Computers and Intelligent Systems

Florian Zitzelsberger, Michael Shafae:
Markerless Performance Capture of Hand Gestures ...........................................1

Jonathan Blackledge, Eugene Coyle, Niall McCoy, Derek Kearney, Keith Sunderland, and Thomas Woolmington:
Analysis of Wind Velocity and the Quantification of Wind Turbulence in Rural and Urban Environments using the Lévy Index and Fractal Dimension ...........................................7

Vu Van Yem, Nguyen Xuan Quyen, and Kyandoghere Kyamakya:
A Chaotic Modulation Method based on Combination of CPWPM and BPSK for Digital Communication .................................................................18

Jonathan M. Blackledge, Matthew D. Blackledge, and Jane. N. Courtney:
Non-Gaussian Anisotropic Diffusion for Medical Image Processing using the OsiriX DICOM……24

Ashish Patel, Michael Shafae:
A Cake Cutting Approach to Rezoning School Districts..........................................................33

Jonathan Blackledge, Dmitri Dubovitskiy, and Fiona Lyng:
Targeting Cell Nuclei for the Automation of Raman Spectroscopy in Cytology........................42

Santhosh Kumar Kadari, Bala Soujanya Bhupathi Raju, and Nguyen Xuan Quyen:
Digital Image Encryption based on Chaotic Behavior of A Modified Tent Map..................52

Jarmo Siltanen, Jyrki Joutsensalo, Timo Hämäläinen, and Kari Luostarinen:
Least Squares Bandwidth Allocation Scheme for 4G Heterogeneous Networks......................58
Abstract—Performance capture is where computer vision techniques are applied to extract information regarding a performer’s body movements and gestures which are then mapped to a synthetic actor. In this paper, a system for controlling a simple virtual puppet using consumer-grade webcams to capture the performance of an untrained puppeteer’s hand. The system operates in real-time by first identifying the user’s hand based on a skin color mask, tracking the hand’s position using flocks of KLT features, and identifying gestures through a machine learning process. The system does not require any markers on the user’s hand.

Index Terms—Computer vision, image processing, performance capture, gesture recognition, hand tracking, hand detection, skin tone classifier

1 INTRODUCTION

Communicating with digital computers has often meant that the human adapted to the limited input and output capabilities of the machine. Slowly but surely, technology has advanced to give more natural and intuitive means for us to express our wishes to our computers. Especially within the realm of entertainment softwares such as games, joysticks and buttons have been the pinnacle of the human-computer interface.

There are entirely new input control methods and technologies, such as the Nintendo WiiMote, Sony PlayStation Move and Microsoft Kinect. The electronic entertainment industry is rethinking the traditional, button and controller based approach to human-machine interaction in video games and entertainment applications. Similarly, more natural user interfaces are heavily researched in the medical and military fields. However, across the board, these new input technologies require very specific and potentially expensive hardware in order to reliably interpret human gestures. Specifically, these devices include sensors like depth-perceptive cameras, gyroscopes, accelerometers, or visual trackers.

This is understandable, considering the difficulties involved with recognizing human gestures, posture and appearance as they are inherently ambiguous and differ between ages and genders. Additionally ethnic and cultural influences add meaning and context to our body language. Moreover, human gestures are not generally associated with the same meaning across all cultures and social settings. Nodding one’s head, for example, is interpreted as a form of acknowledgement by most, but not all cultures. Direct eye contact may be interpreted as a form of attentive listening, but may easily be understood as a provocative gesture.

In this study, we explore the possibilities of using inexpensive and readily available hardware, such as a webcam and a desktop or laptop computer, as a means of detecting and interpreting natural hand gestures to control a virtual puppet. The objective is to identify a potential hand object in a scene, track the position and motion of the hand object, extract gesture information and apply the gesture’s meaning to manipulate a virtual puppet. Most importantly, the entire process is to be done without markers to aid in recognizing a hand object.

This is rather difficult, considering the limitations of computer vision. As visual creatures, we perceive images fundamentally differently from the way a computer perceives an image. “The human brain divides the vision signal into many channels that stream different kinds of information into your brain. Your brain has an attention system that identifies [...] important parts of an image to examine while suppressing examination of other areas. There is massive feedback in the visual stream that is, as yet, little understood." A computer, on the other hand, sees an image as merely a set of discrete samples, potentially falsified by a noise component. Moreover, where as human eyes feed our brains a wealth of information instantaneously, consumer-quality webcams are far inferior. Webcams do not feature automatic control of focus and aperture and colors are severely distorted under varying lighting conditions. Most importantly, there is no automatic pattern recognition or a database of years of experience in identifying objects and extracting important information from this stream of images.
2 PREVIOUS WORK

With respect to computer vision, the primary focus is not a novel computer vision algorithm, but rather to combine a set of existing algorithms to identify a hand object, track the hand object and to extract gestures effectively in real-time. Furthermore, the identification, tracking and gesture extraction is to be done without the use of artificial markers. The problem domain can be divided into three subproblems:

1) Hand Detection
2) Hand Tracking
3) Gesture Recognition

2.1 Hand Detection

One of the most common approaches to solve the hand detection problem consist of building a skin color mask from the input image. In order to build this mask, it is often desirable to convert the input image to a lighting condition agnostic color space. This is a color space which would allow us to separate the lighting and color information from one another. The hue, saturation, and value (HSV) or hue, saturation, and intensity (HSI) color spaces, both a derivation of the RGB color model satisfy this requirement. With HSV, variations in lighting are largely reflected in the saturation component. Sometimes, the normalized RGB or RG chromaticity color spaces may be used to achieve similar results. The normalized components \( R_n \), \( G_n \), and \( B_n \) are computed as follows:

\[
Base = R + G + B \\
R_n = \frac{R}{Base} \\
G_n = \frac{G}{Base} \\
B_n = \frac{B}{Base} = 1 - R_n + G_n
\]

Note, that only \( Base \), \( R_n \) and \( G_n \) must be stored and that \( B_n \) may be derived. If \( Base \) is omitted, the conversion becomes lossy, as it is usually the case with RG chromaticity, a variation of the normalized RGB model.

Once the saturation component has been separated or removed from the input image, a skin color mask may be created by using histogram back-projection or thresholding. With histogram back-projection, a previously created histogram of skin tones is back-projected onto the input image, resulting in a distribution probability image. A distribution probability image’s pixel’s color value reflects the probability of the corresponding pixel in the input image being within the range of the histogram. This approach can be further improved by using an adaptive model, such that the color histogram is created and updated on-the-fly accounting for possible changes in the illumination conditions of the scene. One such approach is detailed in [2].

Another successful approach to creating a skin color mask consists of filtering the input image by a range threshold. This approach results in a binary skin color mask. The sensitivity of this classifier may be adjusted by changing the minimum and maximum values of the threshold applied. These minimum and maximum values are often empirically determined and may be adapted over time. This idea of an adaptive skin color classifier using a threshold is explained in [3].

An entirely different approach to object recognition involves using a Haar cascading classifier. This approach is now widely used in face detection and was first introduced by Viola and Jones [4]. The idea is based on Haar wavelets and uses black and white rectangles as illustrated in Figure 1. Haar features are calculated by subtracting the sum of pixels within the white regions from the sum of pixels in the black region of the black and white rectangle.

This computation can be done in constant time, by first generating the integral image from the input image. Essentially, the value at any given pixel \( ii(x, y) \) on the integral image indicates the sum of all the pixel values preceding \( ii(x, y) \):

\[
ii(x, y) = \sum_{x' \leq x} \sum_{y' \leq y} i(x', y')
\]

Once the integral image has been generated, no more than four array lookups are required to compute the sum of any given rectangle on the input image. As described in [4], the Viola-Jones object detection framework moves a window of a specified size across the input image, calculating the Haar features for every subregion within that window. It then compares the value of the feature against a database of learned objects and non-objects. If a given region falls below the threshold of being classified as an object, early rejection of the window results in a dramatic performance gain. Only if subregions of the window are classified as objects, further processing is done on the window. It must be noted that in order for the Viola-Jones approach to work reliably, a considerably large, trained database of objects and non-objects must be available.

2.2 Hand Tracking

Once the hand detection phase concludes, the user’s hand is tracked in order to map it’s movement to the virtual puppet’s screen-space position. Several algorithms exist for this purpose, most notably Mean-Shift and Optical Flow.

The Mean-Shift algorithm is based on the idea of finding the maxima of a density function, represented
by a number of discrete samples. The algorithm moves a convolution kernel over a predefined search window, detecting maxima within the image region covered by the kernel, and ultimately the search window. The algorithm returns the mode(s) of the density map.

Continuously Adaptive Mean-Shift (CamShift) is an adaptation of the Mean-Shift algorithm, most commonly used in face, head and object tracking. Instead of using a static, precomputed probability distribution, CamShift employs an adaptive approach, by recomputing the probability distribution after every frame. With CamShift, the probability distribution image is commonly generated by creating a skin probability mask via histogram back-project, as explained above. For a more detailed explanation of Mean-Shift and CamShift in particular, see [5].

Optical Flow is not an algorithm per se. The term generally describes a pattern of motion in a stream of images. However, a number of algorithms exist to measure optical flow. These algorithms are based on phase correlation, which can be measured efficiently using a fast Fourier transform. However, some of the most popular algorithms to measure optical flow in computer vision are based on differential methods. One example for such a method is Lucas-Kanade Optical Flow. Lucas-Kanade Optical Flow is an iterative approach, which uses spatial intensity gradient information. This idea is explained in [9] and a Kanade-Lucas-Tomasi (KLT) feature tracker is introduced in [7].

### 2.3 Gesture Recognition

A simple sock puppet can be created by placing one’s hand in a sock. By pinching one’s fingers together the sock puppet’s mouth closes. By opening one’s hand, the sock puppet’s mouth closes. Intuitively, a hand gesture manipulates the sock puppet to take on different forms which in turn are interpreted as such ideas as talking, nodding, smiling, grimacing, etc. In order to achieve the same ends with a virtual puppet, the user’s gesture’s must be first detected and then mapped onto an action the virtual puppet must take.

Doing hand gesture recognition often involves a creative approach to pattern recognition or template matching. A number of algorithms exist in this field, often involving a machine learning process for training a database with recognizable gestures. A comparison of hand gesture recognition techniques is given in [8].

One of the most straight-forward ways of performing template matching is to use the subtraction method. This method involves converting an image of the hand object into a binary mask (for example from the skin color mask), then comparing the hand mask against a set of gesture masks saved in a trained database. In order to perform this comparison, the pixel values on the hand mask are subtracted from the corresponding pixel values on the stored gesture masks. Then, the euclidean distance $d$ is computed. These steps are repeated for every gesture stored in the database and the mask with the smallest euclidean distance is picked as the closest match.

$$d = \sqrt{\sum_{x \leq \text{width}} \sum_{y \leq \text{height}} (i_{\text{mask}}(x, y) - i_{\text{gesture}}(x, y))^2}$$

Another approach has its roots in Principle Component Analysis (PCA). Besides being widely used in image compression, one of the most popular implementations of PCA is Eigenfaces. Although generally used as a face recognition technique, Eigenfaces can be adapted to recognize other features, such as hand gestures. The general idea is to look at all the gesture masks stored in the training set, then generate a mean image as well as and alternative representations as Eigenvectors from the data set. The Eigenvectors will serve as the principle components in this implementation. Every mask in the training set will then be expressed as a vector $v$ of dimension $n$, where $n$ be the number of principle components and the value of the vector component $x$, $0 < i < n$ be an index for similarity between the mask and the respective principle component. The hand mask extracted from the current video frame is then also converted into such a vector and compared against the vectors in the training set. The stored gesture mask with similarity indices closest in value to the current hand mask will then be identified as the recognized gesture.

An alternative method of gesture recognition involves finding the contours of the detected hand object. The positions of all visible fingertips may then be extracted by converting the contours to a convex hull and looking for start and end points of convexity defects. Based on how many fingertips have been found and their distance to each other, some specific hand gestures may then be detected without prior training of a database. This approach is also robust in that it is largely agnostic to the orientation of the hand.

### 3 Implementation

For our particular implementation we needed to select an appropriate algorithm for each of the three steps involved in extracting the necessary information from the input image stream. Our general requirements were as follows:

- **Performance:** the image analysis is required to conclude within less than 30 milliseconds. This leaves us enough frame time to further draw the virtual puppet, while maintaining a minimum frame rate of 15 frames per second.
- **Flexibility:** the implementation needs to be flexible enough to work under somewhat unfavorable lighting conditions, with less than perfect background colors, and with a wide variety of skin tones. Moreover, the system must function well even with a low-quality webcam.
3.1 Hand Detection

In order for our hand detection algorithm to provide the necessary flexibility as proposed in Section 3, we would like to employ an adaptive approach. Due to its ease of implementation and overwhelmingly great results in a practical environment, we chose to implement the Adaptive Skin Color Classifier as proposed in [3], operating on the normalized RGB color space. In order to initialize the Skin Color Classifier, we use a Haar Cascading Classifier as proposed by Viola and Jones [4] to first create a bounding box containing a face found in the scene. Second, a number of discrete skin color samples are taken from the face using an empirically determined mask. Figure 2 illustrates this idea.

The image processing core of our application is organized as a deterministic finite state machine. Once the lower and upper bounds for our skin color mask have been determined we transition to the hand tracking state so to not waste any more processing resources on the skin color threshold computation. However, in order to compensate for drastically changing lighting conditions we monitor the average white coverage of the binary skin color mask. If the mask contains an unreasonable amount of skin color, such as too much or too little white coverage, our image processing core transitions back into the hand detection state, resampling the image for skin colors.

3.2 Hand Tracking

Besides the requirements mentioned in Section 3, our hand tracking algorithm needs to be able to cope with moderately fast movements and the motion blur introduced by such movements. Unfortunately, CamShift is known to be somewhat susceptible to this particular problem. However, we have found the algorithm explained in [9] to perform very well in our case. The idea is based on using flocks of KLT features, collected around the center of the current hand bounding box. Optical flow is measured by a differential algorithm in order to determine the positions of these KLT features in the next frame. If a feature ceased to exist or is now outside the area of the bounding box, a new KLT feature close the the center is picked. Figure 3 illustrates this approach.

3.3 Gesture Recognition

Since our application has fairly specific requirements in terms of gesture recognition, we opted for a more specialized, but easy to implement approach. For this research, we require to primarily detect mouth opening and mouth closing gestures in order to control the virtual puppet. To do so we find the contours of the hand on the binary skin mask, using the tracking bounding box. Figure 4 illustrates this idea. The hand contours are shown as red lines. Once the contours have been found, we convert them into a convex hull. Using the derived convex hull and the original contours, we then search for convexity defects along the contours. As a result, we can derive information about the spaces between the fingers. The depth of the convexity defect gives us an idea of how open or closed the hand is. Moreover, we can estimate
how many fingers are shown, by counting the number of
detected convexity defects, which are the spaces between
the fingers.

From the information we have about the convexity
defects, we can derive a mouth closing and opening
gesture: If one or less fingers are detected, we consider
the mouth closed. If three or more fingers are shown,
the mouth is considered open. We can also derive an
openness factor from computing the mean of convexity
defect depths.

Since based on image noise and quality of the skin
color mask, recognition errors may be made, we can use
a simple pole-filter, such as a low-pass filter, in order to
derive a more stable value for the openness factor.

4 Future Work

Our application suffers from a handful of issues, which
may be solved in a future revision. For example, one of
the problems with our current implementation consist of
the user’s face being part of the skin color mask. This is
especially problematic if the tracked hand moves closer
to the user’s face. In that case, parts of the face might
erroneously be considered part of the user’s hand. KLT
features are then also searched for on the face, which
makes the tracking become somewhat unreliable.

Moreover, parts of the background may falsely be
identified as skin colored pixels, if some background
features are similar in color to human skin. This is es-
pecially problematic under imperfect lighting conditions
or with poor quality webcams. An adaptive background
segmentation algorithm could potentially alleviate this
problem. This algorithm would keep a history of video frames and consider rarely changing pixels as being part
of the background.

Much better image segmentation results may addition-
ally be achieved by using a time-of-flight (TOF) depth-
perceptive camera, such as the Microsoft Kinect.

References

“Skin detection in video under changing illumination conditions,” in Proceedings. 15th International Conference on Pattern Recognition,
and Image Processing GVIP 05, A. A. et al., Ed., vol. 5, International
Congress for Global Science and Technology. Cairo, Egypt: International Congress for Global Science and Technology, 2005,
pp. 324–327.
cascade of simple features,” in Proceedings of the 2001 IEEE Computer
algorithm and multiple quantized feature spaces,” in Proceedings of
the Pan-Sydney area workshop on Visual information processing, ser.
VIP ’05. Darlinghurst, Australia, Australia: Australian Computer
 technique with an application to stereo vision,” in Proceedings of
IEEE Computer Society Conference on Computer Vision and Pattern
Recognition, June 1994, pp. 593–600.
[8] P. Chakraborty, P. Sarawgi, A. Mehrotra, G. Agarwal, and R. Prad-
Abstract—This paper is concerned with a quantitative and comparative analysis of wind velocities in urban and rural environments. It is undertaken to provide a route to the classification of wind energy in a rural and urban setting. This is a common problem and the basis of a significant focus of research into wind energy. In this paper, we use a non-Gaussian statistical model to undertake this task, and, through a further modification of the data analysis algorithms used, extend the model to study the effect of wind turbulence, thereby introducing a new metric for this effect that is arguably superior to a more commonly used and qualitatively derived measure known as the Turbulence Intensity.

Starting from Einstein’s evolution equation for an elastic scattering process, we consider a stochastic model for the wind velocity that is based on the Generalised Kolmogorov Feller Equation. For a specific ‘memory function’ - the Mittag-Leffler function - it is shown that, under specified conditions, this model is compatible with a non-Gaussian processes characterised by a Lévy distribution that, although previously used in wind velocity analysis, has been introduced phenomenologically. By computing the Lévy index for a range of wind velocities in both rural and urban environments using industry standard cup anemometers, wind vanes and compatible data collection conditions (in terms of height and sampling rates), we show that the intuitive notion that the ‘quality’ of wind velocity in an urban environment is poor compared to a rural environment is entirely quantifiable. This quantifies the notion that a rural wind resource is, on average, of higher yield when compared to that of the urban environment in the context of the model used. In this respect, results are provided that are based on five rural and urban locations in Ireland and the UK and illustrate the potential value of the model in the consideration of locating suitable sites for the development of wind farms (irrespective of the demarcation between an urban and rural environment). On this basis, the paper explores an approach whereby the same model is used for evaluating wind turbulence based on the Fractal Dimension using the ‘polar wind speed’ obtained from three-dimensional data sets collected in urban environments.

Index Terms—Wind velocity, wind turbines, non-Gaussian statistics, Lévy index, rural and urban analysis, wind turbulence, Fractal Dimension.

I. INTRODUCTION

A primary factor in the development of a wind farm is an understanding of the potential wind energy associated with the site, i.e. the geographical location of the farm, [1], [2]. This is the key to the economic viability of any wind energy project which must focus on the development of wind farms in effective and efficient regions, subject to the structural and environmental conditions that provide an optimum solution within the engineering and commercial constraints imposed, [3], [4]. Understanding the relationship between on-shore rural and urban environments (and off-shore wind energy schemes), has been, and remains fundamental in the development of the wind industry throughout the world, [5]. For some time now, it has been ‘understood’ by industry experts that the wind velocity in a rural environment is of a higher ‘quality’ and energy yield when compared to the wind velocity in an urban environment, [6]. The term ‘understood’ is often taken for granted, rather than taking all of the facts into consideration and fully justifying the actual results, [7], [8]. In this context, the purpose of this paper is to look at using recently developed stochastic models (originally developed for algorithmic financial trading and used to launch and develop http://tradersnow.com, for example) to investigate a possible correlation between the wind velocity in both rural and urban environments based on a statistical parameter called the Lévy index. This represents a significant departure from conventional statistical analysis of wind velocity data which is typically based on Gaussian-type models where the wind velocity is taken to be a Rayleigh-type distributed. The statistical model considered in this paper is non-Gaussian and is used to provide two distinctive and original contributions:

- quantification of the intuitive understanding that potential wind energy is less in urban regions based on computing the Lévy index;
- quantification of wind turbulence in terms of a new metric that is arguably superior to the conventional Turbulence Intensity based on computing the Fractal Dimension (which is simply related to the Lévy Index).

II. FUNDAMENTALS OF WIND ENERGY

The power generated by a wind turbine is based on a range of design factors but they all relate indirectly to Betz law, which states that the power $P$ in Watts is given by, [9]

$$P = \frac{1}{2} \alpha \rho A v^3$$

(1)
where \( v \) is the upwind speed (i.e. the wind velocity that is incident on the turbine) in metres per second (\( \text{ms}^{-1} \)), \( A \) is the area mapped out by the turbine blades in \( \text{m}^2 \), \( \rho \) is the density of air in \( \text{kgm}^{-3} \) and \( \alpha \approx 0.593 \) is the coefficient of performance. This result is derived by considering the energy generated by a change in the upwind and downwind velocities together with the change in the mass of air that occurs as it ‘travels’ through the turbine. A derivation of equation (1) is given in [10], for example, which includes the idealised conditions upon which this equation is based. This law is the ‘design guide’ associated with the development of wind turbines world-wide. There are two important factors associated with optimising the output power: (i) the diameter of the turbine blades; (ii) the wind velocity which scales of the velocity cubed. The cubic velocity scaling law cannot be maintained in practice over all ranges of wind velocity, and, depending upon the design characteristics of the turbine, there is a natural threshold for the wind velocity beyond which the output power does not increase. This is due to a range of influencing factors including the turbulence phenomena that occur at high wind velocities through interaction with the turbine blades when Betz law breaks down. However, within the framework of Betz law, and given a turbine with a fixed blade diameter, the velocity cubed scaling law is of fundamental importance in determining the output power. Clearly, the wind velocity is time dependent and this dependence cannot, in general, be classified in a deterministic sense. Stochastic models that lead to the design of statistical data analysis algorithms are therefore required that ideally provide a statistical parameter or parameter set that can quantify the wind resource subject to a range of influencing factors.

### III. INFLUENCING FACTORS

There are a range of influencing factors that can affect the performance of a wind turbine and a wind farm. The current industry knowledge is based on the ‘roughness principle’. Commonly found in the rural environment, there tends to be little in the way of large obstacles to cause sufficient turbulence which affect the wind quality, and, in turn, the energy yield of wind turbines. This is due to the relatively laminar flow that is a characteristic of a ‘good site’ as illustrated in Figure 1.

By contrast, in the urban environment, there is generally an abundance of built obstacles representing adverse roughness of the ‘ground truth’ generating turbulence and thereby curtailing the potential energy yield and the output power of wind turbines located in such an environment as illustrated in Figure 2.

Understanding the ‘quality of the wind velocity’ is of particularly interest to those in the wind energy industry, as it allows the developer to identify specific sites to develop, concentrating on which sites produces the greatest energy yields. Within the urban environment, there are numerous factors influencing the wind velocity. The overriding factor of the built environment in the urban setting is that of roughness. There are also the numerous properties of the urban environment and atmospheric influences to be accounted for resulting in an extremely complex environment to accurately model.

Referring to the idealised model illustrated in Figure 3, a standard scaling law for the effect of roughness on the wind velocity at a height \( z \) is given by, [11]

\[
v(z) = \frac{v_f}{\kappa} \log \left( \frac{z - d}{z_0} \right)
\]

(2)

where \( v(z) \) denotes the wind velocity at a height \( z \), \( v_f \) is the friction velocity (which is dependent on the roughness of the ground), \( \kappa \) is the (dimensionless) von Karman constant (typically of the order of 0.41), \( z \) is the height above the earth’s surface, \( d \) is the displacement height and \( z_0 \) is the height above the earth’s surface roughness (where the wind velocity appears to approach zero). The friction velocity depends upon the shear stress \( T \) at the boundary of the flow and is given by, [12]

\[
v_f = \sqrt{T/\rho}, \text{ where } \rho \text{ is the density of air.}
\]

This log-based
scaling law describes the velocity profile of a turbulent fluid flow near a boundary with a non-slip condition and must be fully and accurately introduced when taking into account wind gusts and strong winds (especially in urban environments). Equation (2) is a semi-empirical relationship used to describe the vertical distribution of horizontal mean wind speeds within the lowest portion of the boundary layer.

As the wind speed decreases to zero, closer to ground level, this results in an atmospheric boundary layer. Thus, equation (2) can be accurate up to 200 m. In rural environments, however, the impact of such ‘roughness factors’ are less common and mostly attributed to forestry etc. (as illustrated in Figure 3) which can be removed as required and with the required permission of the appropriate environmental agencies.

![Rural (top) and Urban (bottom) roughness models associated with equation (2)](source: Mertens 2006)

It is intuitively obvious that, whatever the cause, turbulence reduces the energy output from a wind turbine since turbulence dissipates energy over a larger volume (at least for an adiabatic system). It is also clear that turbulence is extremely difficult to model in a fully deterministic sense, based on the principles of fluid dynamics. Thus, in the following section we develop a stochastic model from first principles.

IV. STOCHASTIC MODEL FOR THE WIND VELOCITY

We consider the temporal behaviour of the wind velocity in terms of the space-time evolution of a field $v(x,t)$ working in one-dimension. The temporal behaviour of the wind velocity is then taken to be the time dependent behaviour of this field at a point in space $x$. In [10] and [13], a stochastic model for the wind velocity is developed using a fractional partial differential equation of the type

$$\frac{\partial^\gamma}{\partial x^\gamma} v(x,t) - \frac{\partial}{\partial t} v(x,t) = -\delta(x)s(t), \quad \gamma \in (0, 2]$$

where $\gamma$ is the Lévy index and $s(t)$ is a ‘white noise’ stochastic ‘source function’ with a uniformly distributed Power Spectral Density Function (PSDF) and arbitrary Probability Density Function (PDF). Ignoring scaling constants, it is shown that the Green’s function solution to this equation is

$$v(t) = \frac{1}{t^{1-1/\gamma}} \otimes_t s(t)$$

where $\otimes_t$ denotes the convolution integral over $t$ and $v(t) \equiv v(0, t)$. This solution has the self-affine scaling relationship

$$\Pr[v(\alpha t)] = a^{1/\gamma} \Pr[v(t)]$$

where $\Pr$ denotes the PDF and a PSDF given by (for scaling constant $c$)

$$|V(\omega)|^2 = \frac{c}{|\omega|^{2/\gamma}} \quad \text{where} \quad V(\omega) = \int_{-\infty}^{\infty} v(t) \exp(-i\omega t) dt$$

Following [14] and [15], we now consider an extension and generalisation to this model which is based on developing a solution to the Generalised Kolmogorov-Feller Equation (GKFE) which is derived in the following section. The aim is to show that the solution to the GKFE considered provides a model for the PSDF that is effectively the same as that considered in [10] and [13] and that the wind velocity field can be considered to by a random scaling fractal signal characterised by a Lévy index. In turn, this index is related to the fractal dimension of the signal, and, as discussed later on in this paper, this dimension can be used to characterise the turbulent behaviour of the wind velocity providing an index that is arguably superior to the conventional Turbulence Intensity.

A. Derivation of the Generalised Kolmogorov-Feller Equation

For an arbitrary PDF $p(x)$, Einstein’s evolution equation is, [16]

$$u(x, t + \tau) = u(x, t) \otimes_x p(x)$$

where $u(x,t)$ is a ‘density function’ representing the concentration of a canonical ensemble of particles undergoing elastic collisions. This function is interpreted as a field representing the distribution of physical properties such as the mass, velocity, temperature and pressure, for example.

