Analysis and Recognition at his university. He received his Ph.D degree in Computer Science at the University at Buffalo in 1992. His research is focused on Human Computer Interaction, Pattern Recognition, and Biometrics.

.