

# A Real-Time Automated System for the Recognition of Human Facial Expressions

Keith Anderson and Peter W. McOwan

**Abstract**—A fully automated, multistage system for real-time recognition of facial expression is presented. The system uses facial motion to characterize monochrome frontal views of facial expressions and is able to operate effectively in cluttered and dynamic scenes, recognizing the six emotions universally associated with unique facial expressions, namely happiness, sadness, disgust, surprise, fear, and anger. Faces are located using a spatial ratio template tracker algorithm. Optical flow of the face is subsequently determined using a real-time implementation of a robust gradient model. The expression recognition system then averages facial velocity information over identified regions of the face and cancels out rigid head motion by taking ratios of this averaged motion. The motion signatures produced are then classified using Support Vector Machines as either nonexpressive or as one of the six basic emotions. The completed system is demonstrated in two simple affective computing applications that respond in real-time to the facial expressions of the user, thereby providing the potential for improvements in the interaction between a computer user and technology.

**Index Terms**—Expression, face, motion, real-time.

## I. INTRODUCTION

AFFECTIVE computing addresses issues relating to emotion in computing and has been pioneered by the work of Picard at MIT. [1]. Picard describes how “affective interaction can have maximal impact when emotion recognition and expression is available to all parties, human and computational” and goes on to say that “if one party cannot recognize or understand emotions then interaction is impaired” [2]. The field of affective computing addresses this problem, attempting to enhance interactions between humans and technology through the development of artificial systems responsive to the moods and emotions of a human user.

In humans, one important cue as to our emotional state is our facial expression. It therefore follows that for interaction with technology to be improved and become more natural that it be provided with an ability to display and recognize facial expression. In recent years attempts have been made to do exactly this, with an ability to express emotion being bestowed upon some sociable robots [3] and also on synthetic characters incorporated into pieces of computer software. However, less progress has been made in providing such agents with skills enabling them to understand the expressions of human users.

An effective solution to this problem would have a range of useful applications. For example, allowing a robot to understand the expressions of humans would enhance its effectiveness at performing many tasks [3]. Expression recognition also has a part to play in software or educational tutorials [2], measurement tools for behavioral science [4], as a mechanism for detecting deceit [5] and in the development of socially intelligent software tools for autistic children [6].

In recent years, research into automated recognition of expression has become active, with advancements having been made since the intrusive, although accurate, methods of [7] and [8] that required the tracking of markers applied to the faces of test subjects. Subsequent expression recognition approaches have generally either attempted to recognize expression of Ekman’s six basic emotions (happiness, surprise, sadness, disgust, fear, anger) [9] or to recognize Action Units (AUs) of the Facial Action Coding System (FACS) [10]. This FACS is an anatomically based approach allowing movement of different parts of the face to be described. It consists of 44 unique and visually distinguishable movements called AUs. The following paragraphs describe some of the more significant work carried out in the field of automated expression recognition.

### A. Previous Approaches to Facial Emotion Recognition

Facial motion is a commonly used cue as to expression change, with the work of [11] an important early example. In [11], Mase and Pentland used dense optical flow to estimate the activity of 12 of the 44 facial muscles. The motion seen on the skin surface at each muscle location was compared to a pre-determined axis of motion along which each muscle expands and contracts, allowing estimates as to the activity of each muscle to be made. Recognition rates of 86% were reported. Optical flow has subsequently been used in conjunction with heuristic rules in [12] and with radial basis function networks in [13] to recognize expression of the basic emotions. [14] combined motion information with a face model describing the position of attachment of the 44 facial muscles and the elastic properties of the skin. This allowed the forces involved to be estimated and achieved a recognition rate of 98%.

Alternatives to the use of optical flow include image [15], [16] and feature-based [17]–[19] methods. The feature-based method of [17] used an active shape model to locate facial features and then used shape and texture information at these locations to classify expressions, obtaining a recognition rate of 74% for expression of the basic emotions and also neutral expressions. The image based method of [16] used PCA and Independent Components Analysis (ICA) to recognize asymmetric AUs from the FACS, as well as determining the intensity of these

Manuscript received February 25, 2004; revised August 16, 2004 and January 31, 2005. This paper was recommended by Associate Editor H. Qiao.

The authors are with the Department of Computer Science, Queen Mary, University of London, London E1 4NS, U.K. (e-mail: ka@dcs.qmul.ac.uk; pmco@dcs.qmul.ac.uk).

Digital Object Identifier 10.1109/TSMCB.2005.854502

AUs. ICA was found to perform better than PCA, with an AU recognition rate of 83% achieved on test images consisting of an image showing a single AU. Recognition rates fell to 74% when images of the face showing multiple AUs were included in the test set.

A commonly used approach to improve robustness in classifying expression is to combine the results of several different methods [4], [20], [21]. For example, [4] used PCA, optical flow, and feature measurement to classify AUs of the FACS. By combining these techniques a recognition rate of 92% was obtained, higher than that achieved when each approach was used separately.

### B. Approach Taken Here to Facial Expression Recognition

The work presented here presents a novel solution to the facial expression recognition problem, describing a fully automated real-time expression recognition system that can be used to drive a broad range of applications in the field of human computer interaction. The particular focus of this work is to address problems with previous solutions, specifically their slowness and/or requirement for some degree of manual intervention. These failings have meant that it has not been realistic for technology to respond in real time to the facial expressions of a user.

Real-time recognition of expression is achieved in this system as the approach is restricted to recognising only frontal views of expressions at a single scale. However, these restrictions are not weaknesses in the application domains for which the use of this system is envisaged, namely expression recognition of a user seated in front of a personal computer. In such an environment, the user's gaze is generally directed at the computer monitor and thus recognition from other viewpoints is not of real significance (the problem of pose change has also not been seriously addressed in other systems where there are fewer computational constraints [22]). Also, the distance of the user's face from the camera is fairly constant, with the system described here able to cope with the limited changes in scale that would occur.