Consider a Taylor series for the function $u(x,t+\tau)$, i.e.

$$u(x, t + \tau) = u(x, t) + \tau \frac{\partial}{\partial t} u(x, t) + \frac{\tau^2}{2!} \frac{\partial^2}{\partial t^2} u(x, t) + ...$$

For $\tau << 1$

$$u(x, t + \tau) = u(x, t) + \tau \frac{\partial}{\partial t} u(x, t)$$

and we obtain the ‘Classical KFE’, [17] and [18]

$$\tau \frac{\partial}{\partial t} u(x, t) = -u(x, t) + u(x, t) \otimes_x p(x)$$

Equation (4) is based on a critical assumption which is that the time evolution of the field $u(x,t)$ is influenced only by short term events and that longer term (historical) events have no influence on the behaviour of the field, i.e. the ‘system’ described by equation (4) has no ‘memory’. This statement is the physical basis upon which we introduce the condition $\tau << 1$ thereby allowing the Taylor series expansion of the $u(x, t+\tau)$ to be made to first order. The question then arises as to how longer term temporal influences can be modelled, other
than by taking an increasingly larger number of terms in the Taylor expansion of $u(x, t + \tau)$ which is not of (closed-form) analytical value.

For arbitrary values of $\tau$,

$$\tau \frac{\partial}{\partial t} u(x, t) + \tau^2 \frac{\partial^2}{\partial t^2} u(x, t) + \ldots = -u(x, t) + u(x, t) \otimes_x p(x)$$

We model the effect on a solution for $u(x, t)$ of the series on the left hand side of this equation in terms of a ‘memory function’ $m(t)$ and write

$$\tau m(t) \otimes_t u(x, t) = -u(x, t) + u(x, t) \otimes_x p(x)$$

(5)

where $\otimes_t$ is taken to denote the causal convolution integral over $t$. This is the Generalised KFE (GKFE) which reduces to the Classical KFE when

$$m(t) = \delta(t)$$

Note that for any memory function for which there exists a function or class of functions of the type $n(t)$, say, such that

$$n(t) \otimes_t m(t) = \delta(t)$$

then we can write equation (5) in the form

$$\tau \frac{\partial}{\partial t} u(x, t) = -n(t) \otimes_t u(x, t) + n(t) \otimes_t u(x, t) \otimes_x p(x)$$

(6)

where the Classical KFE is recovered when $n(t) = \delta(t)$.

Any solution obtained to the GKFE will be dependent upon the choice of memory function $m(t)$ used. There are a number of choices that can be considered, each or which is taken to be a ‘best characteristic’ of the stochastic system in terms of the influence of its time history. However, it may be expected that the time history of physically significant random systems is relatively localised in time. This includes memory functions such as the Mittag-Leffler function [19]

$$m(t) = \frac{1}{\Gamma(1 - \beta)t^{\beta}}, \quad 0 < \beta < 1$$

where

$$n(t) = \frac{1}{\Gamma(\beta - 1)t^{2 - \beta}}$$

given that

$$\int_0^\infty \frac{\exp(-st)}{\Gamma(\beta) t^{1 - \beta}} \, dt = \frac{1}{s^{\beta}} \quad \text{and} \quad \int_0^\infty \delta(t) \exp(-st) \, dt = 1$$

B. Green’s Function Solution to the GKFE

Equation (6) can be written in the form

$$\tau \frac{\partial}{\partial t} u(x, t) + u(x, t) = u(x, t) - n(t) \otimes_t u(x, t)$$

$$+ n(t) \otimes_t u(x, t) \otimes_x p(x)$$

so that the Green’s function solution is given by

$$u(x, t) = g(t) \otimes_t u(x, t) - g(t) \otimes_t n(t) \otimes_t u(x, t)$$

$$+ g(t) \otimes_t n(t) \otimes_t u(x, t) \otimes_x p(x)$$

(7)

where the Green’s function is given by

$$g(t) = \frac{1}{\tau} \exp(-t/\tau), \quad t > 0$$

which is the solution to

$$\tau \frac{\partial}{\partial t} g(t) - g(t) = \delta(t)$$

and we assume the initial conditions $u(x, t = 0) = 0$ and $g(t = t_0) = 0$. We can now analyse this solution in Fourier-Laplace space by taking the Fourier transform and the Laplace transform of equation (7) and using the convolution theorems for the Fourier and Laplace transform, respectively, to obtain

$$\tilde{u}(k, s) = \bar{g}(s)\tilde{u}(k) - \bar{g}(s)\tilde{n}(k)\tilde{u}(k) + \bar{g}(s)\tilde{n}(k)\tilde{u}(k)\tilde{p}(k)$$

(8)

where

$$\tilde{u}(k, s) = \int_0^\infty \int_{-\infty}^{\infty} u(x, t) \exp(-ikx) \, dx \exp(-st) \, dt$$

$$\bar{g}(s) = \int_0^\infty g(t) \exp(-st) \, dt$$

$$\tilde{n}(s) = \int_0^\infty n(t) \exp(-st) \, dt$$

$$\tilde{p}(k) = \int_{-\infty}^{\infty} p(x) \exp(-ikx) \, dx$$

From equation (8) we can write

$$\tilde{u}(k, s) = -\frac{\bar{g}(s)}{1 - \bar{g}(s)}\tilde{n}(k)\tilde{u}(k) + \frac{\bar{g}(s)}{1 - \bar{g}(s)}\tilde{n}(k)\tilde{u}(k)\tilde{p}(k)$$

$$= -\frac{\tilde{n}(k)}{\tau s}\tilde{u}(k, t) + \frac{\tilde{n}(k)}{\tau s}\tilde{u}(k, t)\tilde{p}(k)$$

given that $\bar{g}(s) = (1 + \tau s)^{-1}$ and thus obtain the equation

$$\tilde{u}(k, s) = \bar{h}(s)\tilde{u}(k, s)\tilde{p}(k)$$

where

$$\bar{h}(s) = -\frac{\bar{n}(s)}{\tau s + \bar{n}(s)}$$

or, upon inverse transformations

$$u(x, t) = h(t) \otimes_t u(x, t) \otimes_x p(x)$$

(9)

with

$$h(t) \leftrightarrow \frac{\bar{n}(s)}{\tau s + \bar{n}(s)}$$

where $\leftrightarrow$ denotes the Laplace transformation, i.e. mutual transformation from $t$-space to $s$-space.

Consider the iteration of equation (9) defined by

$$u_{n+1}(x, t) = h(t) \otimes_t u_n(x, t) \otimes_x p(x)$$

(10)

for an initial solution $u_0(x, t)$ where $n = 1, 2, \ldots, N$ The equivalent iteration in Fourier-Laplace space is

$$\tilde{u}_{n+1}(k, s) = \bar{h}(s)\tilde{u}_n(k, s)\tilde{p}(k)$$

with initial solution $\tilde{u}_0(k, s)$ so that, after $N$ iterations,

$$\tilde{u}_N(k, s) = [\bar{h}(s)]^N [\tilde{p}(k)]^N \tilde{u}_0(k, s)$$
and upon inverse Fourier-Laplace transformation, has the iterative form

\[ u_N(x, t) = \prod_{j=1}^{N} p(x) \prod_{k=1}^{N} h(t) \otimes_x \otimes_t u_0(x, t) \quad (11) \]

where

\[ \prod_{j=1}^{N} f(t) \equiv f(t) \otimes_t f(t) \otimes_t \ldots \]

denoting the \(N\)th convolution of \(f(t)\).

The criterion for convergence of this (iterative) solution can be considered by introduction of the error function \(e_n(x, t)\) at any iteration \(n\) so that \(u_n(x, t) = u(x, t) + e_n(x, t)\). From equation (10) we can then write (transformation to Fourier-Laplace space)

\[ \tilde{e}_{n+1}(k, s) = \tilde{h}(s) \tilde{p}(k) \tilde{e}_n(k, s) \]

so that

\[ \tilde{e}_n(k, s) = [\tilde{h}(s) \tilde{p}(k)]^n \tilde{e}_0(k, s) \]

and it is clear that, since we require \(\tilde{e}_n \to 0\) and \(n \to \infty\), \([\tilde{h}(s) \tilde{p}(k)] < 1 \forall(k, s)\). The condition for convergence therefore becomes

\[ \|\tilde{h}(s) \tilde{p}(k)\| \leq \|\tilde{h}(s)\| \times \|\tilde{p}(k)\| < 1 \]

or, for Euclidian norms, and, using Rayleigh’s theorem,

\[ \|h(t)\|_2 \times \|p(x)\|_2 < \frac{1}{\sqrt{2\pi}} \]

C. Impulse Response for the Mittag-Leffler Memory Function

Form equation (11), if the initial solution is an impulse (i.e. \(u_0(x, t) = \delta(x) \delta(t)\)) then the Impulse Response Function (IRF), denoted by \(r(x, t)\), is given by

\[ r(x, t) = \prod_{j=1}^{N} p(x) \prod_{k=1}^{N} h(t) \]

with ‘transfer function’

\[ \tilde{r}(k, s) = [\tilde{h}(s) \tilde{p}(k)]^N \]

For a memory function \(m(t)\) modelled by the Mittag-Leffler function (for \(0 < \beta < 1\))

\[ m(t) \leftrightarrow \frac{1}{s^{1-\beta}} \quad \text{and} \quad \tilde{h}(s) = \frac{1}{1 + \tau s^\beta} \sim \frac{1}{\tau s^\beta} \]

so that

\[ h(t) \sim \frac{1}{\Gamma(\beta) t^{1-\beta}} \]

Similarly, if we consider a Mittag-Leffler PDF of the form

\[ p(x) = \frac{1}{\Gamma(1 - \gamma) |x|^{\gamma}}, \quad 0 < \gamma < 1 \]

then the IRF becomes

\[ r(x, t) \sim \prod_{j=1}^{N} \frac{1}{\Gamma(1 - \gamma) |x|^{\gamma}} \prod_{k=1}^{N} \frac{1}{\Gamma(\beta) t^{1-\beta}} \]

D. Temporal IRF for Early Evolutionary Behaviour

The function \(r(x, t)\) is a space-time IRF. A temporal IRF can be considered by integrating over \(x\). Physically, the resulting IRF can be taken to be a characteristic of a time series recorded at an arbitrary point in space. Further, if we consider the early evolutionary behaviour of \(u_N(x, t)\) (i.e. the case when \(N = 1\)), we obtain the simplified expression for the field \(u(t)\) given by

\[ u(t) \equiv \int_{-\infty}^{\infty} u_1(x, t) dx = \frac{1}{\tau \Gamma(\beta) t^{1-\beta}} \otimes_t s(t) \quad (12) \]

where

\[ s(t) = \int_{-\infty}^{\infty} p(x) \otimes_x u_0(x, t) dx \]

This result demonstrates that the model developed in [10] and [13], where the wind velocity is given by equation (3), is a special case of the solution to the GKFE considered here (i.e. equation (12) where, ignoring scaling constants, \(\beta = \gamma^{-1}\) and the field \(u\) is taken to be the wind velocity) and describes the early evolution of a time series governed by the GKFE. In turn, the GKFE is an expression of Einstein’s evolution equation subject to a specialised Memory Function - the Mittag-Leffler function - which yields the fractional diffusion equation. This relationship is compounded further in the following analysis for the case when \(\tau < 1\). Using the convolution theorem, in Fourier space, Einstein’s evolution equation is

\[ \tilde{u}(k, t + \tau) = \tilde{u}(k, t) + \tau \frac{\partial}{\partial t} \tilde{u}(k, t) = \tilde{u}(k, t) \tilde{p}(k), \quad \tau < 1 \]

If we now consider Lévy’s characteristic function (for constant \(a\) and Lévy index \(\gamma\))

\[ \tilde{p}(k) = \exp(-a |k|^{\gamma}) \sim 1 - a |k|^{\gamma}, \quad a < 1 \]

then it is clear that we can write the evolution equation (in Fourier space) as

\[ \tau \frac{\partial}{\partial t} \tilde{u}(k, t) = -a |k|^{\gamma} \tilde{u}(k, t) \]

Using the convolution theorem again, and, together with the Reisz definition of a fractional derivative, i.e.

\[ \frac{\partial^{\gamma}}{\partial x^{\gamma}} f(x) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} |k|^{\gamma} \tilde{f}(k) \exp(i k x)dk \]

we can write

\[ \frac{\partial^{\gamma}}{\partial x^{\gamma}} u(x, t) - \sigma \frac{\partial}{\partial t} u(x, t) = 0 \]

where \(\sigma = \tau/a\). This is a fractional differential operator that has a the temporal IRF (ignoring scaling constants) \(1/t^{1-\gamma}\). Moreover, in this form, it is clear that for \(\gamma = 1\)

\[ \frac{\partial}{\partial x} u(x, t) = \sigma \frac{\partial}{\partial t} u(x, t) \]

an equation which describes flow in one-dimension subject to the continuity equation. In fluid dynamics, for example, the continuity equation states that, in any steady state process,
the rate at which mass enters a system is equal to the rate at which mass leaves the system and is given by (for a three-dimensional space vector \( \mathbf{r} \))

\[
\frac{\partial}{\partial t} \rho + \nabla \cdot (\rho \mathbf{v}) = 0
\]

where \( \rho(\mathbf{r}, t) \) is the fluid density and \( \mathbf{v}(\mathbf{r}, t) \) is the flow velocity vector field. Thus, for a one-dimensional system characterised by a constant velocity field \( \mathbf{v} \) (which is constant over \( x \) and \( t \)) and a density field \( \rho(x, t) \) we obtain

\[
\frac{\partial}{\partial x} \rho + v \frac{\partial}{\partial t} \rho(x, t) = 0
\]

In this sense, the field \( u(x, t) \) for \( \gamma = 1 \) may be taken to describe the flow of mass subject to a constant fluid velocity \( v = \tau/a \). The case of \( \gamma = 1 \) is therefore representative of a steady state process. Moreover, the PDF associated with this process is a Chauchy function since

\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-a | k |) \exp(ikt) dk = \frac{1}{\pi a^2 + x^2}
\]

Similarly, given that

\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-ak^2) \exp(ikt) dk = \frac{1}{2\pi} \sqrt{\frac{\pi}{a}} \exp\left(-\frac{x^2}{4a}\right)
\]

it is clear that the case when \( \gamma = 2 \) describes a Gaussian system, the field \( u(x, t) \) being the solution to the Classical Diffusion Equation

\[
\frac{\partial^2}{\partial x^2} u(x, t) - \sigma^2 \frac{\partial}{\partial t} u(x, t) = 0
\]

We note that, in general [20],

\[
\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-a | k |^\gamma) \exp(ikt) dk \sim \frac{1}{x^{1+\gamma}}
\]

Thus, in terms of using the field \( u \) to model a single or combined velocity field (such as the polar wind speed discussed in Section VII), on the basis of the physical systems described by the cases when \( \gamma = 1 \) and \( \gamma = 2 \) given above, we can expect that for \( \gamma \in [1, 2] \), larger values of \( \gamma \) correspond to more urbanised environments where wind turbulence (which tends towards fully diffusive behaviours but is still fractionally diffusive according to our model) is greater. This idea appears to be validated in the data analysis associated with the case study discussed in the following section.

V. Data Analysis

On the basis of the stochastic model discussed in the previous section, it is possible to estimate the Lévy index, relatively simply. This is achieved using the PSDF method discussed in [21], for example. It is based on exploiting the basic relationship (which ignores scaling factors) [21]

\[
\frac{1}{t^{\tau-1/\gamma}} \leftrightarrow \frac{1}{| \omega |^{1/\gamma}}
\]

where \( \leftrightarrow \) denotes transformation from real to Fourier space (i.e. \( t \)- to \( \omega \) space). Using the convolution theorem, equations (3) and (11) with \( \beta = \gamma^{-1} \), and ignoring scaling by \( [\tau \Gamma(\beta)]^{-1} \), transform to

\[
\bar{u}(\omega) = \frac{s(\omega)}{| \omega |^{1/\gamma}}
\]

Thus, assuming \( s(\omega) \) is a white noise spectrum that can be taken to be a ‘phase only’ function (with unit amplitude),

\[
| \bar{u}(\omega) |^2 = \frac{1}{| \omega |^{2/\gamma}}
\]

This idealised model for the power spectrum is used to estimate the Lévy index based on standard linear regression methods. For the work reported in this paper, and using a MATLAB7 programming environment, the Orthogonal Linear Regression Method based on the m-code available at [22] is used. We note that the power spectrum of a random scaling fractal signal scales as \( [21] | \omega |^{-(5/2+D)} \) where \( D \) is the fractal dimension. Thus, the relationship between the Lévy Index and the Fractal Dimension is

\[
\frac{1}{\gamma} = \frac{5 - 2D}{4}
\]

To accurately model both urban and rural environments, historical data from five rural and five urban wind measurement sites were used. All measurement devices were located at 50m above local ground level to allow an accurate comparison and the locations spread across Ireland and the UK. The measurement devices were all located on lattice towers of the type shown in Figure 4 with industry standard data loggers to store the data. The raw data sets were taken from their raw 10 minute average from industry standard cup anemometers and wind vanes. All data sets were calibrated by industry professionals and the author’s are in receipt of all relevant test certificates to verify the credibility of the calibrations.

Fig. 4. A typical Met Mast used to record wind velocity data at 50m height with a 10 minute average using industry standard cup anemometers and wind vanes. (Source: Wind Prospect Group, 2012.)

The key factor in determining a possible correlation between rural and urban wind velocity is the use of stochastic modelling. The modelling is often based on a statistical analysis of the available wind velocity data which is used to assess optimum regions for the construction of wind farms. In this

1Much of the specific data is confidential and the exact location of the data sources cannot be mentioned in the paper. However, for creditability reasons, the locations are available on receipt of a non-disclosure agreement between the authors and the reader.
TABLE I
MEAN VALUES OF THE LÉVY INDEX \( \gamma \) FOR FIVE rural sites.

<table>
<thead>
<tr>
<th>Site</th>
<th>Location</th>
<th>Longitude</th>
<th>Latitude</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Newport</td>
<td>51.5853</td>
<td>-2.9796</td>
<td>1.3407</td>
</tr>
<tr>
<td>2</td>
<td>Waterford</td>
<td>52.2527</td>
<td>-7.1256</td>
<td>1.3372</td>
</tr>
<tr>
<td>3</td>
<td>Limerick</td>
<td>52.6438</td>
<td>-6.8051</td>
<td>1.3708</td>
</tr>
<tr>
<td>4</td>
<td>Cavan</td>
<td>53.8995</td>
<td>-6.4059</td>
<td>1.4174</td>
</tr>
<tr>
<td>5</td>
<td>Dundalk</td>
<td>53.3979</td>
<td>-6.8051</td>
<td>1.3708</td>
</tr>
</tbody>
</table>

The complex morphology experienced in an urban environment results in a modified flow and turbulence structure in the urban atmosphere in contrast to the flow over ‘ideal or homogenous’ surfaces [28]. Thus, in [27], for example, it is proposed that the TI can be ‘linked’ to the surface roughness parameter via the following equation

\[
TI = \frac{\sigma_v}{\bar{v}}
\]

where \( d \) is the displacement height, which is taken to be equal to 0.66 of the average building height (denoted by \( z_H \)) and \( z_0 \) is the surface roughness length. This equation is predicated on \( z \) (the observation height) being in excess of the wake diffusion height - \( z^* \), which is taken to be above the surface roughness sub-layer and into the inertial sub-layer as illustrated in Figure 5. This result suggests that there is an increasing level of turbulence with increasing roughness and decreasing height relative to the earth’s surface.

VI. TURBULENCE INTENSITY

Urban wind regimes are characterised as having low wind speeds with more turbulent flows which result in limited energy realisation. Research has shown that the lower mean speeds are linked to the higher surface roughness lengths \( z_0 \) prevalent in urban environments, [23] and [24]. The manifestation of turbulence, however, is less well understood. Turbulent flows can be described as those in which the fluid velocity varies significantly and irregularly in both position and time [25]. While turbulently fluctuating flow impacts directly the design of wind turbines, they also influence the productivity of turbines particularly in areas of complex morphologies.

The Turbulence Intensity (TI) is the most common metric used to quantify the effect of wind turbulence as it is generally more useful to develop descriptions of turbulent in terms of statistical properties [26]. TI is defined in [27] as ‘the ratio of wind speed standard deviation to the mean wind speed,
determined from the same set of measured data samples of wind speed, and taken over a specified time’ and should be considered as the standard deviation of the longitudinal wind speed \( \sigma_v \) normalised with the mean wind speed \( \bar{v} \), i.e.

\[
TI = \frac{\sigma_v}{\bar{v}}
\]

The complex morphology experienced in an urban environment results in a modified flow and turbulence structure in the urban atmosphere in contrast to the flow over ‘ideal or homogenous’ surfaces [28]. Thus, in [27], for example, it is proposed that the TI can be ‘linked’ to the surface roughness parameter via the following equation

\[
TI = \frac{1}{\log \left( \frac{z+d}{z_0} \right)}
\]

where \( d \) is the displacement height, which is taken to be equal to 0.66 of the average building height (denoted by \( z_H \)) and \( z_0 \) is the surface roughness length. This equation is predicated on \( z \) (the observation height) being in excess of the wake diffusion height - \( z^* \), which is taken to be above the surface roughness sub-layer and into the inertial sub-layer as illustrated in Figure 5. This result suggests that there is an increasing level of turbulence with increasing roughness and decreasing height relative to the earth’s surface.

![Fig. 5. Wind Speed in the urban context with respect to the boundary layer transitions.](image-url)
Fig. 6. Typical Effects of Turbulence on Power Curves (Source:[31])

low wind speeds (< 3.5m/s). Micro/small wind turbines are designed to commence generating at such wind speeds and in urban environments, mean wind speeds are characteristically low. Thus there is a lack of confidence in the quantification of TI in these environments. Wind speeds below the cut-in speed of a turbine are normally regarded as being non productive; however, this is not the case. In order to have an average wind speed that equals the cut-in speed of say 3.5m/s some values must be above and below 3.5m/s over a 10 minute window so that the mean is 3.5m/s. The question is how erratic is this deviation from the mean and can it be power productive?

Another issue concerning the evaluation of the TI is the qualitative nature of its definition. Given the theoretical model presented in this paper, in the following section we propose a method for evaluating the turbulence intensity based computing the Fractal Dimension of a time series of two-dimensional velocity data. This approach implies that turbulence (as measured by a statistic computed from a wind velocity field) is a self-affine phenomenon and we refer to this metric as the Fractal Turbulence Intensity. In turn, this metric is related to the Lévy Index used to characterise rural and urban environments via equation (13) which provides a ‘link’ between the approach discussed in Section V and that of the following section given the stochastic model developed in Section IV.

VII. FRACTAL TURBULENCE INTENSITY

Observations are made at two urban locations in Dublin, Ireland. St. Pius X National (Girls) School (Site 1), located in Terenure, Dublin 6W (53°20’15.96"N, 6°18’19.02"W) and Dublin City Council Buildings (Site 2), in Marrowbone Lane, located in Dublin 8 (53°20’15.96"N, 6°17’10.27"W) as shown in Figure 7. Site 2 is located closer to the city centre than Site 1 and is therefore more urbanised with a higher associated roughness length. This Site is also characterised by a higher building density in comparison to Site 1 which has a much lower concentration of buildings. As site 2 is closer to the city centre, the buildings consist mostly of office blocks and high-rise residential building. Buildings in the area often reach heights of 20 m and beyond, with some reaching 25 m with topographical complexities located at all angles relative to the anemometer used to record the wind velocity data. Site 1 has a more consistent building morphology and the anemometer is surrounded by a relatively lower average building height that consists mostly of two-storey residential buildings and vegetation which is also at similar heights - see Figure 7.

Fig. 7. Satellite image of Dublin city showing the relative positions of Sites 1 and 2.

At both sites, high-resolution wind speed measurements are taken with a Campbell Scientific CSAT3 three-dimensional sonic anemometer [34]. The observations are at 10Hz at an associated resolution-between 0.5 and 1.0 mm/s, with data that includes date and time-stamp, wind-speed, wind-direction and standard deviation. The CSAT3 measures wind speed employing a right handed orthogonal coordinate system Three orthogonal wind components, which relate to the three dimensions in space, are each measured. Wind entering straight into the anemometer is from the $x$-direction giving wind velocity component $v_x$; wind approaching from the left of the anemometer is from the $y$-direction giving wind velocity component $v_y$; and, wind advancing upwards from the ground...
is from the $z$-direction generating wind velocity component $v_z$. Thus, effectively, the Easterly, Northerly and vertical components of the wind velocity are $v_x$, $v_y$ and $v_z$, respectively, giving a wind velocity vector field $\mathbf{v} = (v_x, v_y, v_z)$. Measurements of this field are taken to an accuracy of 0.01 m/s at a frequency of 10 Hz over a 40 day period from 4/4/2012 to 15/5/2012. Although, on a theoretical basis the Fractal Dimension of any signal is scale invariant so that the sample rate should not matter, in practice, because the computation of the Fractal Dimension uses a Power Spectral Density Function (as discussed in Section V), high data rates in a given sample subset are required to obtain reasonable accuracy. Since turbulence models in general are based on a 10 minute sampling period bench mark, this period is used to compute the Fractal Dimension on a moving window basis, each window consisting of 6000 samples (10 minutes at 10Hz).

The field used to compute the Fractal Dimension from the three-dimensional data available is given by the following model:

$$u(t) = \sqrt{v_x(t) + v_y(t)}$$

This provides a measure of the ‘polar wind speed’ in the horizontal plane which is taken to be the mid $(x, y)$-plane of the three dimensional data field. Application of a combined wind speed model of this type is significant in the sense that, from a physical viewpoint, a turbulence effect is not compounded in a single wind speed direction, any measure of turbulence ultimately having to rely on some multi-dimensional mapping of a fully three-dimensional physical effect. Computation of the longitudinal TI at low wind speeds can have excessive values. This is due to the asymptotic nature of the formula which makes the TI measurement in urban areas particularly problematic with the standard turbulence model. Firstly it is generally accepted that the standard deviation of wind speeds in an urban area is large due to a increased turbulence. Secondly the average wind speed is considerably lower than that of laminar air flows due to the increased surface roughness. The net result is that the TI becomes asymptotically large as the mean wind speed approaches zero. To compensate for this effect it is possible to filter the data by truncating all values of the TI that exceed 1. Using this approach to filter the TI and the data processing method discussed in Section V to compute the Fractal Dimension (and as detailed further in [10]), Figure 9 compares the TI with the Fractal Dimension, normalisation of the data with respect to null entries resulting in the use of 4502 samples.

These results clearly shows that there is correlation between the TI the Fractal Dimension of the horizontal polar wind speed, although it is noted that the Fractal Dimension which, for a Random Scaling Fractal Signal, is a value $D \in [1,2]$, exceeds the upper bound in an analogues way to when $TI > 1$. Figure 10 shows a scatter-plot of the filtered TI denoted by $TI_f$ and the Fractal Dimension and application of linear regression clearly shows that these metrics are correlated, a correlation that, for this the data considered, is compounded in the equation

$$TI_f = 0.1928D + 0.1385$$

**VIII. CONCLUSION**

It is well known that the differences in wind resource in rural and urban environments are curtailed due to the influencing factors such a surface roughness. The aim of this paper has been twofold: (i) to quantify the differences through determination of the Lévy index; (ii) to investigate use of the Fractal Dimension as a measure of the Turbulence Intensity. In the first case, a direct comparison is considered between the urban and rural wind resources at selected location across Ireland and the UK using similar reference heights and fully calibrated equipment so that there is data consistency within the bounds of the practical constraints associated with the technology used to measure the wind velocities. The results confirm that a rural resource is generally of a higher energy yield when compared to the urban resource at least in terms of the Lévy index as computed from the Power Spectrum. This is compounded in lower values of the Lévy index and, as a first study, paves the way for using this non-Gaussian statistical index to evaluate wind resource in general. With regard to the second principal contribution, the fact that conventional turbulence models cannot cater for erratic low mean wind speeds associated with an urban environment requires quantification of alternatives to be considered as given in Section VI.

