Automated people counting using low-resolution infrared and visual sensor

I. J. Amin, A. Al-Habaibeh, F. Junejo, A. J. Taylor, R. M. Parkin

1. Introduction

Research in many areas are being conducted involving automated counting, biology [1], medicine, quality control, industrial machine vision processes and so on. Many automated people counting systems are developed over the years. There are contact based sensors such as pedestrian barriers on entrances into public buildings, gateways. Most commercially available non-contact based counters uses infrared beam, ultrasonic sensors. Specialized human information sensors are also developed for this task [2]. But the most commonly used system still remains was a visual camera [3-5]. The disadvantage with visual counting system is the cost. High spatial resolution visual camera and a frame grabber are required which makes the system expensive.

Even with high spatial resolution cameras the inaccuracy problem still remains for detection of people. Say if a person is wearing same shades of grey as of background it will be difficult to distinguish between the background and same shades of cloths. Also there is no such ways of distinguishing with accuracy a person from different objects. These objects in the background are one of the main concerns that raised false alarms in many automated people counting systems. Where as the background separation is not the easy task. Visual automated counting systems can only work in the presence of ambient lighting such as office environment, sunlight, or other interior types of lighting. In case of emergencies like fire, blackouts the system will malfunction during evacuation of the building thus will render useless during emergencies. Similar case is with exterior use of people counters [5], there will be false alarms during night time by the counting system if there is no special lighting arrangement in the area under consideration.

Thus a system is proposed by using a low cost infrared thermal and visual camera. The visual camera uses an image-processing algorithm can distinguish between people and objects with accuracy of about 12%. This system is developed by [4], even though a visual automated counting system this counting system can be modified easily to accommodate low spatial resolution visual images. This system is based on the background training of visual images using neural network.

2. Thermal Imaging

All objects emit heat by three means: Conduction, convection and radiation. Conduction transfers heat through solid objects. Convection transfers heat through fluids like air and water. Radiation transfers heat through electromagnetic radiation.

Objects continuously radiate heat with certain wavelength. This wavelength depends upon the temperature of the radiating object and its spectral emissivity. As the object temperature increases the radiation also increases. The radiation emitted also includes the infrared radiation emission. This infrared emission consists of electromagnetic wavelengths between 0.7 μm to 1000 μm. Small ranges of infrared emission emitted by the objects are detected by the thermal imagers, which is made visible into the image.

The concept behind the thermal imager infrared emission detection is the black body is a perfect radiator; it emits and absorbs all incident energy. The energy emission for

the blackbody is greatest possible for energy emission for that certain temperature. Radiation power emitted by blackbody as given by Plank's radiation law is:[6]

$$P(\lambda,T)=\frac{2\pi hc^2}{\lambda^2}\left\{\exp\left(\frac{hc}{\lambda bT}\right)-1\right\}^{-1}$$

p= Energy Radiated

λ= Wavelength

T= Temperature (Kelvin)

h= Plank's Constant

c= Velocity of light

b= Boltzman Constant

Real object are not a perfect emitters or absorbers. Thus emissivity (ε) of the real surface is defined as the ratio of thermal radiation emitted by a surface at given temperature to that of a blackbody for same temperature, spectral and directional conditions[7, 8]. Thus emissivity of a blackbody is 1 and all other real surfaces emissivity will be between 1 and 0.

According to Stefan Boltzman Law of emissivity radiation:

$w=\varepsilon\eta T^4$

w = Radiated energy

ε = emissivity

η = Boltzman constant $\left(5.67\times10^{-8}\frac{w}{m^2}K^4\right)$

T = Temperature $\left(\text{Kelvin}\right)$

Thermal imaging converts thermal radiation into digital signal and which is converted into visible image.