There are three main components to this system: a face tracker, an optical flow algorithm and an expression recognition system. The face tracker is a modification of, and an extension to, the ratio template algorithm [23]. Optical flow is determined by a real-time version of a multichannel gradient model [24], whilst the final expression recognition system uses Support Vector Machines (SVMs) [25]. These components are integrated into a single system running at 4 fps on a  $384 \times 247$  image on a 450-MHz Pentium III machine with Matrox Genesis DSP boards. The system is summarized in Fig. 1. The system developed is able to find and recognize the facial expressions of any user who sits in front of a computer equipped with a camera in real-time, even in cluttered and dynamic scenes. It recognizes expression of the six basic emotions, namely happiness, sadness, disgust, surprise, fear, and anger.

### C. Summary of Paper Structure

This paper is divided as follows. Section II provides a short introduction to the system face tracker used. Section III introduces the robust differential based optical flow algorithm used



Fig. 1. System summary. (a) Face in the scene is located using a face tracker. (b) Motion of the face region only is determined using an optical flow algorithm. (c) Processed motion data is input into SVM classifiers. The motion signature is labeled as either one of the six basic emotions or as nonexpressive according to classifier outputs. In this case, the sequence is labeled as happiness as the highest output is happiness (0.98) and this value exceeds the required expression—nonexpression threshold.

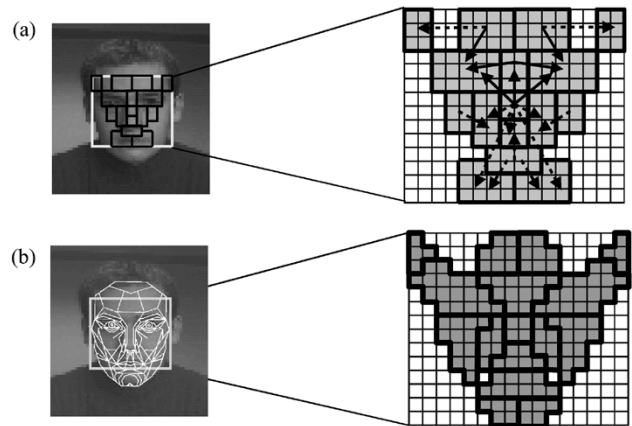


Fig. 2. (a) Spatial face template used by Scassellati [28]. Grey regions indicate areas over which pixel values are averaged, arrows indicate relations. A relation is satisfied if the average greyscale value from the first (arrow tail) to second (arrow head) exceeds a threshold of 1.1. Two types of relation exist, essential (solid arrows) and confirming (dashed arrows). If 10 out of 11 of the essential and 8 out of 12 confirming relations are satisfied, then the location is said to contain a face. (b) Modified spatial face template altered according to a golden ratio face mask [29].

by the system. Section IV then describes the expression recognising component of the system in detail. This includes comparison between different techniques taken to compress the representation of motion data and also a comparison between multi-layer perceptron (MLP) and SVM [25], [26] expression classification techniques. A description of some applications to which the completed system has been put to use is provided in Section V, before finally a discussion and summary is given in Section VI.

## II. FACE TRACKING

The face-tracking component of the system [27] is founded upon a modified version of the ratio template algorithm [23]. The algorithm in [23] operates by matching ratios of averaged luminance using a spatial face model [Fig. 2(a)]. A detailed description of the original ratio template approach can be found in [28].

Our version of the ratio template algorithm [27] is enhanced by the inclusion of biological proportions (the golden ratio) [29] into the face model [Fig. 2(b)], and by examination of ratio of ratio relationships rather than relying simply on the original methods basic ratio measures. We have shown the inclusion of biological proportions improves robustness to illumination change, whilst the examination of higher order relationships helps reject false positives [27]. The information provided

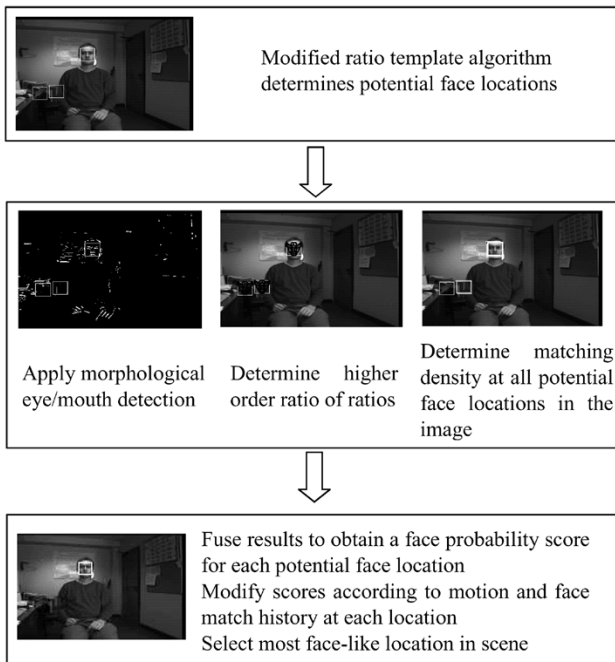


Fig. 3. Face tracker summary.

TABLE I  
SUMMARY OF CONDITIONS UNDER WHICH FACE TRACKER CAN OPERATE

Operating condition	Tolerance
Illumination	Lighting from above. Performance gradually degrades as light source moves to left or right of subjects face
Scale	$\pm 20\%$ from optimal scale
Roll	Head $\pm 10^\circ$ from vertical
Yaw	Head $\pm 30^\circ$ from frontal view around horizontal plane
Tilt	Head $\pm 20^\circ$ from frontal view around vertical plane

by the modified ratio template algorithm is then combined with other image measures, simple morphological eye/mouth detection, image motion data, matching history, and matching density to allocate a single face probability to each location in the scene. Our system does not currently make use of color information and identifies the single most face-like structure in a scene at speeds of  $\sim 14$  fps on a  $384 \times 247$  image using the Matrox Genesis DSP boards, and thus is ideal for inclusion into a real-time system. Fig. 3 summarizes the key stages of the face tracker.