The approach reported in Section VI is an alternative way of computing a Turbulence Intensity that has two advantages. First, it is based on a more fundamental concept of turbulence in terms of the model provided in Section IV and a fractal geometric interpretation thereby providing a
greater conceptual understanding of turbulence compared the heuristic conventional definition of the TI; second, the problem associated with asymptotic behaviour, which is characteristic of the conventional definition of the TI, and occurs at relatively low wind speeds, is eradicated. Moreover, it is noted that the Fractal Dimension of the polar wind speed and the filtered TI are correlated thereby providing evidence that the conventional qualitative and quantitative measure of wind turbulence proposed have an underlying connectivity.

The model, methodology and results reported in this paper now require a quantification procedure to be developed in order to assess and predict the power performance of wind turbines in rural environments and the degradation of this performance in urban environments. This is predicated on the basis, that, with respect to turbulence assessment, the significant reduction in processing overheads associated with computing the Fractal Dimension implies a more efficient means of quantification as well as a conceptual qualification of the model used (at least in terms of the Fractal Geometry of Nature [35] and methods of processing two- and three-dimensional data under a fractal based model [21]). For example, in [10] and [13], the following scaling law is proposed for the mean turbine power output $\langle \log P \rangle_\tau$ (over a period of time $\tau$):

$$\langle \log P \rangle_\tau \propto \frac{1}{\gamma}$$

where $\gamma$ is computed from the wind velocity over the same time period. Quantification of this scaling relationship is now required based on known turbine output power and wind velocity measurements. Finally with regard to urban environments in particular, it may be possible to find a correlation between the Fractal Dimension of the polar wind velocity and the roughness of the local area from a high resolution satellite image of the type given in Figure 8. By computing fractal parameters such as the Image Dimension (Fractal Dimension of a surface), the Information Dimension, Lacunarity and other Multi-Fractal parameters [21], for example, it may be possible to generate a single or combined image roughness measure. A correlation of this measure with the Fractal Turbulence Intensity reported could provide a way of estimating the wind turbulence and hence, subject to quantifying the inverse scaling relationship given above, predict the power performance of wind turbines in a rural environment from an satellite image alone! Such a solution would provide a simple and effective way of prospecting for wind resources in urban environments using on-line facilities such as Google Earth, for example.

ACKNOWLEDGMENTS

The research reported in this paper is supported by the Science Foundation Ireland Stokes Professorship Programme. The authors also acknowledge the support of Dublin Institute of Technology and the Dublin Energy Lab. The wind velocity data provided for the study given in Section V is provided by Wind Prospect Group. The urban data used in the wind turbulence analysis (Section VII) has been provided by Dr. Rowan Fealy, National University of Ireland, Maynooth and Dr. Gerald Mills, University College Dublin, Ireland.

REFERENCES

Jonathan Blackledge holds a PhD in Theoretical Physics from London University and a PhD in Mathematical Information Technology from the University of Jyvaskyla. He has published over 200 scientific and engineering research papers including 14 books, has filed 15 patents and has been supervisor to over 200 MSc/MPhil and PhD research graduates. He is the Science Foundation Ireland Stokes Professor at Dublin Institute of Technology where he is also an Honorary Professor and Distinguished Professor at Warsaw University of Technology. He holds Fellowships with leading Institutes and Societies in the UK and Ireland including the Institute of Physics, the Institute of Mathematics and its Applications, the Institution of Engineering and Technology, the British Computer Society, the Royal Statistical Society and Engineers Ireland.

Eugene Coyle is Head of Research Innovation and Partnerships at the Dublin Institute of Technology. His research spans the fields of control systems and electrical engineering, renewable energy, digital signal processing and ICT, and engineering education and has published in excess of 120 peer reviewed conference and journal papers in addition to a number of book chapters. He is a Fellow of the Institution of Engineering and Technology, Engineers Ireland, the Energy Institute and the Chartered Institute of Building Services Engineers, was nominated to chair the Institution of Engineering and Technology (IET) Irish branch committee for 2009/10 and a member, by invitation, of the Engineering Advisory Committee for the Frontiers Engineering and Science Directorate of Science Foundation Ireland Stokes Professor at Dublin Institute of Technology, and a member, by invitation, of the Engineering Advisory Committee for the Frontiers Engineering and Science Directorate of Science Foundation Ireland.

Niall McCoy graduated with a Diploma in Electrical Engineering in 2008 and Honours Degree in Electrical Services & Energy Management in 2010 from Dublin Institute of Technology. He is a Member of Engineers Ireland, a Chartered Engineer with the Energy Institute, an Accredited Professional Engineer with the South Africa Engineering Institute and joined the renewable energy company Wind Prospect Group in 2007 where he is currently responsible for wind farm developments world-wide such as in Canada, US, South Africa, UK & Ireland, having successfully project managed over 1.4GW of wind farm installations to date. Following a gap-analysis in the wind industry with regard to the development of more quantitative wind energy forecasting models in September 2011, Niall registered for a PhD degree at Dublin Institute of Technology with the provisional title of Wind Power Analysis using Non-standard Statistical Models.

Derek Kearney graduated in 2007 from Ulster University with a First Class Masters Degree in Renewable Energy, and is currently working on a PhD in the area of wind energy. He is a lecturer in the College of Engineering and Built Environment at Dublin Institute of Technology and is programme manager for the MSc courses in Energy Management, Sustainable Electrical Energy Systems and Sustainability, Technology and Innovation. He is also programme manager for the BSc in Electrical Services and Energy Management. His principal area of specialty for both teaching and research is in the area of renewable energy and energy conversion systems. Recent achievements include winner of the 2006 President’s Award for Teaching Excellence. He was chairman of the Honours Degree in Electrical Services and Energy Management for three years and previously worked on mainland Europe for a number of years.

Keith Sunderland is an electrical engineering graduate of Dublin Institute of Technology (DIT) with a first class honours degree in Electrical/Electronic Engineering. He is a member of the (DIT) Power Research Group as well as the Dublin Energy Lab and his research focuses on the applications of micro wind generation. More specifically, his interests are with respect to urban wind profiling and (distribution) network tolerance to increased technology proliferation and his PhD research is in collaboration with the School of Geography, Planning & Environmental Policy at University College Dublin (UCD). He is the DIT member of the Dublin Urban Boundary Layer Experiment (DUBLEx), which is a cross institution collaboration (DIT, National University of Ireland, Maynooth and UCD) investigating urban climatology with respect to energy budgets, CO₂ fluxes and energy applications. He is currently the Assistant Head of Department, Electrical Services Engineering within the School of Electrical Engineering Systems (at DIT).

Thomas Woolmington joined the lecturing staff of Dublin Institute of Technology of the Electrical Services Department in September 2006 and has qualifications in the electrical services and education areas including a national craft certificate (Electrician), a BSc in Electrical Services Engineering and Energy Management and a Postgraduate Certificate in Third Level Teaching and Learning. In April 2009 he co-authored a paper with other members of the Solar Energy Society of Ireland (SESIE) on The Uptake of Micro-gen Training on Third Level Courses Between Ireland and Wales presented at the PV-SAT5 Conference held in Glyndwr University, Wales. In October 2009 he secured funding from Enterprise Ireland to undertake a feasibility study into photometric testing of LED luminaries using proprietary instruments and in April 2011 he registered on a PhD research programme at Dublin Institute of Technology which is concerned with the statistically self-affine modelling of turbulence on the power performance of wind turbines.
A Chaotic Modulation Method based on Combination of CPWPM and BPSK for Digital Communication

Vu Van Yem¹, Nguyen Xuan Quyen¹ and Kyandoghere Kyamakya²

¹School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Vietnam
²Institute for Smart-system Technologies, Alpen-Adria Klagenfurt University, Klagenfurt, Austria

Abstract—In this paper, we present and investigate a chaotic modulation method based on combination of chaotic pulse width-position modulation (CPWPM) and binary phase-shift keying (BPSK) for digital communication. Firstly, binary information is conveyed by a sinusoidal carrier by means of BPSK modulation. Schemes for modulator and demodulator are proposed and their operation is described in detail. In addition, chaotic behavior and its effect on average parameters of the method are investigated. Theoretical calculation and numerical simulation of a CPWPM system with specific parameters are carried out. Simulation results on time and frequency domains as well as bit error rate (BER) performance in additive white Gaussian noise (AWGN) channel are shown in order to verify the operation and performance of the proposed method.

Index Terms—Nonlinear dynamics, chaos, chaotic modulation, CPPM, CPWPM, BPSK, chaos-based digital communications

I. INTRODUCTION

Over the past decade, increasing efforts have been devoted to study the possibility of using chaotic behavior to improve the features of communication systems [1], [2]. Many chaotic modulation methods have been proposed [3]–[5] and most of them exploited chaotic signals generated by dynamical systems to convey information. A robust modulation method named chaotic pulse-position modulation (CPPM) was introduced [6], [7] to reduce the impact of noise and distortion from communication channel on chaotic synchronization. CPPM signal is in the pulse train format, in which the binary information is modulated onto chaotically-varied intervals of inter-pulses. The principal advantage of CPPM is the automatic synchronization without the need of specific hand-shaking protocols [8]. In [9], [10], a chaotic modulation method which is considered as a development of CPPM, named chaotic pulse width-position modulation (CPWPM) was proposed, where the binary information is modulated onto both chaotically-varied intervals of position and width of output pulses. It means that two bits are encoded on a single pulse. There, the position and width of the present pulse are determined by time intervals from its rising edge to the rising edge of the previous pulse and to its falling edge, respectively. Since CPWPM signal is also in the pulse train format as the CPPM signal, the demodulation process can establish and maintain the synchronization state automatically. The CPWPM signal can be considered as a time-modulated baseband binary signal which is able to become the input signal of a binary sinusoidal carrier modulation method such as on-off keying (OOK), binary frequency-shift keying (BFSK) or binary phase-shift keying (BPSK) [11], etc. In practice, the BPSK technique provides a good bit error rate (BER) performance in noise-affected environments [11] with a simple structure and thus it is selected to investigate in the combination with CPWPM in this research. This combination method is called CPWPM+BPSK modulation.

The rest of this paper is organized as follows: In Section II, schemes for CPWPM+BPSK modulator and demodulator are presented and described in detail. Section III investigates the chaotic behavior with test map and from that average parameters of the system are determined. A CPWPM+BPSK system with specific parameters is calculated and simulated, and their results are shown in Section IV. Finally, our conclusion is given in Section V.

II. CPWPM+BPSK MODULATOR AND DEMODULATOR

Schemes for CPWPM+BPSK modulator and demodulator are presented as in Fig. 1b and Fig. 1c, respectively, where each scheme consists of CPWPM and BPSK parts.

A. Chaotic-Position Pulse Generator (CPPG)

Basically, the CPWPM parts in the modulator and demodulator are built around chaotic-position pulse generators (CPPG). Configuration of the CPPG is shown in Fig. 1a. A counter operates in free running mode to produce a linearly increasing signal, \( C(t) = K_1 t \), where \( t \) is the time duration from the reset instance and \( K_1 \) is count-step (the slope of the signal). This linearly increasing signal is reset to zero by each input pulse. Before the reset time \( t_n \), the output value of the counter, \( X_n = K_1 t_n \), is stored in the sample-and-hold circuit (S&H) whose output is fed to the nonlinear converter \( F(\cdot) \). An amplifier with a gain-factor, \( K > 1 \), is used to produce another linearly increasing signal, \( A(t) = K K_1 t = K_2 t \), which has a higher slope compared with that of the input signal. When the magnitudes of the output signals of the amplifier and that of counter reach the same value \( F(X_n^2) \) at the output of the \( F(\cdot) \), two narrow pulses at Outputs 2 and 1 are generated at the times, \( t_n^2 = F(X_n^{2})/K_2 \) and \( t_n^1 = F(X_n^{2})/K_1 \), respectively. It is easy to find that the time \( t_n^1 \) is earlier than \( t_n^2 \) and these times can be controlled by the values of the gain-factor \( K \) and count-step \( K_1 \). With a proper choice of parameters, when Output 1 is connected back to the input to form a closed loop, CPPG will
generate two chaotic pulse trains at its two outputs.

B. Modulator

At the CPWPM part, the binary information is modulated onto the inter-pulse intervals of two pulse trains at the outputs of the CPPG by means of delay modulators in the corresponding feedback loops. At the delay modulators, the input pulses trigger the data source to get the next binary \( S^1_n \) and \( S^2_{n+1} \). Depending on the values of these binary inputs, the output pulses, \( O_1(t) \) and \( O_2(t) \), are delayed by time durations, \( d_2 + m_2 S^2_n \) and \( d_1 + m_1 S^1_n \), respectively. Note that \( d_2 \) and \( d_1 \) are constant time delays that are inserted to guarantee the synchronization of the system; \( m_2 \) and \( m_1 \) are modulation depths which are delayed-time differences between “0” and “1” bits. Therefore, the delayed pulses, \( M_2(t) \) and \( M_1(t) \), at the outputs of Delay modulators 2 and 1 are generated at the times, \( t^*_2 = t^*_1 + d_2 + m_2 S^2_n \) and \( t^*_1 = t^*_n + d_1 + m_1 S^1_{n+1} \), respectively. After that, the modulated chaotic pulse trains are applied to a pulse-triggered edge generator (PTEG) whose output will switch to high (1V) and low (-1V) levels when its inputs 1 and 2 are triggered, respectively. The pulse train with the non-return-to-zero (NRZ) type at the output of the PTEG is also the CPWPM signal which can be expressed as follows:

\[
U_{\text{CPWPM}}(t) = \sum_{n=0}^{\infty} p_{\Delta t_n \Delta T_{n+1}}(t - t_n),
\]

where \( p_{\Delta t_n \Delta T_{n+1}}(t - t_n) \) is the rectangular pulse shaping function defined by

\[
p_{\Delta t_n \Delta T_{n+1}}(t - t_n) = \begin{cases} 
  1 & 0 \leq t \leq \Delta t_n, \\
  \Delta t_n \leq t \leq \Delta T_{n+1}, \\
  0 & \text{otherwise} 
\end{cases}
\]

and \( \Delta t_n, \Delta T_{n+1} \) are the width of the \( n \)th pulse and the position of the \((n+1)\)th pulse, respectively, which are determined by the following intervals:

\[
\{ \Delta t_n = F(K_1 \Delta T_n)/K_2 + d_2 + m_2 S^2_n, \\
\Delta T_{n+1} = F(K_1 \Delta T_n)/K_1 + d_1 + m_1 S^1_{n+1}. \}
\]

The parameters \( m_1, m_2, d_1, d_2, K_1, K_2 \) and the map \( F(\cdot) \) are chosen with proper values to guarantee the chaotic variation in the position and width of the CPWPM pulses. This choice will be investigated in the next section. It can be seen that the CPWPM signal is in the NRZ pulse train format and able to be considered as a time-modulated baseband binary signal which is suitable to become the input signal of BPSK modulation. In our modulator, a simple BPSK modulation part is employed, which consists of a mixer (M), a local oscillator (LO) and a band pass filter (BPF). The output signal of the mixer is the unfiltered CPWPM+BPSK signal, expressed by:

\[
U_{\text{Unfiltered CPWPM+BPSK}}(t) = A_c \cos(2\pi f_c t) U_{\text{CPWPM}}(t)
\]

\[
= A_c \cos(2\pi f_c t) \sum_{n=0}^{\infty} p_{\Delta t_n \Delta T_{n+1}}(t - t_n),
\]

where \( A_c, f_c \) are the amplitude and frequency of the carrier at the output of the local oscillator. The central frequency \( f_{0-BPF} \) and bandwidth \( BW_{\text{BPF}} \) of the BPF are equal to \( f_c \) and \( 2 BW_{\text{LPF}} \), respectively. Here, \( BW_{\text{LPF}} \) is the average bandwidth of the CPWPM signal, which will be determined in the next section. Finally, the output signal of the BPSK modulator is also the CPWPM+BPSK signal at the output of the transmitter.

C. Demodulator

The BPSK demodulation part in the demodulator consists of a carrier recovery block (CR), a mixer (M), a low pass filter (LPF) and a threshold detector (TD). The bandwidth \( BW_{\text{LPF}} \) of the LPF is equal to \( BW_{\text{LPF}} \). The NRZ pulse train at the output of the TD is the recovered CPWPM signal which is then applied to the CPWPM part to recover the binary information. At the CPWPM part, the input pulse train is fed to an edge-triggered pulse generator (ETPG) which is triggered by the rising and falling edges of the input pulses to produce narrow pulses at the Outputs 1 and 2, respectively. The Output 1 of the ETPG is connected to a CPPG. Note that the CPPG in the demodulator is identical to that in the modulator. When the synchronization state is established and maintained, the reproduced chaotic pulse trains at the outputs of the CPPG are also identical to those in the modulator. At Delay detectors 2 and 1, reproduced pulses from the CPPG are compared with pulses from ETPG to determine the delayed-time durations, \( \Delta t_n - F(X^1_n)/K_2 \) and \( \Delta T_{n+1} - F(X^2_{n+1})/K_1 \), respectively. Consequently, the binary information is recovered as follows:
Eq. (5) points out that, to re-establish the synchronization state and correctly recover the binary information, the demodulator only needs to correctly detect three consecutive intervals, \( \Delta T_n \), \( \Delta T_{n+1} \), and \( \Delta T_{n+2} \). Note that the set of values of \( m_1 \), \( m_2 \), \( a_1 \), \( a_2 \), \( K_1 \), \( K_2 \) and \( F(.) \) is considered as a secret key. The binary information is only correctly recovered when a demodulator has full information on these parameters.

D. Filters

Due to the impact of filtering, the input signal of the threshold detector in the BPSK demodulator is seriously affected by frequency and phase-delayed distortions in the BPF and LPF. Therefore, the inter-edge intervals of the recovered pulse train at the output of the TD are changed by these distortions. This can cause errors in recovering binary information in the CPWPM part of the demodulator. The frequency distortion is decreased by increasing the bandwidth of the filters but this leads to the increment of the occupied bandwidth of the CPWPM+BPSK signal in the channel. Since the Bessel filter is a type of linear filter with a maximally linear phase response (a maximally flat group delay) in the pass band [12], [13], the Bessel BPF and LPF are suitable to employed in our propose schemes in order to restrict the effect of the phase-delay distortion.

III. Chaotic Behavior and Average Parameters

Tent map is a discrete-time and one-dimension nonlinear function with the piecewise-linear I/O characteristic curve and it is used for generating chaotic values seen as pseudo-random numbers [14], [15]. In the communication, the tent map is proposed for application in chaotic modulation [6]–[10] with such advantages as the simplified calculation and the robust regime of chaos generation for rather broad range of modulation parameters. Here, the utilization of tent map for chaotic behavior of the CPWPM+BPSK method is investigated. This is very important for design process to guarantee the chaotic behavior of the method. It is noted that the chaotic behavior of CPWPM+BPSK totally depends on that of CPWPM. In other words, the chaos of CPWPM leads to the chaos of CPWPM+BPSK method.

A. CPWPM Tent map

The conventional tent map is iteratively generated through a transformation function \( F(.) \): \((0, 1) \rightarrow (0, 1)\) as given by

\[
\begin{align*}
    x_n &= F(x_{n-1}) = F^{(n)}(x_0) = \\
    &\begin{cases} 
    \alpha x_{n-1}, & 0 < x_{n-1} \leq 0.5, \\
    \alpha (1 - x_{n-1}), & 0.5 < x_{n-1} < 1,
    \end{cases}
\end{align*}
\]

(6)

where \( n \) represents the number of iteration, \( x_0 \) is the initial value, \( x_n \) is the output value at the \( n^{\text{th}} \) step, and the parameter \( \alpha \) controls the state of the map.

According to Eq. (3), the position and width of the \( n^{\text{th}} \) pulse are rewritten as follows:

\[
\begin{align*}
    \Delta T_n &= F(K_1 \Delta T_{n-1}) / K_1 + d_1 + m_1 S_n^1, \\
    \Delta T_{n+1} &= F(K_1 \Delta T_{n-1}) / K_1 + d_2 + m_2 S_n^2.
\end{align*}
\]

(7)

then these intervals can be converted to the following

\[
\begin{align*}
    X_n^1 &= F(X_{n-1}^1) + d_1^1 + m_1 S_n^1, \\
    X_n^2 &= F(X_{n-1}^2) + d_2^2 + m_2 S_n^2
\end{align*}
\]

(8)

From Eq. (6) and Eq. (9), the tent map for CPWPM, called the CPWPM ten map, is derived as

\[
\begin{align*}
    X_n^1 &= F(X_{n-1}^1) + d_1^1 + m_1 S_n^1, \\
    X_n^2 &= F(X_{n-1}^2) + d_2^2 + m_2 S_n^2
\end{align*}
\]

(9)

B. Chaotic Behavior

The equation of the CPWPM tent map above points out that its chaotic behavior depends on not only the control parameter \( \alpha \), but also on the parameters \( \delta_1 = d_1^1 + m_1 S_n^1 \) and \( \delta_2 = d_2^2 + m_2 S_n^2 \). The chaos of \( X_n^1 \) depends on \( \alpha \) and \( \delta_1 \); the chaos of \( X_n^2 \) depends on the chaos of \( X_n^1 \) with a difference, \( X_n^2 - X_n^1 = \delta_2 - \delta_1 \). In other words, the chaos of \( X_n^1 \) leads to the chaos of the system. The Lyapunov exponent of the map is determined by

\[
\lambda = \lim_{k \to \infty} \left( \frac{1}{k} \sum_{n=0}^{k-1} \ln \left| \frac{dF(X_n^1)}{dX_n^1} \right| \right) = \lim_{k \to \infty} \left( \frac{1}{k} \sum_{n=0}^{k-1} \ln a \right) = \ln a
\]

(11)

Based on Eq. (10) and Eq. (11), the behavior of the CPWPM tent map becomes chaotic in \([0, 1]\) with the following condition

\[
\begin{align*}
    \delta_1 &\geq 0, \\
    0.5a + \delta_1 &\leq 1, \\
    \lambda &= \ln a \geq 0,
\end{align*}
\]

(12)

which is equivalent to

\[
\begin{align*}
0 \leq &\delta_1 \leq 0.5, \\
1 < &a \leq 2(1 - \delta_1).
\end{align*}
\]

(13)

Fig. 2. Bifurcation diagram of the CPWPM tent map.
chaotic area is. And, in the $\delta_1 > 0.5$ the chaotic area disappears. It is easy to find that the bifurcation diagram of $X^1_n$ is also the bifurcation diagram of $X^1_n$ after being shifted vertically with a distance, $\delta_2 - \delta_1$.

In the modulation process, the binary bit $S^1_n$ varies between “0” or “1” and thus $\delta_1$ has two values, $d^1_1$ and $d^1_1 + m^1_1$. Based on Eq. (13), the condition in order to guarantee the chaotic behavior in the CPWPM+BPSK method is

$$
\begin{cases}
0 \leq d^1_1 + (M - 1)m^1_1 \leq 0.5, \\
1 \leq a \leq 2\left(1 - (d^1_1 + (M - 1)m^1_1)\right),
\end{cases}
$$

\begin{align}
\text{(14)}
\end{align}

or

$$
\begin{cases}
0 \leq K_1(d_1 + m_1) \leq 0.5, \\
1 \leq a \leq 2\left(1 - K_1(d_1 + m_1)\right).
\end{cases}
$$

\begin{align}
\text{(15)}
\end{align}

C. Average Parameters

In the iteration process, the CPWPM tent map varies chaotically around a fixed point [14] $(X^1_{fp}, X^2_{fp})$ determined by

$$
\begin{cases}
X^1_{fp} = F_1(X^1_{fp}) = (a + \delta_1)/(1 + a), \\
X^2_{fp} = F_2(X^2_{fp}) = a(1 - \delta_1)/(1 + a) + \delta_2.
\end{cases}
$$

\begin{align}
\text{(16)}
\end{align}

In the modulation process, due to the variation between “0” or “1” of input binary bits $S^1_n$ and $S^2_n$, this fixed point is shifted around an average fixed point $(X^1_{av}, X^2_{av})$ as follows:

$$
\begin{cases}
X^1_{av} = (a + (d_1 + m_1/2))/(1 + a), \\
X^2_{av} = a(1 - (d_1 + m_1/2))/(1 + a) + d_2 + m_2/2.
\end{cases}
$$

\begin{align}
\text{(17)}
\end{align}

Due to this feature, the intervals of position and width of the CPWPM signal vary chaotically around average intervals

$$
\begin{align}
\Delta T_{2av} &= \lim_{k \to \infty} \left(\frac{1}{k} \sum_{n=0}^{\infty} \Delta T_n\right) \approx \frac{\Delta T_{2av}}{R_1}, \\
\Delta \tau_{2av} &= \lim_{k \to \infty} \left(\frac{1}{k} \sum_{n=0}^{\infty} \Delta \tau_n\right) \approx \frac{\Delta \tau_{2av}}{R_1},
\end{align}
$$

\begin{align}
\text{(18)}
\end{align}

and its spectrum therefore has an average fundamental harmonic $f_{av-fund}$ and an average bandwidth $BW_{av}$ which are

$$
\begin{align}
f_{av-fund} &= 1/\Delta T_{av} \approx K_1/X^1_{av}, \\
BW_{av} &= 1/\Delta \tau_{av} \approx K_2/X^2_{av}.
\end{align}
$$

\begin{align}
\text{(19)}
\end{align}

The value of the average fundamental harmonic is equal to the average number of pulses transmitted in one second. Since each CPWPM pulse conveys two bits, the average bit-rate $BW_{av}$ of the system is evaluated as follows:

$$
BW_{av} = 2f_{av-fund} \approx 2K_1/X^1_{av}.
$$

\begin{align}
\text{(20)}
\end{align}

IV. Numerical Simulation And Results

In this section, a simplified CPWPM+BPSK system as in Fig. 3 with specific parameters is calculated theoretically and simulated numerically in order to verify the analysis results as well as performance of the presented method.