This study uses a newly developed thermal imager of type IRYSIS IRI 1001,as it offers many advantages such as low cost, wide temperature measurement range and it can be used to capture images if connected to IBM-PC via RS-232C port. The thermal imager is housed in an aluminium casing of 4 inches by 4 inches with optics, pyroelectric detector[9], chopping motor and optical modulator. It has a temperature measurement range of –20 to 90^{o}C with an accuracy of +/-0.1^{o}C [10]. Although it is a low resolution, 16 x 16 pixels thermal imager, but can be used to display images of up to 128x128 pixels using bilinear or bicubic interpolation. Interpolation process estimates values of intermediate components of continuous function in discrete samples. An interpolation technique does not add extra information into the image but can provide better thermal image for human perception. For bicubic interpolation, the output pixel value is the weighted average of the pixels in the nearest 4 x 4 neighbourhood. Mathematically, bicubic interpolation can be described as follows:

Let be a third degree polynomial. The Lagrange polynomial interpolation is given by

$$P(q)=\sum_{i=0}^{3}f_iL_i(q)$$

[11]

Where,

q=Point at which interpolation takes place

P(q)= interpolated value

f_i =Known values on the grid at points (q_i)

$L_i(q)$=Lagrange polynomial, for example

$$L_i(q) = \prod_{k \neq i, k=0}^{3} (q - q_k)/(q_i - q_k)$$

Previous research has shown that low resolution imager gives similar thermal information to the high resolution one, whereas, low cost thermal imager low cost than a typical high resolution thermal imager and much small than conventional thermal imager size. In addition to this low resolution imager is specially designed for embedded system, where data can be directly streamed through an RS232 connection to the computer for on-line monitoring and off-line analysis.

3. Neural Networks

An artificial neural network (ANN) is an information-processing paradigm inspired by the way in which the densely interconnected, parallel structure of the human brain processes information. Neural networks resemble the human brain in the following two ways:

A neural network acquires knowledge through learning.

A neural network's knowledge is stored within inter-neuron connection strengths known as synaptic weights.

Artificial neural networks are collections of mathematical models that emulate some of the observed properties of biological nervous systems and draw on the analogies of adaptive biological learning. The key element of the ANN paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements that are analogous to *neurons* and are tied together with weighted connections that are analogous to *synapses*.

The main advantage of using neural network is the full automation of the learning and classification processes, therefore, they can be implemented in fully automated monitoring systems, such as people counting to recognize and classify different patterns without human involvement, thereby, eliminating any error or lapses associated with human concentration during a repetitive task.

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. Some Neural networks are classified as feed-forward while others are recurrent (i.e., implement feedback) depending on how data is processed through the network. Another way of classifying neural networks types is by their method of learning or training, as some of the neural networks employ *supervised training* while others are referred to as *unsupervised or self-organizing* networks. The selection of supervised or unsupervised network is greatly dependent on the data to be processed for the training of the network.

During supervised learning of an ANN network, an input stimulus is applied that results in an output response. Then this response is compared with desired output i.e. the target response. If the actual response differs from the target response, the neural network generates an error signal, a popular measure of the error '*E'* for a single training pattern, is the sum of square differences i.e.

$$E = \frac{1}{2}\sum_i (t_i - y_i)^2$$

Where,

t_i = desired or target response for ith unit,

y_i = actually produced response for ith unit.

The error “E” is then used to calculate the adjustment that should be made to network’s synaptic weights so that the actual output matches the target output.

In contrast to supervised learning, in case of unsupervised learning does not require a teacher; i.e. no target output is required. It is usually found in the context of recurrent and competitive nets. In case of unsupervised learning, there is no separation of the training set into input and output pairs during the training session, the neural net receives as its input many different excitations, or input patterns, and it arbitrarily organizes the patterns into categories. When a stimulus is later applied, the neural net provides an output response indicating the class to which the stimulus belongs. If a class cannot be found for the input stimulus, a new class is generated. However, it should be noted that even though unsupervised learning does not require a teacher, it requires *guidelines* to determine how it will form groups. Grouping may be based on shape, colour, or material consistency or on some other property of the object.