The face tracker is able to detect frontal views of faces under a range of lighting conditions, although it is ineffective when the subject is illuminated from beneath. The method is able to handle limited changes in scale, yaw, roll, and tilt of the head as summarized in Table I. Its strengths lie in its ease to implement, its speed, and its tolerance to the different illuminations characteristic of an unstructured indoor environment. As the ratio template algorithm uses a spatial face template, it also provides a rough spatial map to allow more focused searching for facial features. This rough map is of importance to the expression recognition system, as it is crucial to know the positions of the different facial features before extraction of their relative motion is possible.

### III. DETERMINING OPTIC FLOW IN FACIAL REGIONS

Once a face has been located in the scene by the face tracker, an optical flow algorithm determines the motion of the face. Motion information is used for the purposes of expression recognition as, first, expressions are inherently dynamic events, and, second, by using motion the task is simplified as it ignores variations in the texture of different people's faces. Hence, the facial motion patterns seen when each of the basic emotions are expressed are similar, independent of who is expressing the emotion. Interestingly, facial motion alone has already been shown to be a useful cue in the field of human face recognition, for example in determining gender [30].

In the system described here, the multichannel gradient model (MCGM) is employed to determine facial optical flow [24]. The MCGM is based on a model of the human cortical motion pathway, and operates in three dimensions (two spatial, and one temporal), recovering the dense velocity field of the image at all locations. The model involves the application of a range of spatial and temporal differential filters to the image, with appropriate ratios being taken to recover speed and direction. Interestingly, there is evidence to suggest that the MCGM is a biologically plausible model of the human cortical motion pathway as it has been shown to correctly predict a number of motion-based optical illusions [24].

The MCGM is chosen for the extraction of motion information for two reasons. Firstly, by using a ratio of differentials in calculating speed, it is robust to changes in scene luminance, thus removing possible contrast normalization problems associated with Fourier energy or template matching methods of recovering optical flow [31]. Secondly, a real-time version of the MCGM has been implemented on a machine using Matrox Genesis DSP boards [32]. This real-time version of the MCGM is optimized for hardware from the original version, but still provides a robust measure of optical flow. It runs at speeds of 18 fps on a  $64 \times 41$  image using the Matrox Genesis DSP boards and thus is ideal for incorporation into a real-time expression recognition system.

The face tracker described in Section II gates the active area of the optic flow calculation. To improve efficiency, rather than calculate optic flow over the whole scene, the tracker provides a location (most probably the face) in the scene and undertakes a local area computation of motion at the putative face location and its immediate surrounds. The region of motion computed provides a border around the face allowing for reasonable horizontal and vertical rigid head movement during expression, otherwise rigid head motions would immediately move the face outside the region of the image where the optical flow algorithm was operating. The border size is easily modified, allowing for more or less rigid head movement, but obviously has an effect on computational speed.

### IV. FRAMEWORK FOR EXPRESSION RECOGNITION

Once facial motion has been determined, it is necessary to place the motion signatures into the correct class of facial expression (either a nonexpression or one of the six basic emotions). The classifiers used in this work were trained and tested

using a subset of the CMU-Pittsburgh AU-coded facial expression database [33], [34] that provided 253 examples of expression of the six basic emotions. This expression set consisted of 57 sequences of the expression of happiness, 49 of sadness, 55 of surprise, 30 of disgust, 33 of anger, and 29 of fear. 70% of this set was used for training/validation of classifiers, with the remaining 30% used for testing. However, before the examples included in the CMU-Pittsburgh AU-coded facial expression database could be used for training/testing of classifiers, it was necessary to pre-process the sequences such that they were in the format required by the system. This pre-processing took the following form.

- Reduce frame rate of sequences from the 40fps of the original CMU-Pittsburgh AU-Coded database to the 4 fps at which the current completed system runs.
- Reduce the size of the faces in the images to the scale detectable by the face tracker and to that used by the expression recognition system (approximately  $45 \times 55$  pixels in size). It should be noted that all images in the database were reduced to the same size, so some variation in head size is present due to normal human differences and the slightly different poses adopted by people in the seats when filmed.
- Label the start frame of each sequence where facial expression begins.
- The sequences were then used to generate motion pattern examples for expression of each of the six basic emotions. In this system, the approach taken to generate data for classifier training/testing was fully automated once the pre-processed sequences were available. The face tracker described in Section II was used to obtain face locations and provided a rough map of face feature locations. The MCGM was then used to obtain the facial motions according to this spatial map. Thus, the facial motion data was generated for training the classifiers without any need for manual labeling of facial feature locations. The system used four consecutive frames of motion data generated by the MCGM to represent each example of a facial expression. As the overall system frame rate was 4 fps, these four frames represent one second of facial motion and consist only of the start phase of a facial expression (i.e., from neutral face to expressive face).

Prior to entry of this motion data into the classifiers used by the expression recognition system, two general approaches were investigated to condense and modify the data. These were the averaging of motion data and the taking of ratios of averaged motion:

- Motion Averaging—Rather than inputting the raw optical flow output directly into the classifiers, the motion (speed and direction) data generated by the MCGM is condensed into a more efficient form by averaging over specific pre-defined regions of the face. Such condensation of data reduces the amount of information entered into the classifiers making the classification task more tractable.
- Ratios of averaged motion—To remove the effects of rigid head translation motion in the system described here, ratios of motion are taken to determine how different facial parts are moving relative to one another. If someone were

to move the head as a whole whilst expressing emotion facially, a dramatic change would be seen in the optical flow output of the MCGM. For example, the optical flow output of someone smiling whilst moving the head downwards would look very different from that of someone smiling whilst raising the head. So by using relative motions the system exhibits invariance to global head translations.

In addition to active expression examples taken from the CMU-Pittsburgh AU-coded facial expression database, classifiers were also trained and tested using a nonexpression training/testing set. This is vital, as for incorporation of the system into actual applications, it is important not only to label an input motion signal to the correct facial expression, but it is also important to know when the input data is not indicative of one of the recognized expressions. The nonexpression set was 4800 frames long and consisted of ten different human subjects recorded under normal indoor illumination using a computer in an unconstrained manner.