The specific parameters of the CPWPM system are chosen as follows: the fundamental sampling period $\tau = 1\mu s$, $K_1 = 0.002/\mu s$, $K = 2.5$, $d_1 = 20\mu s$, $d_2 = 10\mu s$, $m_1 = 50\mu s$, $m_2 = 30\mu s$; the nonlinear converter $F(.)$ uses the tent map with $a = 1.5$. The frequency and amplitude of the carrier in the BPSK parts are $f_c = 1 MHz$ and $A_c = 1 V$, respectively. In order to decrease the phase-delayed distortion, the 3rd-order Bessel BPF and LPF are used. From the calculation result in Subsection IV.A, the parameters of the filters are adopted as follows:

$$
\begin{align}
BW_{BPF} &= 2BW_{av} \approx 2 \times 7.451 = 14.902 KHz, \\
& \quad \quad \quad \quad \quad \quad \quad \quad \quad f_0-BPF = f_c = 1 MHz, \\
BW_{LPF} &= BW_{av} = 7.451 KHz.
\end{align}
$$

A. Theoretical Calculation

Based on Eq. (8), the CPWPM tent map is determined by the following parameters:

$$
\begin{align}
K_2 &= KK_1 = 2.5(0.002/\mu s) = 0.005/\mu s, \\
d_1' &= K_1d_1 = (0.002/\mu s)20\mu s = 0.4, \\
d_2' &= K_2d_2 = (0.005/\mu s)10\mu s = 0.05, \\
m_1' &= K_1m_1 = (0.002/\mu s)50\mu s = 0.1, \\
m_2' &= K_2m_2 = (0.005/\mu s)30\mu s = 0.15, \\
k(d_1 + m_1) &= (0.002/\mu s)(20\mu s + 50\mu s) = 0.14.
\end{align}
$$

With $K_1(d_1 + m_1) = 0.14$, the condition for the chaos of the method according to Eq. (15) becomes $1 \leq a \leq 2(1 - 0.14) = 1.72$. Therefore, we choose $a = 1.5$ to guarantee the chaotic behavior of the CPWPM+BPSK system. Based on the analysis in the Subsection III.C, the average parameters of the system are calculated as follows:

$$
\begin{align}
X^1_{av} &= (1.5 + (0.04 + 0.05))/2.5 = 0.636, \\
X^2_{av} &= 1.5(1 - (0.04 + 0.05))/2.5 + 0.05 + 0.075 = 0.761, \\
\Delta T_{av} &= 0.036/0.002 = 318\mu s, \Delta \tau_{av} \approx 0.671, \\
BW_{av} &= 1/\Delta T_{av} \approx 134.2 KHz, \\
f_{av-fund} &= 3.145 KHz, \\
BR_{av} &= 2 \times 3.145 = 6.29 Kbps.
\end{align}
$$

B. Numerical Simulation

Fig. 4 and Fig. 5 respectively present time domain signals in the modulator and demodulator obtained from simulation of...
the CPWPM+BPSK system in the AWGN channel within the duration from starting time 0 to 5000\(\mu\)s. It is observed that the intervals of the position and width vary chaotically in the ranges 200\(\mu\)s to 500\(\mu\)s and 75\(\mu\)s to 200\(\mu\)s, respectively. The system works in the synchronization state and the binary information is recovered correctly after about 500\(\mu\)s from the starting time.

Fig. 5. Time-domain signals of the demodulator.

Spectrum of the CPWPM signal at the output of the mixer is shown in Fig. 7a. The average fundamental harmonic is \(f_{av-fund} \approx 3.15\text{ KHz}\) and the average bandwidth is \(BW_{av} \approx 7.5\text{ KHz}\). The spectrum of the CPWPM+BPSK signals at the input and output of the 3rd-order Bessel BPF are presented in Fig. 7b and Fig. 7c, respectively. It points out that the occupied bandwidth of the transmitted signal in the transmission channel is equal to \(BW_{BPF} \approx 15\text{ KHz}\). Fig. 5c and Fig. 7d show the waveform and spectrum of the output signal of the 3rd-order Bessel LPF, respectively. We can observe that the distorted CPWPM signal is under the impact of filtering and AWGN.

Fig. 6. Attractor diagram with the average fixed point \((X^1_{av}, X^2_{av})\).

The chaotic behavior of the system is verified by attractor diagram in Fig. 6. In the modulation process, the fixed point is shifted on the bisector and around the average fixed point, \((X^1_{av}, X^2_{av})\) (red point). We can observe that values of the average parameters in the simulation results are completely reasonable to that of the theoretical calculation above. This proves the validation of the theoretical analysis.

Fig. 7. Frequency-domain signals.

The BER performance obtained from simulation of the CPWPM+BPSK system in AWGN channel in comparison to those of the conventional BPSK [11] and CPPM [6] systems is presented in Fig. 8. These BERs are calculated separately as the number of error bits divided by the total number of \(10^8\) bits transmitted. We can see that the BER performance of the CPWPM+BPSK is worse than that of BPSK, but it is much better than that of CPPM. For instance, at the same BER =
10^{-3}$, the $E_b/N_0$ rates for the BPSK, CPWPM+BPSK and CPPM are about 6.4dB, 10.2 dB and 15dB, respectively. This proves that the application of the CPWPM+BPSK to digital communication is feasible.

This paper has presented and investigated the combination modulation method of the CPWPM and BPSK for digital communication. The proposed schemes for the modulator and modulation method of the CPWPM and BPSK for digital telecommunications. The proposed schemes for the modulator and demodulator are described using both theoretical analysis and numerical simulation in terms of time- and frequency-domain signals and BER performance. It can be seen from simulation results that the combination method not only inherits valuable features of the CPPM-based methods such as the high privacy due to its chaotic behavior and the automatic synchronization based on the pulse train format; but also achieve a much better performance in noise- and distortion-affected environments as compared to the conventional CPPM. This makes the CPWPM+BPSK method become a good candidate for applications in security- and narrowband-required digital communication systems.

ACKNOWLEDGMENT

This work is carried out in the framework of the research project supported by the Vietnam’s National Foundation for Science and Technology Development (NAFOSTED) No. ID 102.02-2012.34.

Vu Van Yem received the Ph.D. degree in communications from the Department of Electronics and Communications, TELECOM ParisTech (formerly ENST Paris) France in 2005. From 2006 to 2007, he was a postdoctoral researcher at the Department of Hyper-frequencies and Semiconductor, Institute of Electronics, Microelectronics and Nanotechnology (IEMN), France. He has been qualified to be named as Associate Professor since November, 2009. Currently, he is the head of the Department of Telecommunication Systems and the vice dean of School of Electronics and Telecommunications, Hanoi University of Science and Technology, Vietnam.

Nguyen Xuan Quyen received the Diploma of Engineer and Master degree from the School of Electronics and Telecommunications (SET), Hanoi University of Science and Technology (HUST) in 2006 and 2008, respectively. From 2007 to 2010, he was a lecturer at the Department of Telecommunication Systems of SET-HUST. Currently, he is working toward PhD degree at Institute for Smart Systems Technologies, Alpen-Adria University Klagenfurt in Klagenfurt, Austria. His main research interests are microwave circuits, ultra wide band communications and chaos-based digital communications.

Kyandoghere Kyamakya received the PhD degree in Electrical Engineering at the University of Hagen, Germany in 1999. He then worked three years as a post-doctorate researcher at the Leibniz University of Hannover in the field of Mobility Management in Wireless Networks. From 2002 to 2005, he was Junior Professor for Positioning Location-based Services at Leibniz University of Hannover. Since 2005, he has been Full Professor for Transportation Informatics and Director of the Institute for Smart Systems Technologies at the University of Klagenfurt in Austria.

REFERENCES

Non-Gaussian Anisotropic Diffusion for Medical Image Processing using the OsiriX DICOM

J. M. Blackledge, M. D. Blackledge and J. N. Courtney

Abstract—We present a method for reducing noise in CT (Computed Tomography) and MR (Magnetic Resonance) images that, in addition to other noise sources, is characteristic of the numerical procedures required to construct the images, namely, the (inverse) Radon Transform. In both cases, MR imaging in particular, an additional noise source is due to non-stationary diffusion thereby predating use of the Anisotropic Diffusion Method for noise suppression. This method is based on a diffusion model for noise generation where the Diffusivity is taken to be non-isotropic (inhomogeneous) or anisotropic and is, in the absence of a priori information, computed through application of an edge detection algorithm. We extend this approach to include the effect of fractional diffusion (when the underlying statistical model associated with the diffusion process is non-Gaussian) and derive a corresponding Finite Impulse Response filter. The algorithms developed are implemented using the OsiriX DICOM (Digital Imaging and Communications in Medicine) viewer, a high performance open source image data visualization system for the development of processing and visualization tools. An OsiriX Plugin Image Filter is provided for interested readers and practitioners to apply.

Index Terms—Medical imaging, Digital Imaging and Communications in Medicine, Computed Tomography, Magnetic Resonance Imaging, Non-Gaussian Anisotropic Diffusion, Noise suppression, OsiriX DICOM viewer.

I. INTRODUCTION

The aim of a noise reduction algorithm is to enhance the fidelity of an image by eliminating features that are random and uncorrelated while retaining resolution on those features that are noise free. The reason for this may be to enhance the visual clarity of an image for visual inspection and/or digital image analysis for object segmentation and feature selection. As a general rule of thumb, noise tends to corrupt the higher frequency content of an image where the energy of the data spectrum is usually lower. Thus, a way of reducing noise is by attenuating the high frequency components of the data over a range of frequencies which can be selected and adjusted to provide an optimum result, subject to some predefined image quality criterion. This can be achieved by applying a lowpass filter. A well known and commonly used lowpass filter is the Gaussian function which yields a ‘Gaussian Blur’ and, on a physical basis, is related to classical diffusion processes. This is because the Point Spread Function of an image generated by the diffusion of light is a Gaussian function, a result that can be derived by considering the propagation and interaction of light to be based on a random walk process. The underlying equation for the intensity of light is then given by the diffusion equation.

An important assumption associated with the application of a Gaussian blur filter is that the noise content of the image is additive and homogenous. This is to say that the statistical distribution of the noise is homogeneous throughout the image plane and does not change over the spatial extent of the image in terms of scale, width and/or distribution type. Even if this is the case, and, the noise associated with the generation of an image can be quantified (in a statistical sense), application of a Gaussian lowpass filter reduces at the expense of resolution on correlated features, patterns and textures.

In this paper, we review the Anisotropic Diffusion Method for noise reduction illustrating the relationship between this method and the Characteristic Function that is used to define an anisotropic Gaussian random walk process. This process is then extended to include the ‘non-Gaussian’ case based on Lévy’s (symmetric) Characteristic Function which introduces the ‘Lévy index’. By re-working the Anisotropic Diffusion Method for this case, we derive a new (Lévy index dependent) Finite Impulse Response (FIR) filter which is our original contribution to the field. The algorithms developed are applied to CT (Computed Tomography) and MR (Magnetic Resonance) and implemented in the OsiriX DICOM (Digital Imaging and Communications in Medicine) Viewer to produce a new image filter (an OsiriX Plugin) and an original application in the area of medical imaging.

Since the focus of this work is on noise reduction, a brief discussion of the sources of noise in CT and MR images is given as detailed in [1], for example. We start by providing a brief overview of DICOM which is given in the following section. However, it should be noted that diffusion based models have a range of applications that transcend image noise reduction especially in MR imaging where the process of diffusion is used to generate a form of image analysis known as Diffusion Weighted and Tensor Diffusion (MR) imaging, e.g. [2], [3], [4] and [5]. The approach considered in this paper may have additional applications in this area of imaging, details of which will be published elsewhere.

II. DIGITAL IMAGING AND COMMUNICATIONS IN MEDICINE

The Digital Imaging and Communications in Medicine (DICOM) standard is used for the exchange of medical images and related information. DICOM is a standard with several
levels of support including an image exchange underlying information model and information management services\(^1\) [6].

DICOM was introduced in 1993 after some ten years standards development from the early 1980s when only manufacturers of CT or MR imaging devices could decode the images that the early machines generated. DICOM differs from some, but not all, data formats in that it groups information into data sets. Thus, a file of a chest X-Ray image, for example, actually contains the patient ID within the file, so that the image can never be separated from this information by mistake. This is similar to the way that image formats such as JPEG can also have embedded tags to identify and describe the image.

DICOM has an information model which differentiates it from other standards used in the medical industries sector. The model is based on ‘information objects’ which include definitions (an information template) on the information to be exchanged. Each image type, and therefore information object, has specific characteristics. A CT image, for example, requires different descriptors in the image header compared to an ultrasound image or an ophthalmology image. These templates are identified by unique identifiers, which are registered by the National Electrical Manufacturers Association, the DICOM standard facilitator. Information objects are also known as part of the Service Object Pairs (SOP) Classes. An example of a SOP Class is the CT Storage SOP Class, which allows CT images to be exchanged.

DICOM includes a robust protocol where each DICOM command is acknowledged thereby providing an implicit version control. Strict guidelines are employed requiring each DICOM compatible device to describe its functionality, including supported SOP Classes and transfer syntaxes in a document called the DICOM Conformance Statement. This document allows a user to determine in advance whether or not a specific device is compatible with other devices. Conformance statements contain details about exception and error handling and often contain complete specifications for the information objects (e.g. images) that are exchanged. DICOM uses information management services which include a Modality Work List allowing scheduling information to be retrieved by a modality. The other service in this category is the DICOM Storage Commitment, which transfers the responsibility for images to the receiver, so they can be safely removed from the local disk.

Image quality is the principal issue with regard to the DICOM standardization. DICOM achieves consistency in the image presentation on different monitors, as well as on film, independent of the make or type of characteristics of the medium using the DICOM Greyscale Standard Display Function. This function specifies exactly what luminance or density level should be produced for a certain input value, based on the Barten curve, which maps the values into a range that is perceptually linear. This means that input values are mapped into a space that is perceived as linear by a human observer.

In addition to the principal objectives of the DICOM standard for image visualization, it includes structured reporting which is an object that can be exchanged, very similar to an image, except that instead of pixel data, the message body is a Structured Report. Finally DICOM facilitates security mechanism including access control and authorization rules implemented by the application level software using passwords or biometric access controls coupled with an audit and logging mechanism.

There are a number of DICOM viewers available which allow medical images to be displayed, processed and analyzed. Among the most advanced and versatile of these ‘viewers’ is the OsiriX imaging software which is limited only in its requirement for implementation in a Unix/Linux environment, thereby making use of the advanced graphics facilities that accompany an Apple Mac, for example.

### III. THE OSIRIX DICOM VIEWER

OsiriX is an advanced open-source DICOM viewer for visualizing and processing images produced by a range of medical imaging systems [7] including MR and CT, Position Emission Tomography (PET), PET-CT, ultrasonic B-Scan imaging and so on [8]. Designed for uses under a Unix/Linux operating system, it is arguably the world’s most widely used and versatile medical imaging processing and visualization system and is fully compliant with the DICOM standard for image communication and image file formats.

OsiriX has been specifically designed for navigation and visualization of multimodality and multidimensional images. It incorporates basic 2D viewing, a 3D viewer and 4D viewer (a 3D series with temporal dimension such as Cardiac-CT, for example) and a 5D viewer (3D series with temporal and functional dimensions; for example, Cardiac-PET-CT). The 3D viewer offers all modern rendering modes including Multi-Planar Reconstruction (MPR), Surface Rendering, Volume Rendering and Maximum Intensity Projection (MIP). All these modes support 4D data and are able to produce Image Fusions between different series of images.

The current version of OsiriX provides all aspects of DICOM file support and has a built-in SQL (Structured Query Language) compatible database with an unlimited number of files. Fully compatible network support is provided and the system is available as a 32-bit or 64-bit version. It has an intuitive GUI (Graphical User Interface) with customizable Toolbars and includes multi-slice CT and MR imaging with Regions of Interest (ROI) including Polygons, Circles, Pencil, Rectangles and Points with undo/redo. The 3D Post-Processing facilities of OsiriX include MIP (Maximum Intensity Projection), Volume Rendering, Surface Rendering, ROI, Texture Mapping, Image Registration and Stereo Vision (using glasses).

OsiriX provides custom CLUT (Color Look-Up Tables), 3×3 and 5×5 FIR (convolution) filters (e.g. a ‘Bone filter’), a 4D viewer for Cardiac-CT and other temporal series and Image Fusion for PET-CT examinations, for example, with adjustable blending percentage. With regard to the work reported in this paper, OsiriX provides a complete dynamic Plugin architecture for external functions used for expansion and scientific research. This includes the creation and management of OpenGL views.

\(^1\)An edited version of the DICOM standard, overview and characteristics: a whitepaper.
Figure 1 shows an example of the OsiriX DICOM Viewer illustrating the visualization of a PET image for the whole body and a CT scan of the head and upper torso. The use of combining data from CT scans (which provide high resolution images of both bone and soft tissues) and PET scans (which allows images to be derived on areas of the body that have been doped with radioactive isotopes of Oxygen, Carbon and Nitrogen, for example, with short half-lives) requires pixel and voxel registration using Image Fusion methods for which OsiriX provides an ideal programming environment [9]. OsiriX also has powerful volume rendering, surface and ROI visualization facilities. For example, Figure 2 shows a 3D rendered image from a CT scan of the middle torso.

Fig. 1. Screen shot of PET-CT images using the OsiriX DICOM Viewer.

Fig. 2. Screen shot of a 4D volume rendered CT image using the OsiriX DICOM Viewer.

A version of OsiriX is also available for the iPhone although is functionality is currently limited to low data volume visualization [10].

IV. NOISE SOURCES IN CT, PET AND MR IMAGING

CT, PET and MR images contain characteristic noise fields that are derived from a combination of effects due to a range of disturbances and interference that effect the signal detection and/or are an inherent component of the signals recorded. Because noise is multifaceted, it is not possible to define it uniquely. For this reason, statistical models are required to construct a suitable Probability Density Function for the noise which is statistically compatible with the data. However, as a consequence of the Central Limit Theorem, where the addition of many independent noise fields yields a resultant field that is Gaussian distributed, the noise fields in CT, MR and PET images is often taken to be Gaussian. However, there is a particular component of the noise in these imaging modalities that is unique in that it is a consequence of the image reconstruction method that is required. This is concerned with the underlying model used to describe a set of projections \( P(\hat{n}, z) \) in the plane (where \( \hat{n} \) is the unit vector pointing along the path of a projection) which is given by the Radon transform [11]

\[
P(\hat{n}, z) = \int_{-\infty}^{\infty} f(r) \delta(z - \hat{n} \cdot r) d^2r
\]

where \( f(r) \) is some 2D object function which is related to the physical basis of the imaging modality (e.g. X-ray absorption in the case of X-ray CT). This transform describes a set of parallel projections (line integrals) taken over all angles between 0 and 180°. The inverse Radon transform is given by [11], [12]

\[
f(r) = \frac{1}{2\pi^2} \int_{0}^{\frac{\pi}{2}} d\theta \int_{-\infty}^{\infty} dz \left( \frac{1}{\hat{n} \cdot r - z} \frac{\partial}{\partial z} P(\hat{n}, z) \right)
\]

This transform is the theoretical basis for parallel beam CT and PET although it should be noted that latter ‘fourth generation’ CT and most current systems are based on fan beam project tomography [13]. With regard to MR imaging, image reconstruction is typically undertaken using so called \( k \)-space (spatial frequency space) methods which are related to the Central Slice Theorem for the Radon transform [11].

Irrespective of the method used to reconstruct a CT, PET or MR image, numerical errors occur that are essentially a consequence of the need to map data in a polar or partially polar coordinate system into a Cartesian system. This is because of the intrinsic geometry of the imaging systems that are used where the recorded data has some angular dependence relative to the object function - the body. Although various filters are available to reduce the numerical errors that occur subject to a given resolution (which depends on the angular step change), for example, the fact that some form of numerical error is inevitable yields a characteristic noise field, particularly in the case of CT and PET. In the case of MR imaging, the noise fields are the result of more complex processes which include the effect of diffusion associated with the time taken to record the data and the intensity (amplitude) of the recorded signals which depend on the strength of the magnetic field and the RF fields used to induce proton gyration [14]. Figure 3 provides examples of noise fields that are characteristic of parallel beam and fan-beam (fourth generation) CT. In these examples, the projections are generated from a uniform image (with all pixels...
values set to 1) and the reconstructions (computed using a complete set of projections from 0-179 degrees inclusively in steps of 1 degree) have been histogram equalized for the purpose of visualizing the noise patterns produced.

![Image of noise fields](image)

**Fig. 3.** Example of the characteristic noise fields produced by a parallel beam reconstruction (top-left) using nearest neighbour Polar-to-Cartesian coordinate interpolation and the corresponding 256-bin histogram (top-right) after application of histogram equalization. The equivalent fan beam reconstruction and the corresponding histogram are given in the bottom-left and bottom-right quadrants of the figure, respectively.

V. EVOLUTION OF RANDOM PROCESSES, DIFFUSION AND NOISE

Anisotropic Diffusion describes diffusion processes where the Diffusivity is not a constant but may vary in space (Isotropic Inhomogeneous Diffusion when the Diffusivity is a scalar function of space) and whose spatial variations may be directional (Non-isotropic Inhomogeneous Diffusion when the Diffusivity is a Vector or Tensor function of space). In both cases the governing equation is the Diffusion equation and in this section we briefly look at the origins of this equation. In particular, we investigate the origins of the classical diffusion equation and the fractional diffusion equation. The purpose of this is to inform the reader of the different physical processes that are assumed in the derivation of the classical and fractional diffusion equations in which the fractional diffusion equation can be taken to be a generalization.

A. Einstein’s Evolution Equation

In [15], the following evolution equation is considered

$$u(r, t + \tau) = u(r, t) \otimes p(r), \quad r \equiv |r| = \sqrt{x^2 + y^2 + z^2} \quad (1)$$

where $\otimes$ denotes the convolution integral over all space (denoted by the vector $r$), $u(r, t)$ the ‘Density Field’ which describes the distribution of a canonical ensemble of particles undergoing elastic scattering and $p$ denotes the Probability Density Function (PDF). The space-time evolution of the Density Field $u(r, t)$ depends upon the PDF that characterizes the statistical behaviour associated with the elastic scattering processes in terms of each particle in the canonical ensemble of particles undergoing a three-dimensional random walk.

Einstein’s evolution equation is, in a conventional sense, related to the concept of particle interactions responsible for generating Brownian motion, for example. However, the same evolution equation can be considered in the context of light rays, for example, interacting with a population of random point scatterers where the density field is taken to be a measure of the intensity of the scattered light. In imaging theory, multiple scattering is taken to contribute to the noise term of the fundamental imaging equation where the recorded image is given by the convolution of the Object Function with the Point Spread Function of the imaging system plus noise. The convolution term models single scattering processes according to the Born or Kirchhoff approximations (for volume and surface scattering problems, respectively). However, in general, any imaging system is usually taken to conform to this model and the noise term may be generalized in terms of some product of an effect that is interpreted on a statistical basis. This includes the statistical model compounded in Einstein’s evolution equation, which, in this paper, forms the basis associated with the algorithms developed for reducing the effects of diffusion in CT and MR images.

B. Derivation of the Classical and Fractional Diffusion Equations

The classical diffusion equation can be derived from equation (1) by using a Taylor expansion of the density field $u(r, t + \tau)$ and the convolution integral. This is the basis for Einstein’s derivation of the diffusion equation from which the Diffusivity and Diffusion Tensors are constructed. Here, we consider a different approach which serves to provide a unifying theme for understanding the concept of Gaussian and non-Gaussian diffusion.

Using the convolution theorem, equation (1) can be written in Fourier space as

$$\tilde{u}(k, t + \tau) = \tilde{u}(k, t) \tilde{p}(k), \quad k \equiv |k| = \sqrt{k_x^2 + k_y^2 + k_z^2} \quad (2)$$

where $\tilde{u}$ and $\tilde{p}$ denote the Fourier transforms of $u$ and $p$, respectively ($\tilde{p}$ being referred to as the Characteristic Function, a conventional term in statistics and statistical mechanics), and $k_x, k_y$ and $k_z$ are the spatial frequencies in a Cartesian coordinate system.

Clearly, the properties of the Characteristic Function $\tilde{p}$ determine the statistical characteristics of the ‘system’, which, in turn, defines the evolution of the density field $u(r, t)$ via equation (1). For a homogeneous Gaussian system and for some real constant $a > 0$, $\tilde{p}(k) = \exp(-ak^2)$. We can generalize this result and consider the Lévy Characteristic Function $\tilde{p}(k) = \exp(-ak^\gamma)$ where $\gamma \in (0, 2]$ is the Lévy index. However, this generalization does not take into account the fact that the diffusing properties of the ‘system’ can be inhomogeneous. We therefore generalize further and consider the Characteristics Function

$$\tilde{p}(k) = \exp[-\tilde{u}(k) \otimes_k k^\gamma] \quad (3)$$
where \( \otimes_k \) denotes the convolution integral over all k-space which reduces to the homogeneous case when \( ˜a(k) = a\delta^\gamma(k) \) where \( \delta \) denotes the 3D Dirac delta function.

Given equations (2) and (3), we can now derive the inhomogeneous fractional diffusion equation which is based on:

(i) Taylor expanding \( ˜u(k, t + \tau) \) to first order so that

\[
  ˜u(k, t + \tau) \approx ˜u(k, t) + \tau \frac{\partial}{\partial t} ˜u(k, t)
\]

(ii) Considering a first order approximation to the exponential function of the form

\[
  \tilde{p}(k) \approx 1 - ˜a(k) \otimes_k k^\gamma
\]

(iii) Using the Reisz definition of a fractional Laplacian, i.e.

\[
  \nabla^\gamma \leftrightarrow -k^\gamma
\]

where \( \leftrightarrow \) denotes transformation from real-space to k-space.

Using these results we derive the equation

\[
  \frac{\partial}{\partial t} u(r, t) = D(r) \nabla^\gamma u(r, t) \quad (4)
\]

where \( D(r) = a(r)/\tau \) is the (fractional) ‘Diffusivity’ and we have used the result

\[
  [\tilde{a}(k) \otimes_k k^\gamma] ˜u(k, t) \leftrightarrow [a(r)\delta^\gamma(r)] \otimes_r u(r, t)
\]

and noted that, by Taylor expanding the function \( a \) (within the convolution integral), then, for a unit vector \( \hat{n} \),

\[
  [a(r)\delta^\gamma(r)] \otimes_r u(r, t) = a(r)[\delta^\gamma(r) \otimes_r u(r, t)]
\]

\[
  + \nabla a(r) \cdot \hat{n} \nabla^\gamma [ru(r, t)] + ... \sim -a(r)\nabla^\gamma u(r, t), \; r \to 0
\]

For a Gaussian distributed ‘system’ equation (4) reduces to the classical inhomogeneous diffusion equation when \( \gamma = 2 \). For \( \gamma < 2 \), equation (4) models a ‘system’ where the density function is the ‘product’ of a canonical ensemble of trajectories (particles of light-rays) that have a propensity for propagating over longer distances in an interval of time \( \tau \). This effect is compounded in the long tail distributions associated with the case when \( \gamma < 2 \) in equation (3) subject to the inhomogeneous nature of the system compounded in the function \( \tilde{a}(k) \) which determines the Diffusivity \( D \) (apart from scaling by \( 1/\gamma \)).

VI. NOISE REDUCTION USING ANISOTROPIC DIFFUSION

We consider the application of equation (4) in two-dimensions for the suppression of noise in an image \( u(x, y) \) which has been sampled to form a digital image consisting of a uniformly sampled array of pixels \( u_{ij} \).