3.1. Back propagation Neural Network

Backpropagation Neural Network are one of the most commonly used neural network structures, as they are simple and effective, and has been used successfully for wide variety of applications, such as speech or voice recognition, image pattern recognition, medical diagnosis, and automatic controls. It is a supervised neural network, which consists of “*n”* numbers of neurons connected together to form an input layer, hidden layers and an output layer. The input and output layers serve as nodes to buffer input and output for the model, respectively, and the hidden layer serves to provide a means for input relations to be represented in the output. Before any data has been run through the network, the weights for the nodes are randomly chosen, which makes the network very much like a newborn's brain developed but without knowledge. When presented with an input pattern, each input node takes the value of the corresponding attribute in the input pattern. These values are then “fired”', at which time each node in the hidden layer multiplies each attribute value by a weight and adds them together. If this is above the node's threshold value, it fires a value of “1”'; otherwise it fires a value of “0”. The same process is repeated in the output layer with the values from the hidden layer, and if the threshold value is exceeded, the input pattern is given the classification, once a classification has been given; it is compared to the actual i.e. desired classification, and the error is fed back (backpropagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to producing the desired output. This process is known as "training". The back propagation neural network used in this study uses a sigmoid function in the hidden layer and a linear function in the out put layer. Both functions can be expressed respectively as follows:

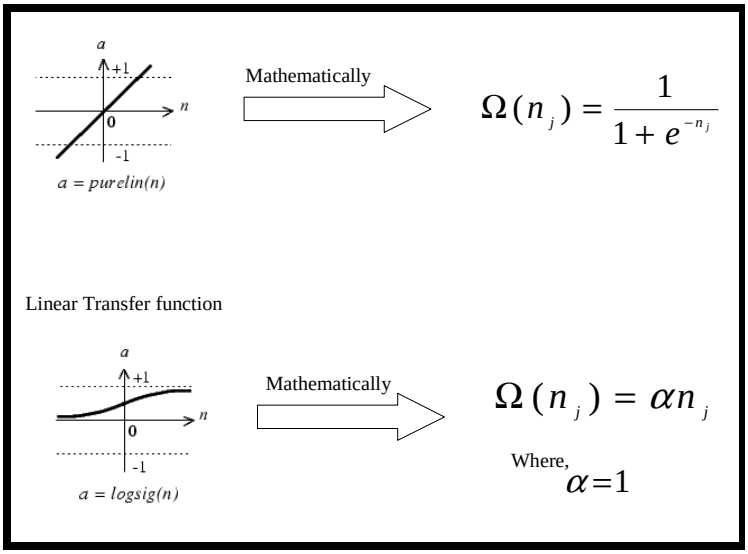

Figure 1 Mathematically expression for the transfer functions

3.2. RAM based Neural Network

Most conventional neural network training procedures, as mentioned above are used to develop the required behaviour in a learning system, having assumed that the 'weight' parameters in which the system's knowledge is stored can be positive or negative, and unboundedly large in size. These analog weights, and the algorithms by which they are adapted, are not well suited to hardware implementation. However, in this study, a sequential (RAM based) neural network has been used, which uses binary weights i.e. 0/1 values, stored in RAM memory blocks which themselves play the role of the 'neurons' in the system; this approach, sometimes called *'weightless neural computing'* has many advantages over other neural network such as fast network training, uses 'one-shot' learning procedures very different to the iterative ones of conventional neural networks as well as they can operate well on low resolution images. In addition to this, in case of RAM based neural networks, the bit-stream communication between RAM neurons, rather than being a hindrance to the system when learning, is actively beneficial in promoting *generalisation* refers to the neural network producing reasonable outputs for inputs not encountered during training (learning), whereas, other networks have to introduce such a 'blurring' of the input (so that in effect a wider range of patterns are seen during training) in a much more artificial way [12].

RAM based Neural Network Architecture:

As shown in figure given below, the basic architecture is as follows:

- the input vector is divided into parts; each part is connected to the address inputs of a 1-Bit-RAM unit.
- The output of all the RAMs within one discriminator are summed up. The number of discriminators needed in a network is determined by the number of class which need to be distinguished by the network.