#### A. Empirical Comparisons to Examine Motion Averaging and Expression Classification Methods

In this section, we detail the experimental work undertaken to select the best representation of facial motion and best method for classification of said motions, for use in the system. To classify expressions most effectively the motion data generated by the MCGM must be introduced to the classifiers in a form that makes distinction between different expression classes as easy as possible. Thus, selection of regions over which to average the motion data and the ratios subsequently taken is of importance. Early work on the system involved carrying out a range of empirical studies using different motion data representations and evaluating classifier performance with the aid of ROC curves.

ROC curves plot False Accept Rates (FARs) against False Reject Rates (FRRs), where FAR is the percentage of negative examples incorrectly classified as positive examples, whilst the FRR is the percentage of positive examples incorrectly classified as negative examples. By varying the threshold level at which the classifiers fire for facial expressions the values of the FAR and FRR change, and by plotting ROC curves the relationship between the two can be seen. The two-dimensional (2-D) representation of an ROC curve is particularly useful as it shows system performance over a range of thresholds rather than an arbitrarily chosen threshold. This is important as in some application domains a false detection is not as costly as a missed detection whilst in others the opposite maybe true. Thus, an ROC curve shows how the performance varies as thresholds are changed and can help a user choose a threshold for their specific application.

In addition to the performance representation provided by the ROC curves, two additional values are provided to aid evaluation of classifier performance. These are the area under the curve (AUC) and the absolute recognition rates. In the work presented here, the AUC is defined as the area under the ROC curve at FARs less than 5%. The absolute recognition rate is the percentage of expressions correctly classified, independent of FARs. An expression is correctly classified if the classifier giving the strongest response to an input signal matches the actual class of the input signal.

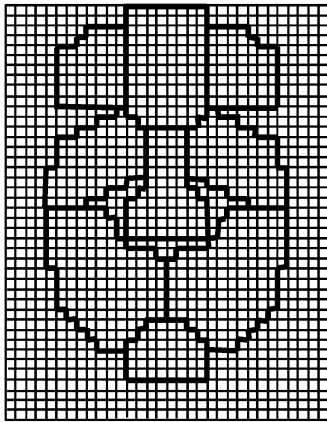
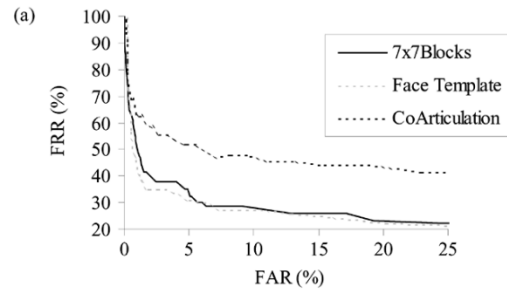


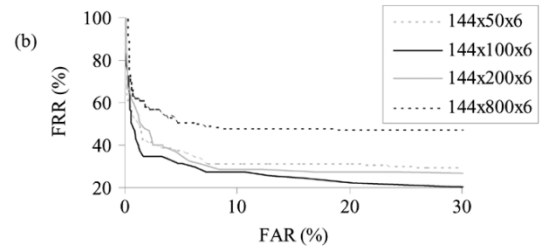
Fig. 4. Regions for motion averaging based on the Co-Articulation regions of Fidaleo and Neumann [35].

The empirical study into motion averaging involved the use of three different approaches. The first approach averaged motion according to the surface structure of the face and used the regions of the ratio template algorithms spatial face map [Fig. 2(b)], with two regions added to include motion of the chin. The second approach averaged motion according to the movement seen on the face during facial expression and involved the use of the co-articulation regions of Fidaleo and Neumann [35] (Fig. 4). Co-articulation regions describe the parts of the face that move together when expressions are made. These anatomically determined regions model the changes seen on the skin surface as a set of nine contiguous regions of skin deformation, and have been used previously for the purposes of animating 2-D cartoon characters [35]. The third approach was a basic blocking strategy where motion was averaged using a  $7 \times 7$  pixel grid placed over the entire face. A range of different sets of empirically determined ratios were used in conjunction with these two potential motion averaging representations and backpropagation MLPs, using the logistic activation function, of different sizes trained with each. The performance of the optimal MLP classifier for each motion averaging strategy is provided in Fig. 5(a). The best motion averaging approach was found to use the ratio template algorithm's spatial face template, with an AUC value and optimal recognition rate (37.5 and 81.82%, respectively) better than those achieved by the other two approaches.

This is contrary to the result that one would initially expect, as co-articulation regions model underlying muscular changes seen in the human face when expressions are made and thus one would predict that this approach would work the best. However, this approach was the least successful, working less effectively even than the basic blocking strategy. The reason for the relative failure of the co-articulation approach result is thought to be that the representations of the 18 spatial locations present in the ratio template algorithms spatial facial map is better at catching the subtle nuances of facial expression when compared to the nine co-articulation regions, allowing for finer discrimination, although at the computational cost of a more detailed description. The 18 regions of the ratio template algorithm's face map is thought to perform better than the 20 regions of the basic  $7 \times 7$  pixel blocking strategy as it more accurately models the surface structure of the face.



Data Representation	Absolute recognition rate %	AUC
CoArticulation regions	63.64	58.0
7x7 blocks	77.92	43.2
Ratio template algorithm's face template	81.82	37.5



MLP Size	Absolute recognition rate %	AUC
288x50x6	70.13	43.5
288x100x6	81.82	37.5
288x200x6	74.02	45.5
288x800x6	53.24	57.6

Fig. 5. (a) Performance of MLP classifiers using different motion averaging strategies. (b) Performance of MLP classifiers of different sizes using the optimal motion averaging strategy of the spatial face template of the ratio template algorithm.

The optimal motion data representation determined by this empirical study is summarized in Fig. 6(a), and as can be seen uses a set of 36 symmetrical motion ratios to remove rigid head motion. It was found that if fewer ratios are used performance is degraded, whilst use of a greater number of ratios does not further enhance performance.