A. Solution for \( \gamma = 2 \)

For \( \gamma = 2 \), equation (4) is

\[
  \frac{\partial}{\partial t} u(x, y, t) = D(x, y) \nabla^2 u(x, y, t)
\]

A numerical solution is considered that is based on solving the above equation using a time step \( \delta = \delta t \). Forward differencing in time and centre differencing in space (for a uniformly sampled grid with sampling intervals of \( \Delta x \) and \( \Delta y \)) we obtain

\[
  \frac{u_{ij}^{k+1} - u_{ij}^k}{\delta} = \frac{D_{ij}}{\Delta^2} (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij})
\]

where \( \Delta = \Delta x = \Delta y \). Rearranging this result, we obtain the iterative solution

\[
  u_{ij}^{k+1} = u_{ij}^k + \lambda D_{ij} \times L_{ij} \otimes_{i,j} u_{ij}^k, \; k = 0, 1, 2, ..., N \quad (5)
\]

where \( \lambda = \delta/\Delta^2 \) is the ‘time space-squared step ratio’, \( N \) is the total number of iterations and

\[
  L_{ij} = \begin{pmatrix}
  0 & 1 & 0 \\
  1 & -4 & 1 \\
  0 & 1 & 0 
\end{pmatrix}
\]

The matrix \( L_{ij} \) defines the Laplacian Finite Impulse Response (FIR) filter of this iterative system where \( \otimes_{i,j} \) denotes the discrete convolution sum over all \( i \) and \( j \). The initial value \( u_{0ij} \) is taken to be the original (noisy) image.

B. Solution for \( \gamma \in (0, 2) \)

A solution to this problem has previously been considered by [16] using a Fourier transform based approach and implemented using a FFT (Fast Fourier Transform) based algorithm. In this paper, we consider an approximation required to implement the method of solution using a FIR filter. The ‘key’ to this approach is to note that \( \nabla^\gamma \equiv \nabla^2 \nabla^{\gamma-2} \) so that the FIR filter given by equation (6) can be used and applied to the computation of \( \nabla^{\gamma-2} u(x, y, t) \). The computation of this term is based on the following two-dimensional Fourier transform relationship [17]

\[
  \frac{1}{r^\gamma} \leftrightarrow \frac{\Gamma_{\gamma}}{k^{2-\gamma}}
\]

where

\[
  \Gamma_{\gamma} = \frac{2^{2-\gamma}\pi\Gamma(1-\frac{2}{\gamma})}{\Gamma(\frac{2}{\gamma})}, \quad r = \sqrt{x^2 + y^2}, \quad k = \sqrt{k_x^2 + k_y^2}
\]

Using the Reisz definition for a fractional Laplacian,

\[
  \nabla^{\gamma-2} u(r, t) \leftrightarrow k^{\gamma-2} ˜a(k, t) = \frac{\tilde{a}(k, t)}{k^{2-\gamma}}, \quad 0 < \gamma < 2
\]

from which, given relationship (7) and using the convolution theorem, it follows that

\[
  \nabla^{\gamma-2} u(x, y, t) = \frac{1}{\Gamma_{\gamma}(x^2 + y^2)^{\gamma/2}} \otimes_{x,y} u(x, y, t)
\]

We can now approximate the convolution kernel to form a FIR filter (of an odd order) where, for a rectangular grid, the singularity is taken to occur at \((1, 1)\). Under these conditions, and, with \( \alpha = \gamma/2 \), we can consider an order 3 FIR filter of the form

\[
  M_{ij}^3(\alpha) = \frac{1}{\Gamma_{2\alpha}} \begin{pmatrix}
  8-\alpha & 5-\alpha & 8-\alpha \\
  5-\alpha & 2-\alpha & 5-\alpha \\
  8-\alpha & 5-\alpha & 8-\alpha
\end{pmatrix}
\]
and an order 5 FIR filter of the form

\[ M_{ij}^5(\alpha) = \frac{1}{2^{12}} \begin{pmatrix} 18^{-\alpha} & 13^{-\alpha} & 10^{-\alpha} & 13^{-\alpha} & 18^{-\alpha} \\ 13^{-\alpha} & 8^{-\alpha} & 5^{-\alpha} & 8^{-\alpha} & 13^{-\alpha} \\ 10^{-\alpha} & 5^{-\alpha} & 2^{-\alpha} & 5^{-\alpha} & 10^{-\alpha} \\ 13^{-\alpha} & 8^{-\alpha} & 5^{-\alpha} & 8^{-\alpha} & 13^{-\alpha} \\ 18^{-\alpha} & 13^{-\alpha} & 10^{-\alpha} & 13^{-\alpha} & 18^{-\alpha} \end{pmatrix} \]

where higher order filters are constructed similarly. In each case, the forward difference in time solution to equation (4) is given by (where \( n \) is the order of the filter)

\[ u_{ij}^{k+1} = u_{ij}^k + \lambda D_{ij} \times L_{ij} \otimes_{i,j} M_{ij}^n(\alpha) \otimes_{i,j} u_{ij}^k \quad (8) \]

where \( u_{ij}^0 \) is taken to be the input image.

C. Computation of the Diffusivity

In the context of processing a digital image \( u_{ij}^0 \), application of equations (5) and (8) depend critically upon the generation of the array \( D_{ij} \). In the absence of any \textit{a priori} information (which is the usual case) this array must be obtained from the image \( u_{ij}^0 \). With regard to using equations (5) and (8) for noise reduction, the value of \( D_{ij} \) at any pixel location \((i,j)\) can be used to control the degree of diffusion, which, in terms of a diffusion process, represents the degree of blurring in the locality of \((i,j)\). For correlated features in an image that are edge dominant and are taken to be relatively noise free (at least, in terms of their high pixel value contrast with regard to low pixel value background noise), we require the degree of diffusion to decrease. This can be accomplished by reducing the value of \( D_{ij} \) in regions of an image that are edge dominant and increasing the value of \( D_{ij} \) in those regions that are background noise dominant. A method of achieving this is to apply an edge detector to the image \( u_{ij}^0 \) to obtain an output \( E_{ij} \), say, and compute

\[ D_{ij} = 1 - E_{ij} \quad (9) \]

where \( E_{ij} \) is (intensity) normalized, i.e. \( 0 \leq E_{ij} \leq 1 \) \( \forall (i,j) \) so that \( D_{ij} \in [0,1] \) \( \forall (i,j) \).

There are a range of edge detection filters that can be applied in this respect to compute \( D_{ij} \) [17]. Further, binarization of equation (9) can be applied via application of a standard thresholding method, i.e. for a \textit{Threshold} \( 0 < T < 1 \),

\[ D_{ij} = \begin{cases} 1, & 1 - E_{ij} > T \\ 0, & 1 - E_{ij} \leq T \end{cases} \quad (10) \]

The issue of which edge detection filter to use and whether or not to binarize the output based on the application of equation (10) or to apply a variable threshold approach lies beyond the scope of this paper. As with other image processing applications, the filter(s) depends on a detailed appraisal of the data. This includes image acquisition, the application focus and user preferences. Edge detection methods are designed to provide a balance between the fidelity of the output in terms of edge accuracy, localization and continuity and the effects of noise leading to spurious edges. In this application, edge detection is being applied for the purpose of non-stationary noise reduction rather than pattern recognition. Further, because the current application is focused on clinical image analysis in real time, it is not appropriate to provide the user with a range of filters for edge detection including parameter settings etc. Rather, we require a hardwired solution that is efficient with minimal computational cost. In this context, second order edge detection via the Marr-Hildreth algorithm [18], for example, or first order edge detection requiring pre-filtering, such as the Canny edge detector [17] is not suitable (at least in a first study as provided in this paper). We therefore consider a well known and simple but effective edge detector, namely, on the Sobel filter which is based on the application of the FIR gradient filters

\[ G_x = \frac{1}{8} \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \quad G_y = \frac{1}{8} \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} \]

For an input image \( u_{ij}^0 \), the output is given by

\[ G_{ij} = | G_x \otimes_{i,j} u_{ij}^0 | + | G_y \otimes_{i,j} u_{ij}^0 | \]

which is then normalized to yield \( E_{ij} \) as follows:

\[ E_{ij} = \frac{G_{ij}}{\max(G_{ij})} \]

from which the Diffusivity can be obtained via equations (9) and (10) using a 50% threshold, i.e. \( T = 0.5 \).

VII. IMPLEMENTATION IN OSIRIX IMAGE PROCESSING SOFTWARE

OsiriX provides the facility for programming filters using the XCode Apple Developer C/C++ compiler [19]. Implementation of the algorithms discussed in the previous sections, in particular equation (8), is easily converted to C-code for implementation in an OsiriX programming environment. The current processor, the OsiriX Anisotropic Diffusion Processor (OADP), is available on-line at [20].

The OADP filter has been designed to be easy to implement and integrate in OsiriX as a Plugin and to be use by a radiologist. No parameter settings are required which comes at the expense of the user being unable to optimize the filter and investigate its properties in a dynamic and image dependent sense. However, these options are not suitable for the medical imaging practitioner who typically expects the operational criteria of a filter to be hard-wired. The current parameters are hard-wired according to Table I. Figure 4 shows an example screen shot after application of the OsiriX OADP filter for a typical CT image using the Plugin application file available at [20]. Figure 5 shows an example of using the OsiriX 3D volume rendering facility after applying a crop to locate a Brain tumour using MR image data to which the OADP filter has been applied. As a final example, Figure 6 shows the result of applying the OsiriX surface rendering option. The

<table>
<thead>
<tr>
<th>Number of Iterations</th>
<th>Time-Space</th>
<th>Lévy Index</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.01</td>
<td>1.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
image consists of two Iso-surfaces with equal transparency weighting, generated using CT data processed with the OADP filter.

VIII. CONCLUSIONS

We have investigated the principles associated with the Anisotropic Diffusion Method for noise reduction and extended the approach to include the case when the underlying statistical models for random walk processes are non-Gaussian. For the case when these processes are taken to be described by a Lévy distribution, a new $n^{th}$-order FIR filter $M_{ij}^n$ (where $n = 3, 5, ...$) has been derived in Section 6.2. The approach to deriving these filters (as presented in Section 5.2) is general and can therefore, in principle, be used to design filters for different Characteristic Functions. However, it should be noted that the Lévy function considered in this paper is approximated in order to undertake the analysis that is applied and that higher order results should be further investigated with regard to the design and implementation of different filters.

Irrespective of whether the Anisotropic Diffusion Method is based on a Gaussian or non-Gaussian model, the algorithms compounded in equations (5) and (8), respectively, are predicated on being able to obtain an estimate of the Diffusivity. This is typically undertaken through application of an edge detection method and, in this paper, we have consider the Sobel edge detection algorithm. However, there is considerable scope for investigating further edge detection methods for this purpose (which lie beyond the scope of this paper). In particular, given the binarization condition that has been applied via equation (10), application of the Marr-Hildreth second order edge detection algorithm may be of value in this context given that edge continuity tends to be preserved in this case. However, it should be noted that the application of second order edge detection for computing the Diffusivity comes at the expense of greater computational overheads and an increase in parameter optimization, in particular, the bandwidth that is applied.

The principal focus of the algorithm developed in this paper has been its implementation in a real time medical image processing environment. For this purpose, the OsiriX DICOM Viewer has been chosen because of the existing and extensive facilities that this system provides, its universal appeal and the programming environment that is available. In this context, the paper has reported on the development of a new OsiriX DICOM filter for noise reduction using a non-Gaussian anisotropic diffusion model. Readers can investigate this approach further by implementing the application available via [20] within the Osirix operational environment [8]. This application includes operational conditions that transcend the results presented in this paper and provides the user with additional options for investigating the performance of the filters considered in this paper.

ACKNOWLEDGMENTS

The research reported in this paper is supported by the Science Foundation Ireland Stokes Professorship Programme. The authors also acknowledge the support of Institute of Cancer Research and Dublin Institute of Technology.
Fig. 5. Example of the OsiriX Anisotropic Diffusion Processor applied to MR image data using the OsiriX 3D Volume Rendering and cropping facility to highlight a brain tumour (white region in the top-half of the image) with CLUT: Spectrum.

Fig. 6. Example of the OsiriX Anisotropic Diffusion Processor applied to CT data using the OsiriX Surface Rendering facility to generate an iso-surface consisting of two surfaces with equal transparencies.

REFERENCES


Jonathan Blackledge graduated in physics from Imperial College in 1980. He gained a PhD in theoretical physics from London University in 1984 and was then appointed a Research Fellow of Physics at Kings College, London, from 1984 to 1988, specializing in inverse problems in electromagnetism. During this period, he worked on a number of industrial research contracts undertaking theoretical and computational research into the applications of inverse scattering theory for the analysis of signals and images. In 1988, he joined the Applied Mathematics and Computing Group at Cranfield University as Lecturer and later, as Senior Lecturer and Head of Group where he promoted postgraduate teaching and research in applied and engineering mathematics in areas which included computer aided engineering, digital signal processing and computer graphics. In 1994, Jonathan Blackledge was appointed Professor of Applied Mathematics and Head of the Department of Mathematical Sciences at De Montfort University where he expanded the post-graduate and research portfolio of the Department and established the Institute of Simulation Sciences. From 2002-2008 he was appointed Visiting Professor of Information and Communications Technology in the Advanced Signal Processing Research Group, Department of Electronics and Electrical Engineering at Loughborough University, England (a group which he co-founded in 2003 as part of his appointment). In 2004 he was appointed Professor Extraordinaire of Computer Science in the Department of Computer Science at the University of the Western Cape, South Africa. He currently holds the prestigious Stokes Professorship under the Science Foundation Ireland Programme based in the School of Electrical Engineering Systems, Dublin Institute of Technology and is Distinguished Professor at the Centre for Advanced Studies, Warsaw University of Technology, Poland.

Matthew Blackledge graduated with a degree in Physics from Imperial College, London in 2006. He moved on to pursue a career in medical imaging and completed a Masters Degree in Medical Physics with distinction from the University of Surrey in 2007. After undertaking research into Diffusion Weighted Magnetic Resonance Imaging (MRI) working with the Cancer Imaging Centre at the Institute of Cancer Research (ICR) and Royal Marsden Hospital, he was awarded a PhD from the University of London in 2012 for a thesis entitled Assessing Response in Bone Metastases using Diffusion Weighted MRI. He is now a Research Fellow at the ICR with research interests that include the acquisition and post-processing of Diffusion Weighted MRI focusing on their application for the detection of brain tumours, full-body medical imaging and combined MRI/CT/PET imaging. Teaching duties include the provision of short course programmes in Medical Physics, Software Engineering and Imaging Processing for Clinical Diagnosis.

Jane Courtney graduated with a degree in Electrical Engineering from University College Dublin (UCD) in 1998. After a year working as a project engineer in industry, she returned to UCD to pursue research in biomedical engineering and graduated with a Masters Degree in 2001 and a Doctorate in 2005. Her research has focused on using digital image processing to develop a novel gait analysis system. Current research and teaching interests include computer vision and image processing, biomedical engineering and power systems engineering. In 2004, she taught on the intervarsity MSc in biomedical engineering. She has been a member of several programme committees including the Medical Imaging and Telemedicine Conference held in Fujian, China, in August 2005, the International Machine Vision and Image Processing conference in 2006, 2007 and 2011 and the Ninth International Conference on Environment and Electrical Engineering held in Prague in May 2010. She was also a reviewer for the European Conference on Circuit Theory and Design held in August 2005 and is currently a reviewer for the Journal of Biomechanics. She is part of the Health Engineering and Rehabilitation Team and People Oriented Technology Research groups at Dublin Institute of Technology (DIT) and an adjunct member of the Vision Systems Group, Dublin City University. From September 1999 to April 2001, Jane worked as an assistant lecturer in the Department of Electrical Engineering in DIT. While pursuing her PhD, she worked as a lecturer in the Department of Electronic and Electrical Engineering at UCD returning to DIT as a full-time lecturer in September 2004.
A Cake Cutting Approach to Rezoning School Districts
Ashish Patel, Michael Shafae
California State University, Fullerton

Abstract—The task of school districting is a process of dividing the city region into school zones which is often done manually with either complete or partial human intervention. The manual process involves extensive human judgement where mistakes may lead to unfair division. We believe school districting is a fair division problem and can be algorithmically solved using computer software that employs cake cutting model. The school region can be represented as a quad tree data structure and players can be represented as software agents who actively take part in cake cutting protocol execution.

Index Terms—Cake cutting, fair division, envy-free division

1 INTRODUCTION

Given a resource, such as a cake, dividing it fairly amongst all the party goers in a fair and envy free fashion can be difficult. Often, a cake is divide by cutting the cake into pieces of approximately equal volume. In one respect, the volume of the cake, the division is fair, but this neglects how the cake recipients value the parts of the cake; the kind and amount of frosting, the cake’s layers, the quantity and variety of toppings. The problem is to divide the cake in such a way that every player in this game gets a piece of cake that is fair and envy free. In other words, every player has a slice of cake and values the slice as being no better or worse than anyone else’s slice.

Cake is a fitting analogy for a resource that different parties value differently. Ask one and it is the heaping quantity of frosting that raises the value of a cake slice. Another will wish to minimize the frosting. And another will inspect what the cake is topped with. In general we will use cake and players in the following sense:

• Cake: Any divisible set of resources that is to be distributed among one or more players. It can represent an heir of an estate, a political area, physical properties, set of work assignments, etc.
• Players: A set of one or more individuals or entities interested in one or more portions of the resources.
• Evaluation: How each piece of resource is evaluated by players is based on personal preferences. In other words, the evaluation matrix of the resources are distinct and unique for each users.

Mostly all the cake-cutting algorithms assume one-dimensional type resources. Therefore, cake has to be cut vertically and not any other ways. This is hardly the case in real situations. Most of the real world problems consist of multi dimensional resources, e.g., if a piece of land has to be divided between few players, using cake-cutting algorithms the resulting individual land pieces will be of the shape of thin stripes. Land divided this way may not be even useful to the individuals even if it is very long.

The primary assumption in a cake-cutting problem is that a piece of cake retains its value when it is cut into two or more pieces. Moreover, small pieces or crumbs can be joined back to the cake. This may not be the case in real world problems, e.g., if a school is to be assigned an area of the city that falls in the school region, the school may not prefer discontinuity in the region which may result from crumbling.

To state the obvious envy free, fair division problems abound in business, government and science. The problem that presents itself is how to divide the cake in a meaningful way so that players do not end up with a fair pile of crumbs and how to build a model of how each player values the parts of the cake. Take for instance how police, fire, and other municipal services are allocated. The disparate regions of a city are composed of citizens of varying political ideologies, socio-economic levels. However a city must allocate it’s fire fighting and police forces in a way that is fair and does not create envy amongst the different regions and demographics of a city.

More importantly, these municipal resources have already been distributed. What is of interest to both computer scientists, mathematicians and social scientists is if, after the fact, we can inspect the division of this shared resource to measure how fair and envy free the political process has been.
2 BACKGROUND

Generally there are $n$ players $1 \ldots n$ and a cake $C$. Every player $p$, where $(1 \leq p \leq n)$ has his own measure $\mu_p$ on the subsets of $C$. These measures satisfy $\mu_p(X) \geq 0$ and $\mu_p(C) = 1$ for all $X \subseteq C$, and $\mu_p(X) + \mu_p(X') = \mu_p(X \cup X')$ for all disjoint subsets $X, X' \subseteq C$. For every $X \subseteq C$ and for every $\lambda$ with $0 \leq \lambda \leq 1$, there exists a piece $X' \subseteq X$ such that $\mu_p(X') = \lambda \mu_p(X)$.

Many cake-cutting protocols are known today, both finite and continuous ones. While a finite protocol always provides a solution after only a finite number of decisions have been made, a continuous protocol could potentially run forever. Among finite protocols, one can further distinguish between bounded and unbounded ones. A finite bounded cake-cutting protocol is present if we know in advance that a certain number of steps (that may depend on the number of players) will suffice to divide the resource fairly independently of how the players may value distinct parts of the resource in a particular case and independently of the strategies chosen by the players. In contrast, in finite unbounded cake-cutting protocols, we cannot predict an upper bound on how many steps will be required to achieve the same goal. A protocol for $n$ players is called frugal, if it only uses $n - 1$ cuts.

The cake-cutting protocol can also be represented using game theory according to Robertson and William Webb [1]. Given cake pieces $X_1, X_2, \ldots, X_n$ to be assigned to players $P_1, P_2, \ldots, P_n$ we may construct a bipartite graph $G$ where players and pieces are both mapped to vertices. A player-vertex is connected to every piece-vertex that they deem acceptable. So the restricted number of cuts (that may depend on the number of players) will suffice to divide the resource fairly independently of how the players may value distinct parts of the resource in a particular case and independently of the strategies chosen by the players. In contrast, in finite unbounded cake-cutting protocols, we cannot predict an upper bound on how many steps will be required to achieve the same goal. A protocol for $n$ players is called frugal, if it only uses $n - 1$ cuts.

The cake-cutting protocol can also be represented using graph theory according to Robertson and William Webb [1]. Given cake pieces $X_1, X_2, \ldots, X_n$ to be assigned to players $P_1, P_2, \ldots, P_n$ we may construct a bipartite graph $G$ where players and pieces are both mapped to vertices. A player-vertex is connected to every piece-vertex that they deem acceptable. So the cake-cutting problem reduces to perfect matching in a bipartite graph $G$ where players and pieces are both signed to players $\mu_p$ on the subsets of $p$ player piece in $S$. In other words, for a real number $\beta$, if $\beta > 1$, we cannot predict an upper bound on the number of cuts. Even and Paz show that for $n \geq 3$ players, there does not exist a perfectly fair protocol that does only $n - 1$ cuts.

2.1 Fairness

The fairness of the protocol is the guarantee given by the protocol to all the players to receive at least some defined fraction of the whole cake. In other words, for a real number $\beta$ with $0 < \beta < 1$, a $\beta$-strategy of a player is a strategy that will guarantee him at least a fraction $\beta$ of the cake according to his own measure, independently of the play of the other $n - 1$ players. A protocol is called $\beta$-fair, if every player has a $\beta$-strategy. Krumke [2] shows that the best possible value for $\beta$ is $\frac{1}{2}$ by proving that, for $n \geq 2$ players, there exists a $\frac{1}{2}$-fair frugal protocol for cake-cutting and there does not exist $\beta$-frugal protocol for cake-cutting with $\beta > \frac{1}{(n-2)}$ [2].

If a restricted number of cuts is used, how much fairness can we guarantee? If $M(n, k)$ represents how close one can come to fair division with restricted number $k$, of cuts, Robertson and Web [3] defines and proves that:

1) $M(n, n - 1) = 1/(2n - 2)$ if $n \geq 2$
2) $M(n, n) \geq 1/(2n - 3)$ if $n \geq 3$
3) $M(n, n + 1) = 1/(2n - 4)$ if $n \geq 4$

2.2 Approximate fairness

Approximate fairness was first introduced by Robertson and Web [3]. A protocol is called $c$-fair if it guarantees each player at least a share of $\frac{1}{cn}$. Jeff and Kirk [4] shows that if protocol requires to be only a $c$-fair protocol, then the complexity of any deterministic protocol for cake-cutting is $\Omega(n \log n)$.

Jeff Edmonds and Kirk Pruhs [5] presented approximate cuts with balanced allocation of the cake with complexity $O(n)$ using a balls and bean model [3, 6].

2.3 Limiting Number of Cuts

The primary assumption in cake-cutting problem is that a piece of cake retains its value when it is cut into two or more pieces. Even though most of the protocols exist with this primary assumption, reducing the number of cuts has been an interesting mathematical research. Reducing the number of cuts also prevents crumbs or small pieces which again need to be distributed in case of full cake distribution which results in more iterations of the algorithm. The restriction of how many cuts are needed often tied with fairness. In other words, how much fairness can we guarantee with a limited number of cuts [3] is one of the areas of research which is discussed in the next section.

If cake is to be divided among $n$ players, no protocol can do with a smaller number of cuts, since $n$ pieces are to be produced [2], thus, minimum of $n - 1$ cuts are needed. Even and Paz show that for $n \geq 3$ players, there does not exist a perfectly fair protocol that does only $n - 1$ cuts.

2.4 Envy-Free

A protocol is envy-free if no single player would like to have the share that is chosen by another player for his own share. In other words, an envy-free protocol guarantees no one will want someone else’s share more than their own. When cake-cutting procedures are to be applied to real-world scenarios, fair finite bounded cake-cutting protocols are favored.

A fair division must be juxtaposed against an fair and envy-free division. Returning to the example of a cake to be divide between two players, if the cake is partitioned and assigned by a non-player, perhaps a parent, a fair division may allocate equal volumes of cake to each player but the partition is in no way envy-free. One player may covet the equally sized piece given to the other player because the color of the cake frosting is more to his liking. Although the division is fair, it is not envy-free.
2.5 Dominant Strategy
A dominant strategy occurs when one strategy is better than another strategy for one player no matter how the other players play. An envy-free cake-cutting protocol is said to be strategy-proof if a cheating player is no longer guaranteed to not envy any other player, whereas all other players are. That is, a strategy-proof envy-free cake-cutting protocol is resistant to manipulation in the sense that for a player to be guaranteed to not envy any other player, he or she is required to play truthfully. In other words, a strategy-proof protocol does not guarantee an envy-free division for the player who is cheating.

2.6 Envy-free division for three players
The Selfridge-Conway discrete procedure [7], presents a solution for three players with at most five cuts. A cake is to be divided between 3 players; P1, P2, and P3. First, P1 cuts the cake into three equal pieces X1, X2, and X3. Since for P2, all three pieces may not hold equal values, P2 orders pieces according to his own measure in the sense that for a player to be guaranteed to not envy any other player, he or she is required to play truthfully. In other words, a strategy-proof protocol does not guarantee an envy-free division for the player who is cheating.

2.7 Super Envy-free Division
A protocol is super envy-free or strong envy-free whenever for a cake division X = X1 ∪ X2 ∪ X3...Xn exists such that μi(Xj) < 1/n for i ≠ j. Barbanel [1] proved that super envy-free divisions exist if and only if the player measures are linearly independent. The result showed only existence and later Robertson and Webb show how to construct a super envy-free division by using a set of pieces that witness the linear independence of the measures [2].

2.8 Degree of guaranteed envy-freeness
Linder [8] introduces the notion of guaranteed envy-freeness which is a measure of how good a cake-cutting protocol can approximate the ideal of envy-freeness while keeping the protocol finite bounded. Here, Linder proposes that the protocol so far only focus on quantifying fairness and not the envy-freeness and thus introduces the notion of degree of guaranteed envy-freeness which represents the total number of envy-free relations among the players. Furthermore, they provide a strategy-proof finite bounded proportional cake-cutting protocol with a degree of guaranteed envy-freeness of 10 for four players. Linder then provides a proportional protocol with an enhanced degree of guaranteed envy-freeness of ⌈n²/2 + 1⌉ for n > 3 players.

3 Previous Work
The cake-cutting problem can be reduced to its simplest form by having just two players. To paraphrase [9], if cake C is to be divided among two players, P1 and P2, a trivial envy free algorithm can be described as follows:
1) P1 cuts the cake in half the value according to his own assessment
2) P2 gets to choose the first piece.
3) P1 gets the remaining piece.

Common operations in devising a fair division algorithm are evaluate and cut. The evaluate query, the player is provided with a piece of cake and is asked to evaluate its value according to his own assessment strategy. The cut operation gives a player a piece or entire cake and is asked to cut it according to a specified proportion.

This trivial algorithm can be adapted to an n-way division, but the running time is unacceptable for most applications. To illustrate, consider a four way division. P1 first cuts 1/4 of the cake and passes it to P2. P2 can either accept this piece or pass it on to P3. This continues until a player either accepts the piece or it is rejected by all players except the one who cut the piece. Should all players reject the piece, the player who cut the piece then accepts the piece and is removed from the queue of players waiting for cake. This process continues until everyone has a piece. Should the player cut a stingy piece, the piece returns to the cutter. A generous piece will be evaluated and accepted by a player down the line.