The 1-Bit-RAM unit, is a device which can store one bit of information for each input address. A control input is available to switch the mode of the RAM between `Write' and `Read' for learning and recall. Initially all memory units are set to `0'. During the learn (`Write') mode the memory is set to `1' for each supplied address; in the recall

(`Read') mode the output is returned for each supplied address, either `1' (if the pattern was learned) or `0' (if the pattern was not learned).
The *discriminator* is the device which performs the generalization. It consists of several RAMs and one node which sums the outputs of the RAMs in recall mode. The discriminator is connected to the whole input vector; each RAM within the discriminator is connected to a part of this vector, so that each input bit is connected to exactly one RAM, The connections are preferably chosen by random.

4. Experiment setup

The experiment is conducted by mounting low-cost visual imaging device and the IRISYS IRI1001 thermal imager. Markers are placed on the floor under consideration so that both infrared imager and visual camera are sharing same information. Visual imager has a much wider field of view than thermal imager. Thus only a cropped visual view is taken into consideration.
Special software is developed using National Instruments LabWindows/CVI [13]. This software communicates with the Visual using USB 1.1 interface and infrared imager using RS-232C. The data is stored offline for further analysis. The software is flexible to stored at different frame rates and different resolutions. It also displays the storing data, which is being stored.
The resolution selected for VGA is 320x240 pixels while infrared resolution is fixed to 16x16 pixels. . The images are taken at 4 FPS even though the infrared imager is capable of up to 8FPS as here analysis is based on mostly individual images than time-based imaging analysis.
Three control experiment scenarios are proposed. Each scenario based on six experiments with differs in position, movement of subjects and different lightning conditions. The background images with no subjects are also taken for each experiment. Each experiment conducted contained around 150 visual and infrared samples of data stored on a hard disk. The length of each experiment varies from 3 to 5 minutes depending upon the subjects involved in the experiments. And during these experiments the data acquisition software is kept in running position.

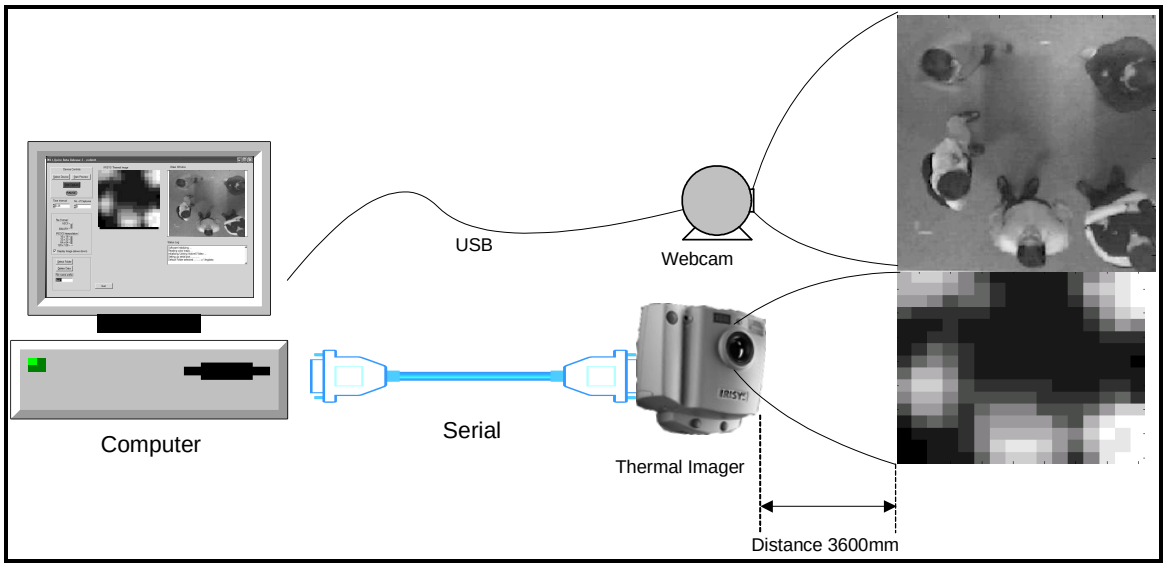

Figure 2 Experimental setup and data acquisition system

5. Image processing strategy

The visual system used is low cost than traditional CCTV cameras and 1/10th its cost. The low cost CMOS sensor used is by the proposed visual system also develops a noise factor, which also a major factor to be considered during the visual analysis. Thus images with simple subtraction with reference scene do not provide a consistent image in our case, which can be thresholded.