The effect of using SVM classifiers, as opposed to MLPs, was then examined, and thus a range of different SVMs were trained using the 36-ratio spatial map representation shown in Fig. 6(a). For the SVMs, the classifier examination involved varying the kernel functions used [25] (RBF, sigmoidal, linear, polynomial), the strategy used for merging the binary outputs of the SVMs to solve the multiclass expression recognition problem [26] (one-against-one, one-against-all), and also by carrying out a grid search to find optimal parameters for the kernel function and error/margin trade off. This allowed for comparison with the optimal MLP approach that was found to work best with networks of size  $288 \times 100 \times 6$ . Fig. 6(b) provides ROC curves and tables contrasting classification performance when the best performing MLPs and SVMs were used.

The use of SVMs using radial basis function kernels and a "one-against-all" merging strategy [26] was determined to be a better candidate for the expression classification task than MLPs. Results show that the ROC curve and AUC value is much improved (30.8 as compared to 37.5) when SVMs are used,

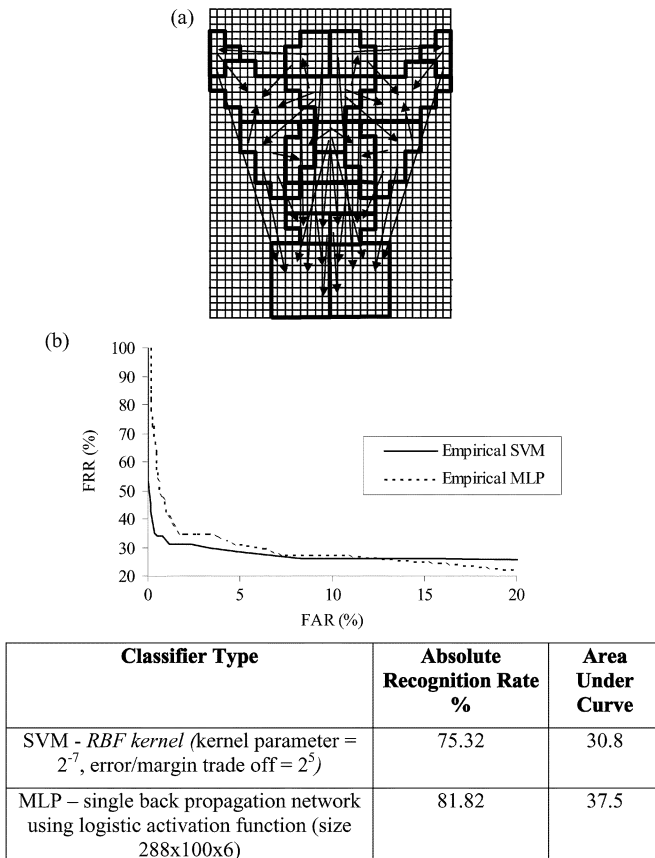


Fig. 6. (a) Optimal data representation as determined empirically. Regions of the spatial face map of the modified ratio template algorithm are used. The 36 arrows indicate ratios taken to determine relative motion, thereby removing the effects of any rigid head movement. (b) ROC curve and table summarizing performance of SVM and MLP classifiers using the optimal data representation.

although the best absolute recognition rate is achieved by the MLPs (81.82%). However, the AUC value and ROC curve is of most importance as the completed real-time system will operate using thresholds such that the FAR is low (to prevent overly frequent miss firing).

Statistical analysis of the difference in performance between SVM and MLP based classification was undertaken. This analysis required the use of moving examples of facial expressions from the widely available benchmark CMU-Pittsburgh AU-coded facial expression database, restricting the test set available to 77 examples. The analysis showed the probability of obtaining our results if both SVMs and MLPs were equally effective at the expression recognition task was 0.10. Based on the ROC curves and statistical analysis, it was decided to use SVMs as opposed to MLPs in our expression recognition system.

### B. Learning the Optimal Data Representation Using Simulated Annealing

In the previous section, we have shown that critical to system success is the representation of the input motion data from appropriate facial areas. Using the SVM architecture detailed in the preceding section we undertook to determine the optimal set of motion features that maximize the systems ability to classify facial expressions, as we felt it possible to improve on the

earlier empirically determined set. The optimization approach used for this work was simulated annealing [36].

An initial set of regions and ratios was chosen at random. This set consisted of 18 square regions  $5 \times 5$  pixels in size positioned randomly in the face and retaining a set of 36 ratios between the regions. A standard simulated annealing algorithm [37] was then run, with random changes made either to the positioning of the regions (moved randomly 1 pixel in either the horizontal or vertical direction) or to the ratios (one region of a ratio pair changed to a different region at random). Our SVMs using RBF kernels were trained using each data representation and the cost of a change to the representation was determined according to the mean squared error to a validation set of both expression and nonexpression examples. The algorithm was terminated when no improvement in error was achieved in 300 successive iterations.

The running of the algorithm produced the data representation shown in Fig. 7(a). Examination of the results of the optimization show that the regions of importance for expression recognition are the mouth corners, eye corners and nose-bridge. Of these areas, the most important region of the face for classification of facial expression appears to be the mouth corners, with a cluster of regions for averaging being positioned here. Only two regions are located at the nose-bridge as it only covers a small area. However, if one examines the ratios themselves it is evident that a large number involve the nose-bridge, demonstrating its importance. Most of the remaining regions are positioned at the edges of the eye. The motion of the chin and forehead seems to be of lesser importance.

The performance of this representation is compared to that achieved by the best empirical approach (the facial mask) in Fig. 7(b). The ROC curves when the empirical and optimized region/ratio sets were used are virtually identical, with AUC values of 30.8 and 31.8, respectively, although the absolute recognition rate is much improved using the data representation learnt by optimization, rising from 75.32% to 80.52%. Such performance was achieved despite the simulated annealing search involving several less degrees of freedom than were available to the ratio template algorithm's spatial face template.

The regions over which data is averaged in the ratio template algorithm's face template vary in size and are not all square, thereby allowing more accurate matching to the size and shape of facial components. However, the simulated annealing approach was constrained to square regions of one size only, 5 pixels  $\times$  5 pixels. Further investigation has not yet been carried out into the effects of allowing regions to change size and shape due to processing and memory constraints (the annealing process previously described already takes  $\sim 2$  days to complete on a 1.7-GHz machine). The expression recognition system described in this section is summarized in Fig. 8.