In successive pair algorithms [3] algorithm, the pieces are first divided among two players and then n – 2 players are added one by one. If there are m existing players, and one more user arrives, then existing m players will make m + 1 cuts of their individual pieces and the new player will choose one piece from each player. This algorithm make n! cuts for n player. This is a simple algorithm for dividing the cake among n players; however, it requires many cuts. For example, for only 5 players it requires 120 cuts.

In the moving knife algorithm, described [9] in algorithm, the knife is moved from the leftmost end of the cake slowly but gradually towards the right. Whenever someone says stop, a cut is performed and the piece to the left of the knife is given to that person and is dropped out of the game. For the remaining portion, the same procedure is applied. This algorithm requires only n – 1 cuts; however, in reality this protocol is infinite because each player has to continuously evaluate the piece on the left of the knife since the value of the cake is 0 to 1 and is discrete. Therefore, there might be infinite steps of evaluation for each player. Hence, this algorithm is not considered in computational algorithms even though it is fairly easy in practice.
In the trimming algorithm, described [9], the first player starts with cutting a piece of cake worth \( \frac{1}{n} \) according to his own measure. After that, the piece is passed to the next player. If the next player thinks it is bigger than \( \frac{1}{2} \) of his own measure, he should cut it and pass the piece to the next player. Performing this exercise, when all the players are done, the piece is assigned to the last person who trimmed the cake and that player drops out of the game. The above steps are performed until all the players are assigned pieces. The crumbs that result from this procedure are added back to the cake. This algorithm requires \( \frac{n(n-1)}{2} \) cuts.

The divide and Conquer [10] is one of the most computationally effective methods for dividing a cake among \( n \) players. If \( n \) is even, ask the first \( n-1 \) players to cut the cake in the ratio of \( 1 : 1 \). If \( n \) is odd, ask first \( n-1 \) players to cut the cake in the ratio \( \frac{n+1}{2} : \frac{n+1}{2} \). And then ask which piece the \( n \)-th player thinks is worth at least its declared value. Once the \( n \)-th person chooses which side he wants to belong to, recursively use the same above method until the problem is reduced down to a cut-and-choose method. In other words, the \( n \)-player game is reduced to 2-or-less player game and then cut and choose is performed to divide the piece of cake into each two players. This method requires \( \lceil (n \log 2n) \rceil \) total cuts which is a significant improvement over the trimming algorithm and the successive pair algorithm. For five players, this divide and conquer algorithm uses 8 cuts. For a higher number of cuts, this algorithm is far better than a successive pair and trimming algorithm.

4 SCHOOL DISTRICTS

Again, the problem that presents itself is how to divide the cake in a meaningful way so that players do not end up with a fair pile of crumbs. Central to this is the challenge to build a model of how each player values the parts of the cake. In Section 1, we touched upon a hypothetical situation where a municipality wishes to devise a fair and envy-free allocation of police resources. In an effort to better understand the application of cake cutting algorithms to a resource allocation problem, we set upon the problem of defining school zone boundaries within a school district.

A school district, in the United States of America, is a form of special-purpose district which serves to operate local public primary and secondary schools [11]. A special-purpose district is an independent governmental unit that exists separately from general purpose local governments, such as municipal or county governments, and exercise substantial administrative and fiscal independence from local government.

A school zone is typically a contiguous geographic region within a school district which is served by a school. In the case that we are studying, the school zone is constrained to be a contiguous geographic region due to the constraints we place on the players within the cake cutting algorithm. Additionally, our cake cutting game has a fixed number of existing schools and the zone for each school is redrawn.

Thus school districting is a cake-cutting problem where the cake is the school district to be divided into contiguous regions which each have a school that serves the population within the region. The players are the schools who are interested in dividing the district’s area. Granted, with the existing definition of the problem, a naive partitioning could achieve a fair division. To make the partitioning envy free, we have the different players have a variety of preferences for attributes such as population, parental education background, median home prices, point-to-point distance from the school, etc. Balancing these factors can create a fair and envy-free allocation of school resources.

Our example school district is the one found near our university, the Fullerton School District. Zoning of schools is accomplished by a school district districting commission. Some of the criteria that such a commission considers in zoning a district are:

- Travel time of students
- District geographic boundaries
- Instructional and non-instructional support services
- Budgeting
- Contractual obligations
- Capital debt
- Enrollments
- Political and municipal boundaries
- Tax rates
- Student demographics
- English language learner percentage
- Small and/or isolated designation
- Expenditure per pupil

The current Fullerton school district zones for elementary schools is shown in Figure 1. The boundary data and other geographic information was provided to us by the Orange County Department of Education, http://www.ocde.us/

Inspecting this map, one can see there are clearly unfair divisions within the district. For example, Beechwood Elementary’s zone is stretched towards the south, which causes a longer travel distance for students stay-
ing in the south of this school zone. Additionally, Laguna Road Elementary school zone has students who live closer to this school who are part of another zone and students that are farther away are part of this school zone.

5 Methodology

In order to represent the population data as cake-like resources, we need to have geo-location information along with the population data. Thus our resources for cake-cutting minimally contains the following information:

- Home location
- Number of school age children
- Community to which this home belongs
- Educational background of parents
- Detailed geographic information, e.g., road, city, lake, river, etc.
- First and second language spoken at home
- General student information, e.g., is special education needed? Handicapped?
- Political, municipal, county, and city boundaries

Schools are the players in the cake-cutting game. Each school has different criteria on how to choose the piece of district area based on its preference and available attributes of the parcel in question. School data minimally encapsulates the following information:

- School name
- School location
- School capacity
- Student diversity preference
- Family education background preference

The cake represents a non-uniform distribution of the resources. Similarly, the population data are also a non-uniform distribution of resources. These attributes are implied from the school preference data and population data. The population data are non-uniform data and each of the schools has it's own metric for evaluating a parcel of the district. The following are a few of the resource attributes that are analogous to cake toppings.

- Distance from home or community to a school. Schools prefer their student to live at minimum driving distance.
- Capacity of the schools and thus willingness to accept students from a given piece of city area.
- Per student funding from city. Some cities may pay higher funding per student and some may pay lower.
- Ethnicity mix of the city area.
- Education background of parents.
- Population of students.
- City the given piece belongs to.
- County the given piece belongs to.
- Student grades and individual capacity for each grades in the school.
- Number of students whose first language is non-English.
- Number of students who need special education.
- Political boundary the given piece belongs to.

The population of a given school district area is a two dimensional resource. The cuts on this resource is tagged with geo-location information which can be used to merge a crumb back into the most appropriate set. Additionally, this resource can only be broken down to atomic units of a single family address or a multi-family domicile such as an apartment building.

The entire school district is represented by a quad tree data structure. During the protocol execution, if the piece of the city region is not manageable, it can be broken down into four subregions. This process recursively happens until the region cannot be subdivided any further. There are many conditions protocol may choose not to subdivide the region any further. Some common examples are:

- The region has only one apartment complex.
- The region has only one gated community.
- The region has only a few home locations, e.g., region has only one home location.
- The protocol wants to limit the segmentation for performance reasons, e.g., a protocol might choose not to subdivide any further if the number of home locations or the number of children in a given quad is less then 10.
- The region represents empty space, e.g., empty land, river, golf course, or business region.

A cake-cutting model consists of players who are interested in one or more parts of the resources. In this case, each school is represented by a software agent. These agents will actively take part in the cake-cutting protocol. Primarily, every player in the cake-cutting protocol and thus every agent in the cake-cutting software, performs three major tasks: evaluate, cut, and choose.

The evaluation operation is accomplished by each agent by evaluating the school’s preferences. We call this metric the school’s preference scale. The preference scale of a parcel under consideration by a given school is calculated as:

\[
\text{Preference Scale} = \sum_{i=0}^{k} \frac{F_i(P_i, W_i, Q_i)}{D} \quad (1)
\]

where \(i\) is the school’s preference index 1 to \(k\), \(P_i\) is the preference value of \(i\)th preference of school, \(W_i\) is the weight factor of \(i\)th preference of the school and \(\sum_{i=0}^{k} W_i\) is the same for all the schools, \(Q_i\) is the set of parcel attributes related to the \(i\)th preference of the school, e.g., if preference is capacity, the related attribute of the parcel will be the number of students in a given parcel, \(D\) is the great circle distance from parcel center to the school, and \(F_i\) is the preference scale numerator function for \(i\)th preference.

Table 1 and Table 2 show an example of preference scale calculations. Table 1 shows the parcel data that is presented to the agents of two schools, Maple and
Laguna. Table 2 shows criteria and preference scale numerator calculations of these two schools. Two preference scale numerator are calculated as below.

\[ Fc = \frac{\text{Capacity}}{\text{Population}} \times \text{CapacityWeightFactor} \]

\[ Fpsf = (PSF_{\text{pro}} - PSF_{\text{req}}) \times \text{Population} \times PSF_{\text{reqWeightFactor}} \]

where, \( Fc \) is the numerator function for capacity criteria, \( \text{CapacityWeightFactor} \) is the weight factor for capacity criteria, \( Fpsf \) is the numerator function for per student funding criteria, \( \text{PSFpro} \) is per student funding provided by the city, to which the given parcel belongs to, \( \text{PSFreq} \) is per student funding required by the given school, and \( \text{PSFREQWeightFactor} \) is the weight factor for \( \text{PSFREQ} \) criteria of the school.

Based on the equations above, Table 1 shows the preference scale numerator values for each preference of each school. The final value of Laguna School’s numerator came out to be negative, which means, if the school gets this parcel, it may end up having a loss, since the per student funding required for Laguna school is $150, whereas, the quad piece presented offers only $140 per student funding. Therefore, Laguna school may encounter a loss of $10 per student. Maple school may like to have this piece because they are getting more funding per student than it requires. Thus Maple will be the winner of this piece and hence this parcel will be allotted to the Maple school.

### TABLE 1
Quad piece data for calculating Preference scale

<table>
<thead>
<tr>
<th>Quad piece data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>15</td>
</tr>
<tr>
<td>Per student funding provided by city (PSFprov)</td>
<td>$140</td>
</tr>
</tbody>
</table>

### TABLE 2
School data for calculating Preference scale and Preference scale numerator calculation

<table>
<thead>
<tr>
<th>School data</th>
<th>Maple</th>
<th>Laguna</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity (C)</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Capacity weight factor</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Per Student Funding Required (PSFREQ)</td>
<td>$120</td>
<td>$150</td>
</tr>
<tr>
<td>PSF weight factor</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preference scale calculation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Value of capacity function ( Fc )</td>
<td>1000</td>
</tr>
<tr>
<td>Value of PSF function ( Fpsf )</td>
<td>6000</td>
</tr>
<tr>
<td>Final preference scale numerator ( (Fc+Fpsf) )</td>
<td>7000</td>
</tr>
</tbody>
</table>

5.1 Rezoning School District

Each school is modeled as a software agent with its own preferences. The school district itself is modeled as a hierarchy of land parcels that can be cut up amongst the competing schools.

Due to the hierarchical nature of our data and the better running time when compared to other \( n \)-way cut-and-choose algorithms, a divide and conquer algorithm is the best choice. In this algorithm, every time the cut is made, the complexity is reduced by \( \frac{1}{2} \). A similar approach can be considered when using the quad tree data structure. Here we subdivide the parcels recursively until they cannot be subdivided any further. After that, agents evaluate each piece based on their preference and available attributes of the resource pieces. At the end of the execution of the protocol, the agent who has the highest claim for a given piece, is the winner of that piece.

For any agent, the set of won pieces are then assigned to that agent’s school, which we can call a cut operation. The formal cake-cutting operations are performed as below.

5.2 Mark Operation

A mark operation is different from the cut operation. This operation simply divides the region into smaller parcels such that players can evaluate. These marks are logical cuts and do not change the state of the parcel in question until after the proposed division is evaluated by the players.

Since the geographic information of the school district is stored in a quad tree, a typical subdivision would take any given parcel and recursively divide it into four equal parts. Since we wish to minimize the formation of crumbs, attributes of the parcel are considered to fairly divide the parcel. The following are a few of the attributes that influence the subdivision of a parcel.

1) City boundary: If city boundary information is available, the region can be subdivided exactly from the boundary.
2) Major road: If the road information is available and a major road is present near the middle of a region, the region should be subdivided exactly from that road. This prevents fragmentation which may result in the larger number of crumbs.
3) Lake, river, golf course, cemetery, business area, etc.: In a case where a region has a mix of residential and other non-residential areas like river, lake, etc., subdivision is performed from the identified border.

5.3 Evaluate Operation

Each agent is given a parcel of the school district region and is asked to evaluate it based on its preferences. This operation involves an agent’s decision on how much it values the given piece of city region. The agent may have to query a few or more of the following questions to the given piece.

1) How many students are there in a given region?
2) What is the average education background of parents in a given region?
3) What cities does this region include?
4) What is the ethnicity-mix of a given region?
5) How far is this region, or what is the average driving distance for homes in this region to the school?
The result of the agent’s evaluation of the piece is called preference scale. The preference scale is a measure of how much the agent values a given parcel. The preference scale is a normalized value between 0 to 1.

5.4 Choose Operation
The choose operation is performed automatically at the end of the protocol without the agent’s intervention, since an agent’s job is only to evaluate a given parcel. At the end of the protocol, the indivisible pieces of school district area are assigned to an agent that has the highest preference scale. The group of subregions of school district area are assigned to an agent who is the winner, which is equivalent to the cut operation.

6 EXPERIMENTAL RESULTS
First and foremost, the rezoning of a district can only be as good as the input data regarding the schools’ preferences and the geographic and demographic data of the district’s constituents. That being said, we have gone to great lengths to collect the best data possible from the Fullerton School District as well as from other state and federal sources to have the best model possible. However, we believe the data collected is far from ideal.

The data we collected can be broken down into two coarse grain categories, population data and school data.

The population data of the Fullerton city region are collected from GIS department of City of Fullerton. The raw data set contains 47,000 home addresses. This dataset does not include business locations, geography information, and detail population information. Addresses from this dataset are then geo-coded to represent the exact location of each home in the form of latitude and longitude.

6.1 School Data
With the assistance of the Orange County Department of Education, we replicated the boundaries of the Fullerton Elementary Schools in our own GIS software. Figure 2 shows the replicated school map in our application. Since school data are bare geolocations and boundaries, the capacity of each school is added manually to the school data. The distance criteria is built-in and can be derived from distance from the school location to home location. The school data are then used to create agents that will actively take part in the protocol and perform evaluation operations on the given piece of school district area provided by the protocol. The output data produced by the application are in the form of XML file.

6.2 Constraints
Because of limited availability of data, following are the constraints considered in the prototype.

6.3 Protocol Execution
The divide and conquer protocol for school districting applies two criteria for evaluating the city region or quad piece: distance and school capacity. The following are few results of experiments with these two criteria.

When only the distance from school to the parcel is considered, the agent prefers to have parcels that are closer to the school. As the distance from school to the parcel increases, the preference of the agent for that parcel decreases. The distance criteria for all the schools remains the same. Distance is a primary criteria in calculating the preference scale. The redistricted school with only distance criteria is as shown in Figure 3.

School capacity criteria can be added as a relative capacity scale, e.g., if school A’s actual capacity is 100 and school B’s actual capacity is 300 and school A’s relative capacity can be said to be 1 and school B’s relative capacity said to be 3. When capacity criteria are added, schools are not guaranteed to be allocated as many students. The capacity criteria adds an additional factor for winning the parcel for a given school compared to other schools, e.g., if a given parcel is of exactly the same distance from school A and B, then schools’ capacities will play a significant role in preference scale value provided by both of the school agents. Figure 4 shows how adding the capacity criteria to the school preference affects school districting. Table 4 shows the division of school zones with only distance criteria and
Table 4 shows the division with distance and capacity criteria. Since distance is built in as geographical property of the data and also a primary criterion; it cannot be changed. Other criteria can be changed as per schools’ request. Figure 5 shows how changing the capacity of Laguna School affects the division. As the capacity of Laguna School is increased from 1000 to 2000, the allocated number of homes increases from 615 to 1986.

7 CONCLUSION & FUTURE WORK

Cake cutting algorithms have been an interesting area of research in the domain of mathematics since 1947. However, applications of cake cutting algorithms to quotidian social and economic problems has had little exploration.

First of all it is important to note that cake cutting, when dealing with $n$-way envy free, fair division, has two main drawbacks. There is initially the computational cost both in terms of space and time. Second, there is the practical consideration of how to divide the resource to avoid or minimize crumbs. Lastly, modeling player desires is far from a trivial task.

However, $n$-way envy free, fair division is something that is taxing and difficult to accomplish on its own. Cake cutting offers not a solution but a way to measure how good or poor a negotiated division is when compared against an ideal born of mathematical rigor and dispassionate computation. In the study we have conducted of the school zones in the Fullerton School District, one can compare the existing division of resources against a number of idealized schemes, assess and build strategies that can lead to better utilization of resources.

We are currently considering where to take this work next. In our current system we would like to add greater
TABLE 4
Division with two criteria: Distance and School Capacity

<table>
<thead>
<tr>
<th>School Name</th>
<th>Students Allotted</th>
<th>School Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SunsetLane</td>
<td>2434</td>
<td>1500</td>
</tr>
<tr>
<td>Filer</td>
<td>1631</td>
<td>1500</td>
</tr>
<tr>
<td>FernDrive</td>
<td>3395</td>
<td>2000</td>
</tr>
<tr>
<td>Laguna</td>
<td>615</td>
<td>1000</td>
</tr>
<tr>
<td>GoldenHill</td>
<td>2257</td>
<td>1800</td>
</tr>
<tr>
<td>Hermosa</td>
<td>2217</td>
<td>2000</td>
</tr>
<tr>
<td>Beechwood</td>
<td>2533</td>
<td>2000</td>
</tr>
<tr>
<td>Valencia Park</td>
<td>1272</td>
<td>1200</td>
</tr>
<tr>
<td>Pacific Drive</td>
<td>2021</td>
<td>1800</td>
</tr>
<tr>
<td>Richman</td>
<td>2234</td>
<td>2000</td>
</tr>
<tr>
<td>Maple</td>
<td>778</td>
<td>1300</td>
</tr>
<tr>
<td>Raymond</td>
<td>2763</td>
<td>2000</td>
</tr>
<tr>
<td>Commonwealth</td>
<td>3166</td>
<td>2500</td>
</tr>
<tr>
<td>Acacia</td>
<td>1503</td>
<td>1000</td>
</tr>
<tr>
<td>Rolling Hills</td>
<td>2833</td>
<td>1800</td>
</tr>
<tr>
<td>WoodCrest</td>
<td>1873</td>
<td>2000</td>
</tr>
<tr>
<td>OrangeThorpe</td>
<td>2453</td>
<td>2000</td>
</tr>
<tr>
<td>Total</td>
<td>35978</td>
<td>29400</td>
</tr>
</tbody>
</table>

TABLE 5
Division with two criteria: distance and school capacity. As capacity of Laguna school changed, the division result changes. Compare the division result with Table 4

<table>
<thead>
<tr>
<th>School Name</th>
<th>Students Allotted</th>
<th>School Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SunsetLane</td>
<td>2204</td>
<td>1500</td>
</tr>
<tr>
<td>Filer</td>
<td>1631</td>
<td>1500</td>
</tr>
<tr>
<td>FernDrive</td>
<td>3147</td>
<td>2000</td>
</tr>
<tr>
<td>Laguna</td>
<td>1986</td>
<td>2000</td>
</tr>
<tr>
<td>GoldenHill</td>
<td>2025</td>
<td>1800</td>
</tr>
<tr>
<td>Hermosa</td>
<td>1873</td>
<td>2000</td>
</tr>
<tr>
<td>Beechwood</td>
<td>2288</td>
<td>2000</td>
</tr>
<tr>
<td>Valencia Park</td>
<td>1272</td>
<td>1200</td>
</tr>
<tr>
<td>Pacific Drive</td>
<td>2021</td>
<td>1800</td>
</tr>
<tr>
<td>Richman</td>
<td>2234</td>
<td>2000</td>
</tr>
<tr>
<td>Maple</td>
<td>778</td>
<td>1300</td>
</tr>
<tr>
<td>Raymond</td>
<td>2763</td>
<td>2000</td>
</tr>
<tr>
<td>Commonwealth</td>
<td>3166</td>
<td>2500</td>
</tr>
<tr>
<td>Acacia</td>
<td>1503</td>
<td>1000</td>
</tr>
<tr>
<td>Rolling Hills</td>
<td>2833</td>
<td>1800</td>
</tr>
<tr>
<td>WoodCrest</td>
<td>1873</td>
<td>2000</td>
</tr>
<tr>
<td>OrangeThorpe</td>
<td>2453</td>
<td>2000</td>
</tr>
<tr>
<td>Total</td>
<td>35978</td>
<td>29400</td>
</tr>
</tbody>
</table>

geographic detail and more controls into how schools select their zones. Additionally, complicating the cake cutting game with charter schools as well as student feeder pattern (how students move from primary to secondary schools) would open the door to more interesting software agent behaviors. Furthermore, from our initial prototype application we intend to analyze the current zoning and attempt to derive a measurement of fairness and envy-freeness.

REFERENCES

Targeting Cell Nuclei for the Automation of Raman Spectroscopy in Cytology

Jonathan Blackledge, Dmitri Dubovitskiy and Fiona Lyng

Abstract—Biological cell analysis has, and is still, an important aspect in medical research and clinical diagnosis. Although cytologists routinely undertake a diagnosis using optical microscopy, human factors make this routine unreliable especially when it involves many consecutive tasks that are monotonous, time consuming and focus on pattern matching tasks where the patterns concerned are not always entirely clear and/or do not necessarily belong to a well defined class. Raman Spectroscopy provides the potential to generate a fundamental representation on the status of cellular conditions through the characteristics of a Raman spectrum generated by the back-scatter of a laser pulse incident on the cell nucleus. However, this approach requires the nucleus of the cell to be accurately targeted from a complex of many hundreds of such cells within a conventional optical field of view as defined by the resolving properties of a microscope. This requires specialist digital image processing methods to be developed and in this paper we discuss a new approach to the processes of object detection, recognition and classification for target detection in cytology using Raman Spectroscopy. In particular, we report on a system designed for the inspection of slides used in a cervical cancer screening system known generally as a ‘Pap-smear’ test. After providing a short introduction to the pattern recognition in general, we present a unique procedure for automating the targeting process based on an analysis of the principal issues associated with object recognition which include the basic model used and segmentation algorithms derived from the model.

Index Terms—Pattern analysis, segmentation, object recognition, image morphology, fuzzy logic, cervical cancer screening, Raman spectroscopy

I. INTRODUCTION

Approximately 471,000 women are diagnosed with invasive carcinoma of the cervix each year and 233,000 die from the disease, worldwide. Cervical cancer is among the most common female cancers in many countries of the developing world. Sexually transmitted infection by certain strains of the human papilloma virus is a major cause of cervical cancer; smoking has also been linked to the disease. In Ireland, for example, on average 800 women are diagnosed with cervical pre-cancers and 180 women are diagnosed with cervical cancer. It is estimated that 2,900 women are living with cervical cancer and 31,200 women have had cervical pre-cancer. The incidence of Gynaecological cancers has been predicted to increase by 98% by 2020 (from 855 to 1676) [37]. A National Cervical Screening Programme, Cervical Check, was launched in September 2008 and initially the Irish government outsourced the screening to Quest Diagnostics, US. In April 2010, Sonic Healthcare, the largest pathology laboratory provider in Australia/New Zealand and Europe and the third largest provider in the United States were awarded half of the screening contract and have since set up a pathology laboratory in Dublin. The technology described in this paper, if shown to be accurate in future clinical trials, could provide a cost effective and high-throughput screening system which could influence the future of screening services world-wide.

Cervical cancer is preceded by a precancerous condition called CIN (Cervical Intraepithelial Neoplasia) which can be easily treated if detected. It is therefore important to identify CINs by screening women. The screening test is called a cervical smear. A clinician removes a small sample of cells from the surface of the cervix and spreads the sample onto a glass slide, the material being ‘fixed’ in alcohol. The slide sample is typically stained and examined using an optical microscope, reports being provided on any abnormal cells or cell clusters. This technique has a false negative error rate of 15-59% [4]. There is undisputed evidence that cervical cancer is related to human papillomavirus (HPV) infection [5]. Despite the recent introduction of Gardasil and the planned introduction of Cervarix, both HPV vaccines, routine cervical screening is still recommended as the vaccines do not protect against all high risk HPV strains and some women may not benefit from the vaccines if there is a pre-existing infection or if they do not receive the complete number of doses.

Although staining the sample can provide valuable colour identifiers for automatic image recognition, the approach discussed in this paper is entirely stain independent due to the influence that staining has on the generation of a Raman spectrum [6]. Colour independent pattern recognition methods are therefore considered to detect and target abnormal cells (in particular the cell nucleus). The use of automated recognition systems of the type reported here has the potential to provide a new level of stability and robustness in screening procedures, thereby making national cervical screening programmes a possibility in countries where they are not currently practically viable.

II. CURRENT METHODS

Currently, a Papanicolaou (Pap) smear is used to screen for Cervical Intraepithelial Neoplasia (CIN) and cervical cancer in the general female population. There are two main approaches; manual screening and automated screening. The Papanicolaou
test (also called Pap smear, Pap test, cervical smear, or smear test) is a screening test used to detect cancer and pre-cancer in the ectocervix. A tool is used to collect the cells from the cervix and after placing on a slide and staining the cells are examined for abnormalities under the microscope by highly trained personnel. The Pap test is an effective, widely used method for early detection of cervical cancer and pre-cancer. However, it is widely acknowledged that sensitivity values are low (with sensitivity and specificity values of 72% and 94%, respectively [7]). Since the mid 1990s, liquid based cytology has been introduced where the cells are placed in a liquid medium which preserves the cells and this sample is processed in the cytology laboratory into a monolayer of cells on a slide. The sample is then stained and examined by microscopy and by highly trained personnel as before. There are two main types of liquid based cytology, SurePath (TriPath Imaging / BD Biosciences) and Thin Prep (Cytyc Corp / Hologic).

A more recent innovation is the introduction of automated imaging systems (ThinPrep Imaging System or Focal Point Slide Profiler) that automatically scan the slide for large and dark nuclei which are features of abnormal cells. Generally the slide is scanned twice, once by the imaging system and then manually by the cytologist. The imaging system identifies areas of interest based on cellular DNA content that are subsequently reviewed by the cytologist. The ThinPrep Imaging system has been shown to be at least equivalent to manual screening [8] while some diagnostic areas demonstrated an increased sensitivity and specificity [9]. A marked improvement in productivity has consistently been found. After automated screening each sample is reviewed by a cytologist and if any abnormalities exist, the sample is passed to the Chief Medical Scientist and then to a pathologist. This means that each sample is screened at least twice and potentially four times if there is a suspected abnormality. Despite the introduction of automated imaging systems, some clinical laboratories even choose to manually screen each sample in addition to the automated screening.

Raman spectroscopy is a powerful tool that can generate a biochemical fingerprint of a sample in a rapid and non-destructive manner. It is sensitive to subtle biochemical changes occurring at the molecular level allowing spectral variations corresponding to disease onset to be detected. There is convincing evidence that Raman spectroscopy can be used as a diagnostic tool to identify spectral changes in malignant and premalignant cells. In recent years, Raman spectroscopy has been used in the detection of a variety of cancers including, breast, lung, brain, colon, liver, oral, oesophageal, prostate and cervical cancer [1], [2]. However, finding the abnormal cells prior to Raman spectroscopy can be challenging and can require specialist personnel to correctly identify the abnormal cells. In order to automate the approach, a system which combines Raman spectroscopy with novel automated image processing is required. This cell detection approach can provide more complete information about whole cells. Further, a learning algorithm based on a pathologists/cytologists experience and expertise can also be employed to provide a greater degree of accuracy for the system. The specific advantages of such a technology are:

• higher throughput and increased speed due to automated cell detection;
• higher accuracy due to Raman analysis of biochemical fingerprints;
• significantly reduced work-load resulting in lower costs to health service providers.