Visual analysis is very similar to that of done in [4] but the equipment used is low cost. The thermal imaging analysis is also done separately. Each analyses results is then further compared to increase the accuracy of the system as well as the system that will be developed is able to count during smoked filled room and other emergencies situations which is not the case with conventional visual counting system. Later on both visual and infrared data is analysed separately and then combined results are discussed.

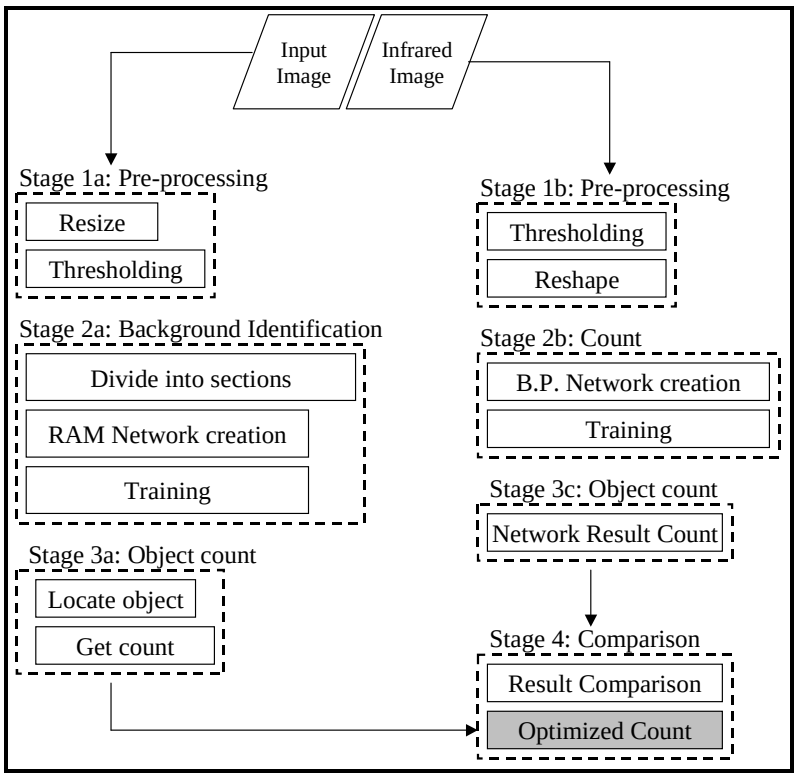

Figure 3 Overview of modified people counting system

5.1. Visual Analysis

The visual analysis is done using background identification technique used in [4] because this system does not use a standard approach which does not take light variations into account. This system by [4] is independent of light intensities in an image. Thus this process is chosen for visual analysis for the development of our system with some modifications like we do not require location information in an image, as it is not necessary in the proposed application, if required infrared imaging will do that with much accuracy. The visual counting system developed should be

Accuracy	Approximately 10%
Error	Maximum of +/-1 error in 4 to 10 people in a scene
Lightening conditions	Adaptable to any indoors lighting conditions
Adaptability	Most scene in indoors buildings

Table 1 Design guidelines for Visual counting system

5.1.1. Stage 1a: Pre-Processing

The pre-processing stage for visual analysis consists of resize and thresholding. The initial image acquired from the experiment is 288x288 pixels are reduced further up to 72x72 pixels. As the reduction in the resolution means a faster processing and faster counting rate with negligible amount of degradation in the thresholding result. For example the initial image of 288x288 pixels thresholding takes about 4.5 seconds using a considerably faster processing speed while 72x72 pixels take only about 2.5 seconds using MatLAB. This will obviously improve significantly when final development of system using programming languages like C or C++.