## V. PROTOTYPE AFFECTIVE COMPUTING APPLICATIONS

To demonstrate the usefulness of the system described previously, two simple applications were developed. These applications had thresholds set such that the FAR of the classifiers was at 1%. The first application was a chatroom system that automatically inserts emoticons (symbolic abbreviations for emotional state such as :) for happy). This application is based on the JChat

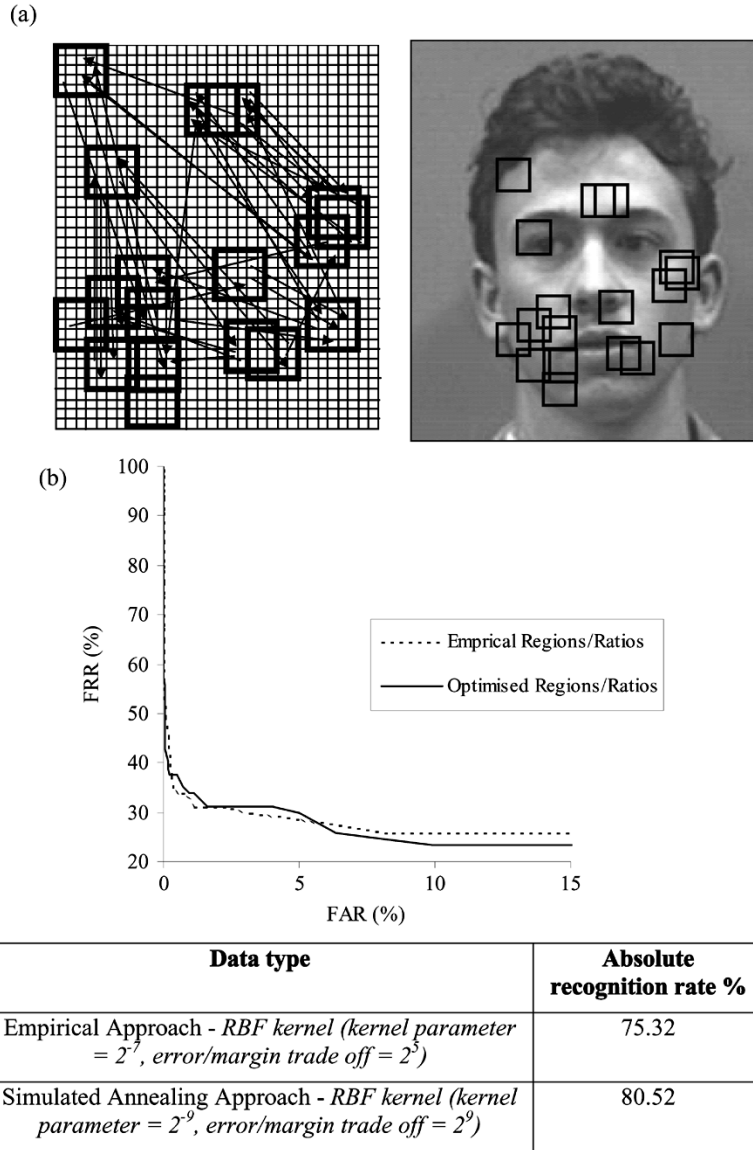


Fig. 7. (a) Regions and ratios learnt by initial simulated annealing approach. (b) Comparison between performance when optimized and empirical regions and ratios are used.

application [38] and is called EmotiChat. An example sequence of the application in use is provided in Fig. 9. The user is shown seated in front of the monitor and is reading the text. The expression recognition system is tracking the motions of the face, and when for example the smile is seen, the system recognizes this and triggers the insertion of a “happy” emoticon into the text of the EmotiChat application.

The second prototype application developed monitors the expressions of a user and automatically triggers a desktop application, such as a web browser or media player in response. For example, the system could respond to a sad expression on the user’s face by playing some happy music to cheer up the user or by opening an appropriate web page. Results of a survey provided to 100 students indicate that such applications are perceived of as being useful.

The working system was examined to characterize the sensitivity of this expression recognition method to changes in user position. Our methodology was developed without explicit inclusion of methods to account for large variations in face

scale, yaw or rotation. Though the system was trained only using frontal views of facial expressions, with each image of the training sequence pre-processed to a single size, there was some variation in head size within the training sequence due to normal human variation. To examine the systems actual tolerances we placed several subjects in front of the active system and gradually changed the camera viewpoint. The results of this study suggest that the expression recognition system can effectively handle changes in yaw  $\pm 20\%$  from a fully frontal view and scale changes  $\pm 20\%$  from optimum (this equates to a distance range of around one meter in the set-up used for the experiment). For applications involving the use of a desktop computer, where a user’s gaze is directed in main at the monitor and their distance from the monitor is fairly constant, it is thought that such tolerances are adequate.

## VI. DISCUSSION AND CONCLUSIONS

This paper has presented results for a fully automated real-time expression recognition system able to distinguish between