In order to accomplish this, advanced pattern recognition for optical microscopic images are required based on a suitable imaging model.

III. PATTERN RECOGNITION

Pattern recognition is a component of image analysis which involves the use of digital image processing methods designed in an attempt to provide a machine interpretation of an image, ideally, in a form that allows some decision criterion to be applied [12], [13]. It uses a range of different approaches that are not necessarily based on any one particular theme or unifying conceptual framework. This is because there is no complete and unique theoretical model available for explaining and simulating the processes of human visual image comprehension. Hence, machine vision remains a subject area in which automatic inspection systems are advanced without having a fully operational theoretical framework as a guide. For this reason, numerous and non-related algorithms for understanding two- and three-dimensional objects in a digital image have and continue to be researched in order to design systems that can provide reliable automatic object detection, recognition and classification in an independent environment (e.g. [14], [15], [16] and [17]).

Pattern recognition can be thought of as the process of linking parts of the visual object’s field with stored information or ‘templates’ with regard to a pre-determined significance for the observer. There are a number of questions that need to be considered in the development of any machine vision system. These include: (i) what are the goals and constraints? (ii) what type of algorithm or set of algorithms is required to complete the system? (iii) what are the implications for the processes, given the types of hardware that might be available? (iv) what are the levels of representation required? The levels of representation are dependent on what type of segmentation process can and/or should be applied to an image. These are recorded as place tokens and stored in a database. Regions of pixels with similar intensity values or sets of lines are obtained by isolating the edges of an image scene and computed by locating regions where there is a significant difference in the intensity. Such sets are subject to inherent ambiguities when computed from a given input image and associated with those from which an existing data base has been constructed. These ambiguities can only be overcome by the application of high-level rules based on how humans interpret images, but the nature of this interpretation is not always able to be clearly defined. Parts of an image may tend to have an association if they share size, figural similarity, continuity, shading and texture. For this reason, it is necessary to consider how best to segment an image and what form this segmentation should take. For example, optical microscopy involves the use of image processing methods that are often designed in an
attempt to provide a machine interpretation of a biological image, ideally in a form that allows some decision criterion to be applied, such that a pattern of biological significance can be recognised [12], [13]. Compared to image processing, computer vision is more than automated image processing. It results in a conclusion, based on a machine performing an inspection of its own. The machine must be programmed to be sensitive to the same aspects of the visual field as humans find meaningful. In this context, segmentation is concerned with the process of dividing an image into meaningful regions or segments. It is used in image analysis to separate features or regions of a pre-determined type from the background and it, in most cases, the first step in automatic image analysis and pattern recognition. Segmentation is broadly based on one of two properties in an image: (i) similarity; (ii) discontinuity. The first property is used to segment an image into regions which have grey (or colour) levels within a predetermined range. The second property segments the image into regions of discontinuity where there is a more or less abrupt change in the values of the grey (or colour) levels.

Pattern recognition can be considered to be a form of machine understanding based on assigning a particular class to an object. The tasks of construction and application of formal operations for numerical or character representation of objects in a real or idealized world is the basis for pattern recognition. This depends on establishing equivalence relations that express a fit of evaluated objects to any class with independent semantic units. The recognition classes of equivalence can be set by the user in the construction of an algorithm, which uses selective representations or external padding information on a likeness and difference of objects in the context of a solved task. This is the basis for phrase ‘recognition with the teacher’. For a typical object recognition system, the determination of the class is only one of the aspects of the overall task. In general, pattern recognition systems receive data in the form of ‘raw’ measurements which collectively form a stimuli for the generation of a ‘feature’ vector [18], [19]. Uncovering relevant attributes in the elements present within the feature vector is an essential part of such systems. An ordered collection of relevant attributes which most clearly represent the underlying features of the object is assembled into the feature vector. In this context, learning amounts to the determination of the rules of associations between the features and attributes of a pattern.

A. Practical Image Recognition Systems

Practical image recognition systems generally contain several stages in addition to the recognition engine itself. The recognition represents information processing that is realised by some converter of the information having an input and output. On input, such a system establishes information on the properties of an object. On output, the information shows which class or feature of an object is to be assigned.

When a computerised system decides on the task of classification without engaging external learning information, it is called automatic classification - ‘recognition without the teacher’. The majority of algorithms for pattern recognition require the engagement of a number of computational procedures which can be provided only with high-performance computer equipment [20].

There are two principal methods for object recognition using either a parametric or non-parametric approach. Statistical voting and alphabetic proposition methods have been reviewed in [21], [22] and [23]. The main disadvantage with this method is that classes have to be clearly defined so that no overlapping is allowed. Methods based on a principal of separation and potential functions can be found in [36] and [16]. These methods require a large amount of training data or preliminary information about system to be generated which makes the recognition process less flexible. In general, there is no method or, at least, an operation system, which considers objects from the point of view of a superposition of global scenery. This leads to the following problem: how can we evaluate an object in terms of it being part of the ‘bigger picture’ without losing specific details on its particular texture for precise recognition? This includes the incorporation of concepts from Fractal Geometry [38], [39], [40], and [41] and Fuzzy Logic [44], [45] and [46].

B. About this Paper

The main technical objectives associated with the approach developed in this paper are:

- To develop image processing algorithms to identify suspicious cells in a cervical smear sample;
- to develop an Application Program Interface (API) to control a Raman microscope and integrate this with Raman spectrometer and image processing software;
- to develop a user friendly Graphical User Interface;
- to test the system for automated cervical cancer screening.

In this context, we consider an approach to object detection in an image scene that is based on a new segmentation algorithm using a Contour Tracing Algorithm [48] and a Space Oriented Filter. Because some parts of the image need enhancement, a novel self-adjustable filter for isolated feature sharpening is also developed. The segmented object is then analysed in terms metrics derived from both a conventional Euclidean geometric and textural perspective, the output fields being used to train a fuzzy inference engine and the recognition structures being based on some of the technologies for image processing, analysis and machine vision reported in [23], for example. The approach considered is generic in that it can, in principle, be applied to any type of imaging modality used to develop a membership function. The unique approach reported in this paper (which represents the main contribution to the field) is to develop a focal membership function, obtained by focusing a microscope at different levels in the image plane. It is this approach that is used to differentiate the nuclei of cells from the surrounding material in the image plane and thereby generate the target required. However, to place this approach and the methods developed in the context of the application required, we begin with a short overview of Raman Spectroscopy in Cytology.
IV. RAMAN SPECTROSCOPY IN CYTOLOGY

A Raman spectrogram is generated by the inelastic scattering of monochromatic (laser) light and is characteristic of the vibrational, rotational, and other low-frequency modes of a material that has been penetrated by the light (it is a volume scattering effect). The interaction of the molecular structure with laser light, yields photons with different energies. This results in the generation of a wavelength spectrum that is characteristics of the vibrational modes of the molecular structure of a material. The ‘material’ can include biological entities such as cells and cell nuclei. Raman scattering is a form of weak scattering and a principal task in Raman spectroscopy is to separate the elastic scattered light from the inelastic ‘Rayleigh scattered’ light which is a dominating effect [25]. Compared to Rayleigh scattering (i.e. an inelastic volume scattering process in which the material properties are taken to be inhomogeneous), modelling inelastic scattering processes such as Raman scattering is significantly complex. Nevertheless, even in the absence of a full working physical model, if the characteristics of a Raman spectrum can be correlated with material states that are biologically significant, and, this correlation is statistically significant from one measurement to the next, then the method can be used effectively for diagnostic purposes.

Raman spectroscopy has been used in the detection of a variety of cancers [1], [2] including cervical cancer; [24], [26] and [27], and shown to be a powerful diagnostic tool for cervical tissue sections and cervical smear samples, e.g. [28], [29] and [30]. This clinical application of Raman spectroscopy for cervical cancer has been the subject of both a UK and International Patent [31].

Fig. 1 shows a typical example of the type of Raman spectra that are generated by cell types in a cervical cytology sample. Subtle differences in the signature associated with Raman scattering from nucleic acids, proteins and lipids can be observed. It is these signatures that form the basis of Raman spectroscopy in cytology.

Irrespective of the spectral analysis algorithm(s) used, a principal problem is to maintain consistency so that the results are statistically significant, a significance that predicated on a physically self-consistent system. This includes targeting the cell nuclei automatically. In this respect, and, in terms of a clinical application, the ideal instrument for this purpose is a microscope based system where the user ‘points and clicks’ on an area of interest, whereby a spectrum or a number of spectra from individual cells/tissue of the type illustrated in Figure 1 are automatically recorded. These spectra can then be compared to a large library of pre-recorded spectra from a wide sample base including all grades of cervical intraepithelial neoplasia (CIN I, II and III). An algorithm is then used to classify the spectra into the most appropriate group and a clinically significant classification returned via a simple user interface. This includes the introduction of a learning algorithm based on the experience and expertise of a pathologist/cytologist which can be employed to provide a greater degree of accuracy, abnormal cells being identified using image processing algorithms which are subsequently targeted for Raman spectral analysis.

There are a variety of ways in which the spectral signatures illustrated in Figure 1 can be classified providing consistency is maintained from sample to sample. For example, Figure 2 shows a Principal Component Analysis [33] for Raman spectra that, in this example, provides point clusters associated with the coordinates of the first three Principal Components. This analysis provides the basis for implementing the approach adopted in [34], for example, which applies ‘Fuzzy Logic’ to categorise parameters that include those associated with signals derived from the boundaries of ‘fuzzy objects’ in medical images.

Fig. 2. Results of a Principal Component Analysis for the Raman spectra given in Figure 1 showing a three-dimensional principal component scatter plot (left) and a two-dimensional principal component scatter plot (right), [32].

The rationale for the research reported in this paper is to develop a second generation system which eliminates the need for the pathologist/cytologist to ‘point and click’. The system uses image processing algorithms to identify ‘suspicious cells’ in a cervical smear sample. One approach is to exploit the well known fact that the cell nucleus increases in size in abnormal cells because of increased DNA content. However, the same effect can be used to identify suspicious cells by detecting epithelial cells in the mixed cell population of a cervical smear. As illustrated in Figure 3 [35], a number of different cell types, including red and white blood cells, can be present as well as the squamous epithelial cells. The DNA content increases in abnormal cells resulting in larger and denser nuclei and...
an increased nucleus-to-cytoplasm ratio. Using this effect, the epithelial cells can, in principle, be detected on the basis of their larger size and an image created showing all the detected epithelial cells which is then saved with a patient identifier code. In this respect, if the nucleus-to-cytoplasm ratio and optical density of nuclei are measured in cervical smear samples then the spatial co-ordinates of the suspicious cells can be recorded. Using these co-ordinates, Raman spectra can be obtained from each suspicious cell and compared to the library and classified.

Although a feasible solution for stained samples (where there is usually clear colour contrast between the cell nuclei and the cytoplasm), for application to Raman spectroscopy, this approach cannot be taken due to the effects that staining has on the Raman spectrum. Even if the approach can be adapted from non-stained samples, the processing time required to obtain the coordinates of a cell with a nucleus to cytoplasm ratio that is ‘larger than normal’ can be excessive, requiring specialist image processing hardware.

**Fig. 4. Example of a standard cytological slide.**

**V. TECHNOLOGY OVERVIEW**

In this section, we briefly review the principal components associated with the application. Cytological cells are prepared using the ThinPrep technology and ‘fixed’ on a slide of the type shown in Figure 4. The cell sample is fixed as a monolayer within the red circle shown in Figure 4. Each slide has a unique Identification Number and a OCR (Optical Character Recognition) system controls the ‘order’ of the slides. The slides are stored in a cartridge and loaded into an optical microscope. For the work reported in this paper an Olympus BX51 microscope is used together with a Prior motorised stage which selects and scans the slides. The principal purpose of the image recognition system developed to date is to return the relative coordinates that define a target (suspect) cell, in particular, the nucleus of that cell. For this purpose each slide needs to be calibrated. This begins by searching for three features allocated to each slide which form part of the standard ThinPrep slide feature identified by the three blue rings shown in Figure 4.

Image acquisition depends on the technology that is best suited for integration with the particular application. For pattern recognition in histopathology, for example, high fidelity digital images are required for image analysis whose resolution is, at least, compatible with the image acquisition equipment used for human inspection, e.g. an optical microscope. The microscope used for analysing a cervical smear slide has to be equipped with a ‘C-Mount adapter’ and digital camera. The images used in the current application as discussed in this paper are, in general, relatively noise free and are digitised using a standard CCD camera. Nevertheless, it is important that high fidelity images are obtained which are homogeneous with regard to brightness and contrast through application of an optical diffuser, for example. Unless consistently high quality images can be generated that are compatible with the sample images used to design a given computer vision system, then that same system can be severely compromised.

The cells obtained via a cervical smear do not cover the whole surface area within the circle given in Figure 4. A low magnification lens is therefore used to segment regions of interest which include cell clusters where there is a high population density of cells that can be be inspected. The focus is on the identification of the edge features which is an important component of cell recognition, in general. This identification provides information on the basic topology of a feature from which an interpretative match can be achieved. Some edges can be detected only in terms of a representative view of a whole image and have no connection with local pixels. Nevertheless, the segmentation of an image into a complex of edges is a useful pre-requisite for object identification and the solution requires an analysis of the whole scene.

Although many low-level processing methods can be applied for this purpose, the problem is to decide which object boundary each pixel in an image falls within and which high-level constraints are necessary. In many cases, a principal question is which comes first, recognition or segmentation?

**VI. IMAGING MODEL**

Suppose we have an image which is given by a function \( f(x, y) \) and contains some object described by a set (a feature vector that may be composed of integer, floating point and strings) \( S = \{ s_1, s_2, ..., s_n \} \). We consider the case when it is necessary to define a sample which is somewhat ‘close’ to this object. This task can be reduced to the construction of some function determining the degree of proximity of the object to a sample - a template of the object. Recognition is the process of comparing individual features against some pre-established template subject to a set of conditions and tolerances. The process of recognition commonly takes place in four definable stages: (i) image acquisition and filtering (as required for the removal of noise, for example); (ii) object location (which may include edge detection); (iii) measurement of object parameters; (iv) object class estimation.

For the current application, cell location is undertaken via the computation of a set of weight coefficients \( k_{x,y} \) that, for
each pixel are defined in terms of the equation

\[ f_{m,n} = f(x,y)k_{x,y} \]

where

\[ k_{x,y} = \left[ \begin{array}{ccc} k_{x-1,y+1} & k_{x,y+1} & k_{x+1,y+1} \\ k_{x-1,y} & p_{xy} & k_{x+1,y} \\ k_{x-1,y-1} & k_{x,y-1} & k_{x+1,y-1} \end{array} \right] \otimes p_{obj}(x,y) \]

and \( \otimes \) denotes the convolution integral (over both \( x \) and \( y \)), the matrix values being user defined. This result yields local dependency between the current pixel \( f_{m,n} \) and the object pixels, global evaluation being determined by \( p_{obj}(x,y) \) which is the probability that the pixel could be a part of an object. This probability is calculated from a Fuzzy Logic Membership Function which has a feed-back to the current object location. The function \( p_{obj}(x,y) \) is a two dimensional matrix and recalculates local values dynamically using the object table location \( f_{m,n} \). The construction of this matrix is based on the following procedures:

1. The intensity level of the object(s) is computed. This level uses only those pixels which have not been recognised as a part of the object. The object level, denoted by \( L_{obj} \), is initially set to be lower than the background level \( L_{bgr} \), and, as the recognition process continues, so long as \( L_{obj} = L_{bgr} \), all objects are recognised as having been indexed according to the equation [47] and [48]:

\[ L_{bgr} = \text{mean}[f(x,y) - f(m,n)] \]

2. In order to obtain \( L_{obj} \), a probabilistic min-max equation (which has been experimentally tested for different images) is used given by [47]:

\[ L_{obj} = \begin{cases} L_x, & L_x \leq L_y; \\ L_y, & \text{otherwise}. \end{cases} \]

where

\[ L_x = \frac{1}{2} \left( \max_y \left( \min_x f(x,y) \right) - \min_x f(x,y) \right) \]

\[ + \min_x f(x,y) \]

and

\[ L_y = \frac{1}{2} \left( \max_x \left( \min_y f(x,y) \right) - \min_y f(x,y) \right) \]

\[ + \min_y f(x,y) \]

In order to maintain simplicity, we do not include in this equation that component which is responsible for dividing those previously defined objects in \( f_{m,n} \). For more complex images, a filter can be used to restrict a region of interest \( a \ priori \) depending on the light conditions and point of ‘evaporation’. The approach considered is generic in that it can, in principle, be applied to any type of imaging modality. The system developed for the application of cell location includes features that are based on the textural properties of an image which is an important theme in the field of pattern analysis for biophotonics.

In the current application, it is the cell nuclei that need to be identified and are the principle regions of interest. This is because in the application of Raman spectroscopy the cell nucleus is the principal target area [10] and [11]. During the process of cell nuclei recognition, all cells are indexed from high to low probability of abnormality. Raman spectrometers are then generated for those targets with the highest probability of abnormality thereby confirming or otherwise the condition of the cell [6]. The method used for targeting the cell nuclei is discussed in the following section.

VII. TARGETING CELL NUCLEI

The approach reported in this paper is predicated on Liquid Based Cytology (LBC). A clinician takes a sample in the same way as in a PAP test, but using a very small brush instead of a spatula. The head of the brush is broken off and immersed in a small vessel of liquid instead of smearing the sample directly onto a slide. This approach is better at preserving the cells and so the results of the test are generally more reliable. At present, about one in twelve PAP smears have to be repeated because the results are inconclusive due to poor readability. With LBC, far fewer tests have to be repeated. Although re-training programmes are needed to accompany the introduction of this technique, the approach reported in this paper is nevertheless based on the assumption that digital images are acquired from LBC slides.

A. Cell Categories

There are two principal types of cervical cancer: Squamous cell cancer and Adenocarcinoma. They are named after the type of cell that becomes cancerous. Squamous cells are the flat skin-like cells that cover the surface of the cervix. They are the most common type of cervical cancer. Adenocarcinoma cells are glandular cells which produce mucus. The cervix has these glandular cells along the inside of the passageway that runs from the cervix to the womb (the endocervical canal). Adenocarcinoma is the cancer of these cell types. It is less common than squamous cell cancer, but has become more commonly recognised in recent years. Only about one in five to one in ten cases of cervical cancer are adenocarcinoma and are associated with a similar precancerous phase. It is treated in the same way as squamous cell cancer of the cervix.

B. Depth of Focus Analysis

In a simplistic three-dimensional sense, most cells consist of some basic generic features which can be classified. The most important of these is that: (i) the Cytoplasm has a relatively flat textural surface whereas the Nucleus is not flat but has relatively significant depth; (ii) the border between the Nucleus and the Cytoplasm has a distribution of textures in depth, i.e. a border pattern with changes in its textural properties as a function of depth. Images of this ‘depth dependence’ can be acquired by considering different focal depths. An example of this is given in Figures 5, 6 and 7 which show three images
C. Nuclei Detection using Three-Dimensional Cell Structure

There are two imaging methods that can be considered based on: (i) using a two-dimensional image generated in a single focal plane - the mid-focal-plane, for example; (ii) considering a three-dimensional image obtained by changing the focal plane to yield a set of images at different focal depths. While the first approach is the most conventional it does not necessarily provide the most accurate solution to the current problem as discussed in [42] and [43]. This is because a mid-focal-plane image may not capture the border between the nucleus and the cytoplasm in a way that is unambiguous due to textural three-dimensional characteristics of the nuclear-cytoplasm interface.

The second approach considers the three-dimensional cell structure which, in general, may not be of clinical significance at least in a conventional sense. However, by using this approach to differentiate between the three dimensional nature of the cell nucleus relative to the flatness of the Cytoplasm we show that it is possible to extract the nucleus, detect a region of interest and thereby apply a Fuzzy Logic based Membership Function.

Suppose we have a three-dimensional array composed of a set of images (of the type given in Figures 5, 6 and 7) which is denoted by a function \( f(x, y, z) \) and contains some object described by a set of features \( S = \{ s_1, s_2, ..., s_n \} \). We consider the case when it is necessary to define a sample which is similar to this object in terms of a matching set. A conventional method consists of calculating some function of a point-wise coincidence between the map of the object and the image together with a search for the maximum value of this function. In terms of a ‘similarity function’, this method can be represented in terms of metrics that include the sum of square deviations, the sum of the modulus of deviations or as a pair of sum of multiplications of brightness values (function of the greatest transparency), for example.

The first two similarity functions compute the ‘smallness’ of a functional pair (instead of searching for a maximum a search is launched to obtain the minimum). However, in this application, not all fragments of an nucleus edge are equally important and hence, a broadly distributed functional evaluation matched with weighted coefficients is undertaken. The selection of weight coefficients is calculated from a given set of samples with two fuzzy logic sets for the nucleus and for the cell.

Normally, fuzzy logic systems for image analysis provide a decision using a knowledge database by subscribing different edges. In this application the nucleus edge is distributed only in depth and the fuzzy set is defined such that it is not necessary to use a positive feedback learning procedure for second stage object modelling [44].

The computation of a particular value of the Membership Function \( p_{obj}(x,y,z) \) is obtained according to the equation

\[
p_{obj}(x,y,z) = \int_{xy}^{xyz} (f_{x,y}L_{obj} - L_{bgr} + \text{edge}_{xyz}) \, dx \, dy \, dz
\]

for a closed border of the object, a schematic diagram being given in Figure 8. The function \( \text{edge}_{xyz} \) is an edge detection function. In Figure 8, and, by way of an example only, the
maximum value of $p_{obj}(x,y,z)$ corresponds to the top surface of the nucleus. However, this maximum value may change for different cells because not all cells are fixed in a single layer. Hence, it is necessary to undertake a search for this local maximum. For a three dimensional polygon, the local maximum can occur within the Cytoplasm and thus, the object segmentation function is limited to the locality of a particular region of interest. Irrespective of the area allocation, the algorithm is applied recursively until nuclei fail to be detected. The shapes of these nuclei (as fixed on a slide) are not continuous and so it is not possible to develop a model based on a deterministic logic. However, Fuzzy Logic is well suited for this application.

**D. Membership Function Allocation**

We use the approach discussed in the previous section to allocate a Membership Function which provides a fast and reliable solution based on changing the focal depth over a number of consecutive steps $n$ generated by a stepping motor. We consider the function

$$V(L_{bgr}) = \frac{2\pi(R_{n+1} - R_n)}{Z_{n+1} - Z_n}$$

where $R_n$ is the radius at step $n$ and $R_{n+1}$ is radius at step $n+1$ computed by taking the average length from the Centre-of-Gravity to the edge of the nucleus segmented using the edge detection function. In correspondence, $Z_n$ is the depth of focus for step $n$ and $Z_{n+1}$ is equivalent at the next consecutive step. By computing

$$dZ = \frac{k}{V(L_{bgr})}$$

where $k$ is a correction coefficient specific to the stepping motor (obtained by calibration of the microscope) the point when $dZ < \text{Threshold}$ is taken to represent the upper bound of the nucleus. The value of the threshold and the step lengths need to be established experimentally so that they are within the bounds on the extent of the nucleus. The size and shape of the upper surface of the nucleus is mapped on to a reference image for morphological processing.

Two special filters are used for this purpose, namely, the ‘Detour by Object Contour’ and the ‘Convex Hull Spider’ algorithms presented in [48]. These filters are used for generating a uniformly closed boundary to define the edge of the cell nucleus - the ‘edge detection function’. The ‘Centre-of-Gravity’ of the closed boundary is taken to be the centre of nucleus and its coordinates used to target the point at which a Raman spectrum is generated. This is illustrated in Figure 9 which shows a screen shot from the GUI used for targeting a cell nucleus based on the approach discussed in this paper. The computational time associated with the system developed to date depends on the image resolution and normally varies from between three to six seconds for each cell using a MATLAB-Linux environment. For real time recognition, the approach considered can be implemented by means of a FPGA (Field Programmable Gate Array) to yield a computational performance that is compatible to the video frame rate associated with the application.

**VIII. Conclusion**

The work reported in this paper is part of a wider investigation into automating the application of Raman Spectroscopy in Cytology with a focus on the detection of abnormal cervical cells using LBC based on [6]. We have been concerned with the task of developing a methodology concerned with two key tasks: (i) the partial analysis of an image in terms of the textural properties that characterise that structure - the cell nucleus; (ii) the use of a fuzzy logic engine to classify the nucleus edge. The specific contribution to pattern recognition considered has been addressed and involves using the texture contrasts that occur in depth between the nucleus of a cell and the Cytoplasm. By exploiting this property it is possible to isolate and thereby target the nucleus of a cell in order to generate a Raman Spectrum. This approach has been developed as part of a scheme to automate the process of Raman spectral diagnosis in Cytology and forms the basis of a British Patent [49] filed in October, 2012.
ACKNOWLEDGEMENT

The work reported in this paper was supported by an Enterprise Ireland Commercialisation Fund. Jonathan Blackledge is supported by the Science Foundation Ireland Stakes Professorship Programme. The authors are grateful to Dr James Condron of the TeaPOT Research Group http://teapot.dit.ie/, Professor Hugh Byrne, Head of the Focal Institute http://www.dit.ie/focas/ and Mr Tom Flanagan, Head of Commercialisation, Hothouse http://dit.ie/hothouse/ at the Dublin Institute of Technology. The paper includes references to both hardware and algorithmic development. The hardware implementation was supported by Enterprise Ireland including the microscopy, the motorised stage control interface and the associated software engineering including the GUI. The cell recognition algorithms were developed by Dr Dubovitskiy prior to him commencing engineering including the GUI. The cell recognition algorithms supported by Enterprise Ireland including the microscopy, the algorithmic development. The hardware implementation was supported by Enterprise Ireland including the microscopy, the motorised stage control interface and the associated software engineering including the GUI. The cell recognition algorithms were developed by Dr Dubovitskiy prior to him commencing engineering including the GUI.

REFERENCES

Jonathan Blackledge holds a PhD in Theoretical Physics from London University and a PhD in Mathematical Information Technology from the University of Jyvaskyla. He has published over 200 scientific and engineering research papers including 14 books, has filed 15 patents and has been supervisor to over 200 MSc/MPhil and PhD research graduates. He is the Science Foundation Ireland Stokes Professor at Dublin Institute of Technology where he is also an Honorary Professor and Distinguished Professor at Warsaw University of Technology. He holds Fellowships with leading Institutes and Societies in the UK and Ireland including the Institute of Physics, the Institute of Mathematics and its Applications, the Institution of Engineering and Technology, the British Computer Society, the Royal Statistical Society and Engineers Ireland.

Dmitri Dubovitskiy holds a Diploma and B.Sc. in Aeronautical Engineering from Saratov Aviation Technical College (Russia), an M.Sc. in Computer Science and Information Technology from the Bauman Moscow State Technical University and a Ph.D. in Computer Science from De Montfort University (UK). He is a Director and Principal Engineer at UK-based Oxford Recognition Limited which he co-founded in 2005 to undertake research and development, specializing in the application of artificial intelligence for computer vision. He is currently working at DIT in the Focas Institute on an Enterprise Ireland funded research contract developing cell detection algorithms combined with Raman spectroscopy for cervical cancer screening.