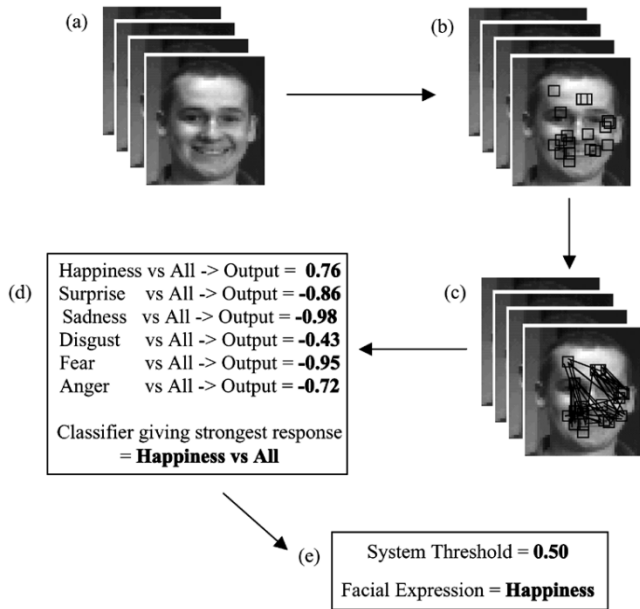


Fig. 8. Summary of expression recognition approach. (a) Obtain four consecutive frames of facial motion information over an area of  $52 \times 58$  pixels using the multichannel gradient model (this includes a  $10 \times 10$  pixel border around the whole face to allow for some vertical and horizontal rigid head motion). (b) Average motion (speed and direction) data over 16 facial regions. These regions can be those learnt using an optimization approach or determined empirically. (c) Take 36 ratios of averaged motion to determine relative movement, thereby removing rigid head motions. (d) Enter 288 velocity inputs (144 direction, 144 speed) into six “l-against-all” SVMs using radial basis function kernels. Determine which classifier gives the strongest response. (e) If winning classifier output exceeds threshold, label motion as a facial expression. If output is below threshold label motion as nonexpressive. This threshold can be altered (thereby changing the FAR-FRR ratio) depending on the application domain for which the system is to be used.

expressions of happiness, sadness, surprise, disgust, fear, and anger. The system makes use of the constraints naturally occurring in the user scenario for which it was developed, specifically a single user at a fairly well fixed distance and aspect to the camera. The novel components of this work include the integration and extension of existing techniques to produce a complete expression recognising system that can be used by anyone, the full automation of the expression recognition systems training process, the optimization approach used to learn facial regions for the expression recognition task and the use of ratios to remove the effects of rigid head motion. We have also shown the system to operate effectively in a number of prototype interactive user applications.

The system described in this work is able to operate in real-time as each component stage of the system has been selected due to its relative computational simplicity. Once the tracker has found a face in the scene, computation of the 23 convolutions of the ratio template algorithm is sufficient to track the face, whilst the whole scene is searched for new, more face-like, objects using the complete tracker (e.g., with morphological filtering) just once every 16 frames. The real-time version of the multichannel gradient model used to determine optical flow has been heavily optimized, measuring motion using up to third-order derivatives (rather than the sixth-order derivatives of the full version [24]) and eight orientations (rather than 24). Rather than inputting raw optical flow data, the expression recognition system employs a motion averaging strategy that allows outputs to be



Fig. 9. Example of EmotiChat application in use. (a) Two users log into the EmotiChat system, a chatroom application linked to the real-time facial expression recognition system. (b) Whilst chatting, a statement is made and the chatroom user smiles. (c) The real-time expression recognition system correctly classifies the motion signature as a smile and automatically inserts an emoticon into the chatroom text.

obtained using just six SVMs with around 100 Support Vectors each. This can be contrasted with other approaches that use the optical flow in conjunction with computationally heavy 3-D face models [14] or more complicated classifier architectures [13].

It should be noted that the absolute best recognition rate (81.82%) achieved is lower than those cited by some other systems discussed in Section I (e.g., the optimum recognition rate of 98% for the approach described in [14]). Nonetheless, the performance is comparable to other techniques that use a basic optical flow approach (e.g., 84.5% in [4], 80%–94% in [12], and 88% in [13]), despite our system being constrained due to its incorporation in a real-time, fully automated system. Full automation of both the training and testing phases leads to slight inaccuracies in the location of the face by the face tracker, unlike other approaches where facial features are manually located [12], [13]. As a real-time system running at 4 fps, our system is restricted by use of only four frames of motion to

represent each facial expression, by use of a slightly less accurate real-time optical flow algorithm, due to faces being only  $\sim 30 \times 40$  pixels in size, and by employing motion averaging strategies over a limited number of facial regions. Such restrictions lead to loss of data relating to more subtle facial motions that could help the classifiers differentiate between the facial expressions and thus will have reduced overall performance.

Additionally, our system was not solely trained to distinguish between different expressions, but also to distinguish between expressive and nonexpressive sequences. Most of the other systems described in Section I did not consider this issue and our system's optimal recognition rate has been affected by their inclusion in the training set. Our system also found classification of the fear and anger expressions particularly troublesome, and thus if figures for these expressions are ignored, correct classification rates of around 90% are achieved. Although a more detailed comparison with past approaches is desirable, this direct comparison with the literature cannot be undertaken due to differences in the test databases used, the expressions recognized (e.g., AUs from the FACS, some or all of the Ekman's basic expressions), and omission of data indicating the computational speed of such systems.

Nonetheless, there is still considerable potential for further investigation. For example, in the work described previously in this paper the data representation for each expression is identical. However, different facial components do not move in the same way and have varying importance in characterising different facial expressions. Thus, although the technique used in this work uses a rapid "one size fits all" approach, it is possible that the introduction of different data representations to test for each individual expression could enhance performance further.

The work presented here attempts to recognize the six emotions universally associated with unique facial expressions. However, there is no reason to believe that these expressions are those exclusively of importance when human subjects interact with technology. Further study into the expressions that actually manifest themselves during human computer interaction would be useful in selecting expressions to be learnt in automated expression recognition. Also, it would be interesting to explore whether this set of expressions is altered and/or people become more expressive facially if they know that technology can understand and respond to their facial cues.

Finally, a detailed investigation into the efficacy of learning expression features is also warranted. Some preliminary investigations have been made here into the use of simulated annealing for learning facial regions to average motion data over for the purposes of expression recognition. This approach performed to a similar level as the optimal empirical SVM approach (with an improved absolute recognition rate) even when fewer degrees of freedom were available. Further investigations are therefore merited into learning facial regions when the size and shape of regions are not restricted, as well as the number of regions and ratios used.