Fiona Lyng holds a BSc (Physics and Chemistry, 1st Class honours) from Trinity College Dublin and Dublin Institute of Technology awarded in 1991 and a PhD (Radiation Biology and Electron Microscopy) obtained in 1995 from University College Dublin. This was followed by a Postdoctoral Fellowship (awarded by the Ernst Schering Research Foundation) at the Synchrotron Radiation Department, Daresbury Laboratory, UK, from 1995 to 1998, studying cell signalling pathways by unique microscopical techniques utilising synchrotron radiation. She then worked at the Cellular Physiology Research Unit in University College Cork from 1998 to 1999 continuing her research from Daresbury Laboratory and commissioning and training users on a Bio Rad MRC 1024 confocal microscope. She was awarded a Postdoctoral Fellowship from Dublin Institute of Technology in 1999 and an Arnold Graves Fellowship in 2000. She has been the Radiation and Environmental Sciences Centre Manager since 2003 and was awarded an Honorary Professorship of Dublin Institute of Technology in 2012.
Digital Image Encryption based on Chaotic Behavior of A Modified Tent Map

Santhosh Kumar Kadari  
Institute of Smart-System Technologies  
University of Klagenfurt, Austria  
skadari@edu.uni-klu.ac.at

Bala Soujanya Bhupathi Raju  
Institute of Smart-System Technologies  
University of Klagenfurt, Austria  
bbhupath@edu.uni-klu.ac.at

Nguyen Xuan Quyen  
School of Electronics and Telecommunications  
Hanoi University of Science and Technology, Vietnam  
quyenmx-fet@mail.hut.edu.vn

Abstract—This paper investigates an image encryption algorithm which is based on the chaotic behavior of a modified tent map where a secret key of 100-bit is used in order to produce two values of initial and control parameters which are input to the map. The number of chaotic elements generated by the map is equal to the number of pixels. The order of these elements is arranged by means of sorting their chaotic values and the position of pixels is scrambled in the encrypted image based on this order. In order to verify the performance of the algorithm, an experimental simulation is implemented using MATLAB and its results are reported.

Keywords-tent map, secret key, image encryption

I. INTRODUCTION

COMMUNICATION through image transmission and reception has been rapidly increasing in recent years. The image security is an important factor in order to achieve successful communication and it is required in many applications such as multimedia communication, video conferencing, cable TV, military image databases, medical image analysis, etc [2, 3]. Image encryption is a robust way to guarantee the privacy and security of communication [3, 4].

There are many traditional encryption algorithms have been developed such as Data Encryption Standard (DES), International Data Encryption Algorithm (IDEA), Advanced Encryption Standard (AES), etc [2-6]. These algorithms are not suitable for real time image encryption due to their weakness of large computational time and high computing power [3]. To avoid this drawback, several researches have been carried out to develop image encryption algorithms based on chaotic engineering [2-10].

In this paper, we introduced an image encryption algorithm which is based on the chaotic behavior of a modified tent map. The advantage of using this map is to generate chaotic values in the wide range. In the algorithm, a secret key of 25 hexadecimal characters is used and the size of each character is 4-bit length. The contribution of secret key is utilized to get control and initial parameters which determine chaotic elements generated at the output of the map. The number of these elements is equal to the number of pixels of the image. The order of chaotic elements is sorted out through which the position of pixels is scrambled in the image encryption. It is not easy to recover the original image if an unauthorized decoder has no information on the secret key, the equation of the map and the rule of the algorithm.

This paper is organized as follows. Section II explains the chaotic modified tent map. Section III elaborates the design of the image encryption algorithm. MATLAB simulation results are presented and described in terms of the image histogram in Section IV. Correlation Coefficient and PSNR analysis are discussed in Section V and Section VI. Finally, our conclusions are given in Section VII.

II. MODIFIED TENT MAP

A chaotic modified tent map has been investigated in detail in [1]. In this section we just present its brief description. The map is iteratively generated through a transformation function $F(X): (-1, 1) \rightarrow (-1, 1)$ as given by:

$$
X_{n+1} = F(X_n) = \begin{cases} 
\mu X_n & \text{if } \frac{1}{|\mu|} < X_n < \frac{1}{|\mu|} \\
-\mu(X_n - \frac{2}{|\mu|}) & \text{if } -1 < X_n \leq -\frac{1}{|\mu|} \\
\mu(X_n - \frac{2}{|\mu|}) & \text{if } 1 \leq X_n < \frac{1}{|\mu|} 
\end{cases}
$$

(1)

where $n$ represents the time step, $X_n$ is the output value at the $n$th step and parameter $\mu$ controls the state of the map. The sign of the output value alternates in each iteration, i.e, $X_{n+1} X_n < 0$. 
In other words, for an equal initial state, the absolute value of the output sequence is equal for both the tent map and the modified tent map while the output of the modified tent map alternates between positive and negative values. Since the output values of the tent map and the modified tent map have equal absolute values, the modified tent map is interchangeable with the tent map for generating chaotic values.

The bifurcation diagram of the modified tent map is shown in Fig. 1 which shows that the dependence of the chaotic behavior of the map on the control parameter $\mu$.

![Fig. 1. Bifurcation diagram of the modified tent map](image)

With $|\mu| > 3$, the map could not maintain the state in the input range, which could not be used for chaotic applications. With $-1 < \mu < 1$, the map does not demonstrate chaotic behavior. Regardless of the initial state of the map, the output will ultimately settle in zero. Chaotic behavior can also be observed with $1 < \mu \leq 2$. With $-2 \leq \mu < -1$, the map also shows chaotic behavior. In this mode, the output alternates between positive and negative values with each iterative step. In this region, the system does not have an asymptotic density distribution. With $2 < |\mu| < 3$, the system is chaotic and the bifurcation diagram represents the steady state density distributions.

### III. IMAGE ENCRYPTION ALGORITHM

The image encryption algorithm consists of several steps that are elaborated as follows:

Step 1: Consider an image of size $M \times N \times C$ where $M$, $N$ denote respectively rows and columns of intensity level and $C$ denotes color combination. The pixel values of the image range from 0 to 255 and the total number of pixels are multiplication of $M$, $N$.

Step 2: Separate the original image into $R$, $G$, $B$ matrices and convert each matrix into single array vector of size $1 \times MN$.

Step 3: The algorithm uses a secret key $K$ consisting of 25 hexadecimal characters (0–9 and A–F).

$$K = k_1, k_2, k_3, ..., k_{25} \quad \text{(Hexadecimal format)}$$

Step 4: Each character in the secret key is converted into its binary equivalent form of four bits and the entire secret key is of 100-bit long.

$$K = k_1, k_2, k_3, ..., k_{25} \quad \text{(Binary format)}$$

Step 5: In the proposed algorithm, modified tent map is used to achieve the goal of image encryption. The map is presented in equation (1).

Step 6: Control parameter $\mu$ and initial parameter $X(1)$ are deduced from the secret key.

Step 7: The initial parameter $X(1)$ is calculated by using the following equation.

$$X(1) = (X_{01} + X_{02} + X_{03}) \mod 1.$$  \hspace{1cm} (5)

The value of initial parameter is in between -1 and 1 and it is given to the chaotic modified tent map. $X_{01}$, $X_{02}$, and $X_{03}$ are computed using the following equations.

$$X_{01} = \left(\frac{k_{131} \times 2^0 + k_{132} \times 2^1 + ... + k_{134} \times 2^3}{k_{181} \times 2^0 + k_{182} \times 2^1 + ... + k_{184} \times 2^3}\right)2^{24}$$ \hspace{1cm} (6)

$$X_{02} = \left(\frac{k_{231} \times 2^0 + k_{232} \times 2^1 + ... + k_{234} \times 2^3}{k_{121} \times 2^0 + k_{122} \times 2^1 + ... + k_{124} \times 2^3}\right)2^{24}$$ \hspace{1cm} (7)

$$X_{03} = \left(\frac{k_{111} \times 2^0 + k_{112} \times 2^1 + ... + k_{114} \times 2^3}{k_{61} \times 2^0 + k_{62} \times 2^1 + ... + k_{64} \times 2^3}\right)2^{24}$$ \hspace{1cm} (8)

Step 8: The control parameter $\mu$ is calculated by using the following equation.

$$\mu(1) = (\mu_{01} + \mu_{02} + \mu_{03}) \mod 3.5.$$ \hspace{1cm} (9)

The value of control parameter is in between 2 and 3 and it is given to the chaotic modified tent map. $\mu_{01}$, $\mu_{02}$, and $\mu_{03}$ are computed using following equations.

$$\mu_{01} = \left(\frac{k_{131} \times 2^0 + k_{132} \times 2^1 + ... + k_{134} \times 2^3}{k_{181} \times 2^0 + k_{182} \times 2^1 + ... + k_{184} \times 2^3}\right)2^{24}$$ \hspace{1cm} (10)
\begin{align}
\mu_{02} &= \left( k_{21} \times 2^0 + k_{72} \times 2^1 + ... + k_{74} \times 2^3 \right) / 2^{24} \quad (11) \\
\mu_{03} &= \left( k_{61} \times 2^0 + k_{62} \times 2^1 + ... + k_{64} \times 2^3 \right) / 2^{24} \quad (12)
\end{align}

The equation (13) generates control parameter value in between -2 and -3.

\[ \mu(1) = (\mu_{01} + \mu_{02} + \mu_{03}) \mod 4.5. \quad (13) \]

In this paper, we have considered the positive control parameter value.

Step9: Based on the initial and control parameters, the map generate chaotic elements with size of \( M \times N \times C \).

Step10: Divide the chaotic elements into three parts.

Step11: Change the order of chaotic elements by using ascending order.

Step12: Scramble the order of \( R, G, B \) vectors based on the order of chaotic elements. Each vector follows the order of each part of the chaotic elements.

Step13: Convert the size of \( R, G, B \) matrices \( 1 \times MN \) into \( M \times N \) and combined them to get the encryption image.

The encrypted image is input image in the decryption process and the remaining all steps are similar to the encryption process. The original image is exactly recovered when the secret key, initial, and control parameters are correctly used.

IV. EXPERIMENTAL RESULTS

In this section, we take two different color images for testing and simulation is done using MATLAB software to evaluate the efficiency of the proposed image encryption method. Experimental results include Image Encryption Analysis and Image Histogram Analysis.

A. Image Encryption Analysis

The encryption results of two images such as Lena and Micro controller are showed in Fig.2 and Fig.3. For encryption and decryption process, we have used the same secret key as \( K = 1A2BC7DEF093A4568984512B37 \), the initial and control parameters are generated from the secret key as \( X(1) = 0.2747 \), \( \mu = 2.7747 \).
B. Image Histogram Analysis

The corresponding histograms of Lena and Micro controller images are showed in Fig. 4 and Fig. 5. The image histogram represents pixel intensity versus pixel distribution.

(a) Original Image Histogram
(b) Encrypted Image Histogram
(c) Decrypted Image Histogram

Fig. 4. Lena Image histogram analysis

(b) Encrypted Image
(c) Decrypted Image

Fig. 3. Micro controller Image encryption analysis

(a) Original Image
(b) Encrypted Image
(c) Decrypted Image
V. CORRELATION COEFFICIENT ANALYSIS

In this section, we have provided the correlation between original and encrypted images. The correlation analysis is used to analyze the similarity between two adjacent pixels. For a normal image, pixels are usually highly correlated with its adjacent pixels in any direction (horizontal, vertical or diagonal). These high-correlation properties can be quantified as the correlation coefficient for comparison. The following equations are used to calculate the correlation coefficient.

\[ E(x) = \frac{1}{N} \sum_{i=1}^{N} x_i \]  \hspace{1cm} (14)

\[ E(y) = \frac{1}{N} \sum_{i=1}^{N} y_i \]  \hspace{1cm} (15)

\[ D(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - E(x))^2 \]  \hspace{1cm} (16)

\[ D(y) = \frac{1}{N} \sum_{i=1}^{N} (y_i - E(y))^2 \]  \hspace{1cm} (17)

\[ \operatorname{cov}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - E(x))(y_i - E(y)) \]  \hspace{1cm} (18)

\[ \gamma_{xy} = \frac{\operatorname{cov}(x, y)}{\sqrt{D(x)D(y)}} \]  \hspace{1cm} (19)

The Table 1 gives the correlation coefficient values for the Red, Green, and Blue components of the encrypted images with respect to their original images. The correlation analysis proves that the proposed encryption technique is good because all the values are tending towards zero correlation. The correlation coefficient near zero indicates that the correlation among the pixels is very little. However, the pixels in the original image were highly correlated.

VI. PSNR ANALYSIS

This section discusses the Peak Signal-to-Noise Ratio (PSNR) which is used to measure the quality between the original and an encrypted image. The higher the PSNR value, the better the quality of the recovered image. The PSNR value is evaluated based on the (minimum square error) MSE. The MSE represents the cumulative squared error between the encrypted and the original image, whereas PSNR represents a measure of the peak error. The lower the value of MSE, the lower the error is. The following equations are used to calculate the MSE and PSNR.

\[ \text{MSE} = \frac{D}{M \times N} \]  \hspace{1cm} (20)

\[ D = \sum_{m,n} \left[ I_o(m,n) - I_e(m,n) \right]^2 \]  \hspace{1cm} (21)

\[ \text{PSNR} = 10 \log_{10} \left( \frac{255}{\sqrt{\text{MSE}}} \right) \]  \hspace{1cm} (20)

where \( I_o \) and \( I_e \) are the original and encrypted images. \( M \) and \( N \) are the number of rows and columns in the images.

<table>
<thead>
<tr>
<th>Image</th>
<th>R</th>
<th>G</th>
<th>B</th>
<th>Average Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>0.0008</td>
<td>0.0023</td>
<td>0.0053</td>
<td>0.0028</td>
</tr>
<tr>
<td>Micro-controller</td>
<td>0.0009</td>
<td>0.0006</td>
<td>0.0003</td>
<td>0.0006</td>
</tr>
</tbody>
</table>
The following Table 2 shows the PSNR values for the Red, Green, and Blue components of the encrypted images with respect to their original images. From the Table 2, it is difficult to retrieve the original image from the encrypted image because the resulted PSNR values are very low (<10dB).

Table 2: Peak Signal-to-Noise Ratio for different images

<table>
<thead>
<tr>
<th>Image</th>
<th>R (dB)</th>
<th>G (dB)</th>
<th>B (dB)</th>
<th>Average Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>8.8212</td>
<td>8.9261</td>
<td>9.3258</td>
<td>9.0243</td>
</tr>
<tr>
<td>Microcontroller</td>
<td>6.1267</td>
<td>7.7546</td>
<td>7.863</td>
<td>7.2458</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

This paper has presented an image encryption algorithm based on the chaotic modified tent map which aims to improve the privacy and security of image communication. The proposed algorithm of image encryption is described using both theoretical analyses and experimental simulation. The simulation results show that the algorithm is more secure due to wide range of chaotic values generated by the map and the input elements to the map, i.e. control and initial parameters, are generated from the secret key. The correlation and PSNR analysis has been proved that the proposed image encryption algorithm is more secure against different attacks.

REFERENCES


Santhosh Kumar Kadari received the Bachelor of Engineering from Jawaharlal Nehru Technological University, Mancherial, Andhra Pradesh, India and Master degree from Alpen Adria University Klagenfurt, Klagenfurt, Austria, in 2007 and 2010,respectively. Currently, he is working toward PhD degree at Institute for Smart System Technologies,Alpen-Adria University Klagenfurt in Klagenfurt,Austria. His main research interests are image processing, nonlinear dynamics and Chaos-based digital communications.

Bala Soujanya Bhupathi Raju received the Bachelor of Engineering from Anna university, Chennai, India in 2007. She is currently pursing Master degree at Institute for Smart System Technologies, Alpen Adria University Klagenfurt, Klagenfurt, Austria. Her main research interests are image processing, Embedded Systems, Nonlinear dynamics and Digital communications.

Nguyen Xuan Quyen received the Diploma of Engineer and Master degree from the School of Electronics and Telecommunications (SET), Hanoi University of Science and Technology(HUST) in 2006 and 2008, respectively. From 2007 to 2010, he was a lecturer at the Department of Telecommunication Systems of SET-HUST. Currently, he is working toward Ph.D degree at Institute for Smart System Technologies, Alpen-Adria University Klagenfurt in Klagenfurt,Austria. His main research interests are microwave circuits, ultra wideband communications and chaos-based digital communications.
Least Squares Bandwidth Allocation Scheme for 4G Heterogeneous Networks

Jarmo Siltanen, Jyrki Joutsensalo, Timo Hämäläinen, and Kari Luostarinen
1 JAMK University of Applied Sciences
2 University of Jyväskylä
3 Metso Paper

ABSTRACT

4G multiservice networks will carry different kinds of applications in the near future. Bandwidth requirements changes rapidly and due that the network resource management will play an important role to guarantee the use of the limited resources the most efficient way. We have approached the channel capacity allocation problem by developing a QoS-aware channel allocation method. In our model, the channel may be wired or wireless, so this method can be adapted in multi-technique networks. The algorithm allocates resources between two network nodes via several different capacity routes. Maximal capacity of the data network infrastructure is tried to exploited by dividing user’s needs to the different CoS classes. The performance of the algorithm is justified from the least squares and maximum likelihood point of view.

I. INTRODUCTION

Demand for Internet access and interactive services will describe the use of mobile devices in the future. This results into a situation, where a large amount of traffic flows to the mobile terminal and the overall traffic will be quite asymmetric. We will present a model how to select users traffic flows, so that the total link capacity will be maximized. Our allocation scheme is suitable to different kind of situations in which a user population uses dynamically a same limited resource. Our primary focus is the context of networked services, e.g., provided through the different kind of network links (wired and wireless).

Bandwidth management is important issue when allocating capacity among traffic flows. If we consider e.g. wireless IP networks and use differentiated service architecture [1], then we need intelligent properties to data link layer to allocate channel capacity in most efficient way. Data link layer bandwidth allocation model is presented e.g. in [5]. This paper presents a TDMA-based centralized MAC model to provide fair throughput among bursty data flows in a Rayleigh fading channel. Other dynamic channel allocation models is presented e.g. in [2], [3], [9].

Our work has same kind of features as above and other studies [6], [7], [10], [11], but the most significant difference is that our scheme is suitable also for wired network environment. This issue makes our model very interesting for all-IP networks. Our distributed route selection functionalities, which employs the idea of maximum likelihood method at each network node, can be implemented e.g. in the IP supporting BS or deeper into network (edge router).

The rest of the paper is organized as follows. In Section 2, we discuss the least squares maximum likelihood channel allocation. In Section 3, we study the performance of our algorithm and also present and discuss the simulation results. Section 4 concludes this work, highlighting the advantages of the proposed method.

II. LEAST SQUARE ROUTING

A. Least square model and algorithm

Here we introduce a least square channel allocation algorithm, which appears also to be optimal in the maximum likelihood sense under certain conditions. In our scenario, m users are connected from City 1 to City 2 via n channels. The channels may be optical fibers, wired or wireless channels, TDMA, CDMA, and so on. All the channels are integrated together. One user may use several channels, if it has very large utility. For example, if some user needs 1Gb/s for transmitting virtual reality images, and if one channel may offer only 100 Mb/s service, then it needs at least 10 channels for parallel use.

Let r be the vector containing different data rates (or generally utilities):

\[ r = [r(x_1), \ldots, r(x_m)]^T \text{ bits/s} \]  

(1)

Here \( x_i \) is the price paid by customer \( l \) to the network operator, and \( r(x_i) \) is the data rate offered to the customer \( l \). Thus the total data rate offered to all the users in the network at the time instant \( t_0 \) is

\[ C = \sum_{i=1}^{m} r(x_i) \text{ bits/s} \]  

(2)

Here \( C \) is the total channel capacity of the entire network. Channel capacity \( C \) is the maximal number of information that the entire network \( B \) can transfer without errors. This is the basic definition given by classical Shannon information theory [4]. In practice, suboptimal estimate of \( C \) is obtained by observing the data flow history in the network. In this paper, we do not take a stance on the strict estimation procedure of the channel capacity.

In the true application environment, the vector \( r = r(t) \) varies with respect to the time \( t \) due to the new contacts and leaves.

Let \( c \) be the capacity vector:

\[ c = [c_1, \ldots, c_n]^T \]  

(3)
The vector $c$ contains available capacities of different channels. Again, due to the non–idealities of the network infrastructure, such as fading or multipath propagation in the wireless channels, the capacity vector $c = c(t)$ depends on time. Here it is assumed that the network is in the (high load) steady state in the sense that

$$E(\sum_{j=1}^{n} c_j) = \sum_{i=1}^{n} r_j$$  \hspace{1cm} (4)

In general, the data streams between two points via several channels, and thus the data rate of the user $l$ can be modeled as a linear combination

$$r(x_l) \approx \sum_{j=1}^{m} w_{lj}c_j$$  \hspace{1cm} (5)

under the constraints

$$\sum_{i=1}^{n} w_{ij} = 1$$  \hspace{1cm} (6)

for all $j$, and

$$w_{ij} \geq 0$$  \hspace{1cm} (7)

for all $i, j$. Eq. (4) justifies the use of the constraint (6), because

$$\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij}c_j = \sum_{j=1}^{n} c_j \sum_{i=1}^{m} w_{ij} = \sum_{j=1}^{n} c_j$$  \hspace{1cm} (8)

The data rates can be modeled by the matrix form

$$r = Wc + w$$  \hspace{1cm} (9)

where $r$ is known by the network manager, $c$ is an estimate of the capacity vector, and the unknown weight matrix $W$ is tried to be optimized. Here $w$ is an error vector due to the estimation errors of $c$. Our goal is to get estimate for $W$ so that the error $w$ is minimized. The optimization problem is the following:

$$J(W) = \|r - Wc\|^2 = \min$$  \hspace{1cm} (10)

where $\|\cdot\|$ is 2-norm. Here $W$ contains the unknown weights to be optimized. If $w$ is modeled as an additive white Gaussian noise, then it is well known eg. from [8], that the maximum likelihood problem for estimating optimum $W$ reduces to the least squares equation (10). The problem reduces to the form

$$-2r^T Wc + c^T W^T Wc = \min$$  \hspace{1cm} (11)

The derivative of $J$ with respect to the matrix $W$ is

$$\frac{\partial J(W)}{\partial W} = -2(Wc - r)c^T$$  \hspace{1cm} (12)

and the gradient algorithm for optimization weights is as follows:

$$\dot{W}(t + 1) = \dot{W}(t) - \mu(t)(\dot{W}(t)c - r)c^T$$  \hspace{1cm} (13)

where $\mu(t)$ is a small positive number. It is important to emphasize that the selection of $\mu(t)$ is very critical. If $\mu(t)$ is too large, the weights may "explode", but if $\mu(t)$ is selected to be too small, the convergence may be too slow, and the computation time is wasted.

In practice, the channel allocation is implemented as follows:

1. Give some initial guess to $\dot{W}(0)$.
2. Update Eq. (13).
3. Set $w_{ij} = sgn(w_{ij})w_{ij}$ and $w_j = \frac{w_j}{\|w_j\|}$ so that the constraints (6) and (7) are fulfilled.
4. Stop when converged using some error criterion, otherwise go to the Step 2.

The gradient algorithm is adapted continuously, because the network "lives". To detect convergence, Eq. (10) can be checked, and different initial guesses to $\dot{W}(0)$ can be given, when convergence seems to fail.

B. Weighted algorithm

In practice, users that pay more must be served better. Therefore, optimization criterion may be modified. We may use e.g. the weighted criterion

$$J_2(W) = (r - Wc)^T R (r - Wc)$$  \hspace{1cm} (14)

where

$$R = \text{diag}(r_1, \ldots, r_m)$$  \hspace{1cm} (15)

gives different weights to the different data rates and utilities. When Eqs. (14) and (15) are used, the iteration rule

$$\dot{W}(t + 1) = \dot{W}(t) - \mu(t)R(\dot{W}(t)c - r)c^T$$  \hspace{1cm} (16)

can be derived.

C. Computational complexity

Let us consider the computational complexity of the algorithms. Let us consider algorithm (13). $a_1 = Wc$ needs $O(mn)$ operations. $a_2 = a_1 - r$ needs $O(m)$ operations. $A_3 = a_2c^T$ needs $O(mn)$ operations. $\dot{W} + \mu A_3$ needs $O(mn)$ operations. If the network is iterated by using totally random initial guess, it needs $i$ iterations to converge. Thus it is needed $O(mni)$ operations. Therefore the choice of $\mu(t)$ may have a dramatical effect to the computational complexity.

In the tracking mode, it is assumed that the network is in some optimal (global) or suboptimal (local) equilibrium point. Then network makes few iterations when it detects contact or leaving. In that case, the computational complexity is $O(mn)$ operations/iteration.

III. SIMULATIONS

Example 1. In this example, we have the following scenario and parameters:

- The number of users is $m = 200$.
- There are $n = 20$ channels between City 1 and City 2.
The capacities of the channels lie uniformly in the interval $(0, 1)$. This is a normalization convention without loss of generalization.

168 simulations were made. (We used some Matlab routines, and there are 168 prime numbers between 1 and 1000.) In each simulation, different random initial guess to the weight matrix $W$ was given.

Figure 1 illustrates the simulation results using the algorithm (13). The horizontal axis presents the number of iterations $(t = 1, \ldots, 500)$, while the vertical axis presents the logarithmic error

$$E_1 = 10 \log 10(||r - Wc||^2) \quad (17)$$

In the figure, there are 168 curves corresponding to different random initial guesses. At the beginning of the iteration $(t = 0)$, the error is about 0 dB, and for other initial guesses, error decreases rapidly, while for other guesses, error does not decrease at all. However, it is easy to detect the behavior of the error, and give new initial guess, if no convergence is detected.

In the next simulation, we have fixed the error target $E_1$ in Eq. (17), and estimated what is the probability that the algorithm (13) achieves fitting error which is below that target. Totally random initial guess was given. In Fig. 2, we see that even in the case $E_1 = -150$ dB, the probability of resolution is 0.71.

In the last simulation, we made similar test for the weighted algorithm (16) with weighted error criterion

$$E_2 = 10 \log 10[(r - Wc)^T R (r - Wc)] \quad (18)$$

Now probability of resolution is low for strict error targets, but rapidly increases at about $E_2 = -70, \ldots, -40$ dB. Figure 3 shows the result.

In these simulations, and in some cases, the algorithms have serious oscillation problems. However, it is not difficult to detect oscillations, and so the network manager can reinitialize the weight matrix $W$, if it detects oscillation or too large estimation errors. The detection is performed by checking directly the error value $E_1$ or $E_2$.

IV. CONCLUSIONS

In this paper, we have presented link allocation algorithms based on a least square principle. In our scenario, Router 1 has several links or channels to the Router 2. These channels can be e.g. TDMA or CDMA channels, optical fibers, ordinary wired electric channels, and so on. The least square algorithm has an iterative nature, and thus it can be realized in the environments where contacts and leaves continuously occur, i.e. the network lives. Next, we will develop our model so that the below open issues can be studied:

- Derive optimal forgetting factor $\mu$.
- Adaptive tracking, when the network parameters vary with respect to the time.
- Avoiding oscillations.
- Hierarchic implementations, by which we mean that the weights of the largest utilities are adapted first.

We will also expand the model to cover large number of network nodes and consider how the model can be implemented in real environment. Especially Service Level Agreement is considered.
REFERENCES