## REFERENCES

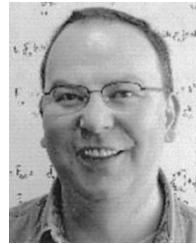
- [1] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1995.
- [2] ———, "Toward agents that recognize emotion," in *Actes Proc. IMAGINA*, 1998, pp. 153–155.
- [3] C. Breazeal, "Emotion and sociable humanoid robots," *Int. J. Human-Comput. Stud.*, vol. 59, pp. 119–155, 2003.
- [4] M. S. Bartlett, J. C. Huger, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, pp. 253–263, 1999.
- [5] M. S. Bartlett, G. Donato, J. R. Movellan, J. C. Huger, P. Ekman, and T. J. Sejnowski, "Face image analysis for expression measurement and detection of deceit," in *Proc. 6th Annu. Joint Symp. Neural Computation*, 1999, pp. 8–15.
- [6] B. Ogden, "Interactive Vision in Robot-Human Interaction," Progress Report, 2001.
- [7] S. Kaiser, T. Wehrle, and S. Schmidt, "Emotional episodes, facial expressions, and reported feelings in human-computer interactions," in *Proc. 10th Conf. Int. Society for Research on Emotions*, 1998, pp. 82–86.
- [8] W. Himer, F. Schneider, G. Kost, and H. Heimann, "Computer-based analysis of facial action: A new approach," *J. Psychophys.*, vol. 5, pp. 189–195, 1991.
- [9] *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Eds., Wiley, Chichester, U.K., 1999. P. Ekman, Basic emotions.
- [10] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.
- [11] K. Mase and A. Pentland, "Recognition of facial expression from optical flow," *IEICE Trans. E*, vol. 74, pp. 408–410, 1991.
- [12] Y. Yacoob and L. Davis, "Recognizing facial expressions by spatio-temporal analysis," in *IEEE CVPR*, 1993, pp. 70–75.
- [13] M. Rosenblum, Y. Yacoob, and L. Davis, "Human emotion recognition from motion using a radial basis function network architecture," in *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994, pp. 43–49.
- [14] I. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 757–763, Jul. 1997.
- [15] C. Padgett and G. W. Cottrell, "A simple neural network models categorical perception of facial expressions," in *Proc. 20th Annu. Cognitive Science Conf.*, 1998, pp. 806–807.
- [16] B. Fasel and J. Lüttin, "Recognition of Asymmetric Facial Action Unit Activities and Intensities," IDIAP Research Rep., 1999.
- [17] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 743–756, Jul. 1997.
- [18] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Netw.*, vol. 16, pp. 555–559, 2003.
- [19] M. Garghesha and P. Kuchi, "Facial expression recognition using artificial neural networks," *Artif. Neural Comput. Syst.*, pp. 1–6, 2002.
- [20] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li, "Automated facial expression recognition based on FACS action units," in *Proc. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 390–395.
- [21] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, pp. 97–115, 2001.
- [22] B. Fasel and J. Lüttin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, pp. 259–275, 2003.
- [23] P. Sinha, "Perceiving and Recognising Three-Dimensional Forms," Ph.D. dissertation, M. I. T., Cambridge, MA, 1995.
- [24] A. Johnston, P. W. McOwan, and C. P. Benton, "Robust velocity computation from a biologically motivated model of motion perception," in *Proc. Roy. Soc. Lon.*, vol. 266, 1999, pp. 509–518.
- [25] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Disc.*, vol. 2, pp. 121–167, 1998.
- [26] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Tran. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
- [27] K. Anderson and P. W. McOwan, "Robust real-time face tracker for use in cluttered environments," *Comput. Vision Image Understand.*, pp. 184–200.
- [28] B. Scassellati, "Eye finding via face detection for a foveated, active vision system," in *Proc. 15th Nat. Conf. Artificial Intelligence*, 1998, pp. 969–976.
- [29] S. R. Marquardt. (2003). MBA website. [Online]. Available: <http://www.beautyanalysis.com>
- [30] H. Hill and A. Johnston, "Categorising sex and identity from the biological motion of faces," *Current Biol.*, vol. 11, pp. 880–885, 2001.
- [31] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. f Comput. Vis.*, vol. 12, pp. 43–77, 1994.
- [32] J. L. Dale, "A Real Time Implementation of a Neuromorphic Optic Flow Algorithm," Ph.D. dissertation, University College London, London, U.K., 2002.

- [33] J.J.J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, "Detection, tracking, and classification of subtle changes in facial expression," *J. Robot. Auton. Syst.*, vol. 31, pp. 131–146, 2000.
- [34] J. F. Cohn, A. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding," *Psychophysiology*, vol. 36, pp. 35–43, 1999.
- [35] D. Fidaleo and U. Neumann, "CoArt: Co-articulation region analysis for control of 2D characters," in *Proc. IEEE Computer Animation 2002*, 2002, pp. 12–17.
- [36] V. Cerny, "Thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm," *J. Optimiz. Theory and Applic.*, vol. 45, pp. 41–51, 1985.
- [37] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. New York: Wiley, 1989.
- [38] T. Timoteo. (2003). J Chat. [Online]. Available: [www.ansurgen.org/board/ttimoteo/jchat/](http://www.ansurgen.org/board/ttimoteo/jchat/).



**Keith Anderson** was born in the United Kingdom in 1978. He received the B.Sc. degree in biochemistry from the University of Bristol, Bristol, U.K., in 1999, the M.Sc. degree in information technology from Queen Mary College, London, U.K., in 2000, and the Ph.D. degree in computer science from Queen Mary College in 2004.

He is currently with Orbis, a software development company in London.



**Peter McOwan** was born in the United Kingdom in 1962. He received the B.Sc. degree in physics from Edinburgh University, Edinburgh, U.K., in 1984, the M.Sc. degree in medical physics from Aberdeen University, Aberdeen, U.K., in 1985, the Ph.D. degree in computer generated holography from King's College London, London, U.K., in 1990, and the M.Sc. degree in experimental methods in psychology from University College London in 1995.

After holding a Wellcome Trust Mathematical Biology fellowship at University College London, he was a Lecturer in the Department of Cybernetics, University of Reading, Reading, U.K., between 1996 and 1997, a Lecturer in the Department of Mathematics and Computer Science at Goldsmiths College, University of London, between 1998 and 1999, and has since been with the Department of Computer Science, Queen Mary College, London, where he currently holds the position of Reader and is Director of Teaching. His research interests are in the area of biologically inspired computing and cognitive science.

He is a member of the IEE, the Institute of Physics, and the Optical Society of America.