After resizing the reference images from each experiment conducted is taken. Reference images are just background images with no people in the scene. These reference images are thresholded by not the constant greyscale value but adaptive local thresholding is applied. The neighbouring pixel will allow the intensity of pixel to be compared with each other. If the comparison of these pixels is high up to a certain value set by another variable'α' the pixel is turned black otherwise white. Thus it can be mathematically expressed as:

$$[q]_{(i,j)} = \begin{Bmatrix} 1 \text{ if } \alpha > \left| r_{(i,j)} - r_{(i-1,j-1)} \right| \\ 0 \text{ if } \alpha \le \left| r_{(i,j)} - r_{(i-1,j-1)} \right| \end{Bmatrix}$$

Where

α = Thresholding value

r = original image

q = output thresholded image

i = 1 to 72

j = 1 to 72

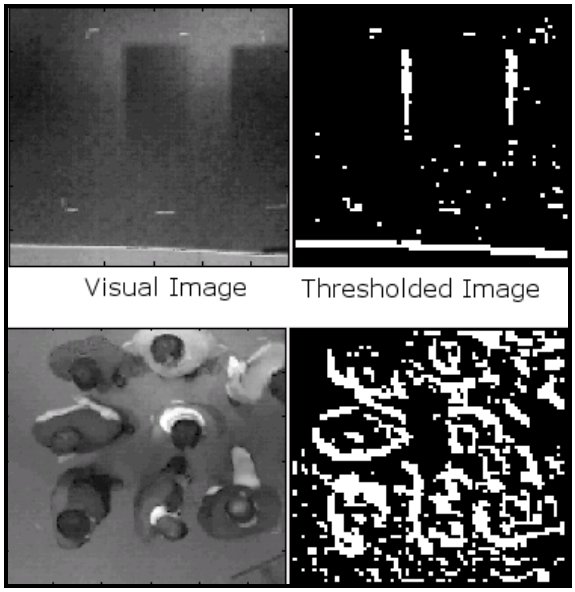

Figure 4 Above two images show no people standing while the below two images shows eight people standing in the scene

Here ’α’ is the global thresholding value of the image being processed. To calculate this value 1/3rd pixel values of images are randomly selected. The difference of intensities of these pixels is taken from their diagonal member. Here two constants ‘c’ and ‘d’ are introduced in threholding value of ’α’. After summing up all these intensity difference values, the final value is multiplied with constant ‘c’, which is less than 1. The value acquired is then summed up with the constant value of ‘d’. Thresholding expression is mathematically expressed as:

$$\alpha = d + c\sum\left|q_{(k,l)} - q_{(k-1,l-1)}\right|$$

Where

α = Thresholding value

d = Constant

c = Constant

q = Grayscale image

k = 1 to 72 (random values)

l = 1 to 72 (random values)

The optimal values of ‘c’ and ‘d’ are found by experimenting with the visual images taken during the experiment.

5.1.2. Stage 2a: Background Identification

The background identification is based on the RAM based neural network creation and training of that network. Only background images are trained using this network.

5.1.2.1. Network Creation

The thresholded image is dived into 4x4 sections. In this 4x4 section of an image randomly 4 pixels groups are selected. This random selection remains constant over the period. These randomly selected 4 pixels groups are used as the addresses of

RAM. For each 4x4 section created and randomly selected 4 pixels sections creates a single classifier. The group of 4 pixels is used as an address of the RAM created.

5.1.2.2. Training & testing

Training of a RAM Based network is done by reading the 4 pixels from each group in 4x4 section outputs 1 to the RAM of that certain address as shown in. For example if the 4 randomly selected pixels value is 0101 then it outputs 1 to the corresponding memory output of that address. Then it starts summing up all values the memory addresses, which are specific for each individual 4 pixels group. Thus for every section the image seen it outputs 1 into the RAM of that section address. It goes on until all the background samples are trained for that network. There is no use of running the samples again through the network which the network already seen as the result will always be the same of that particular image.

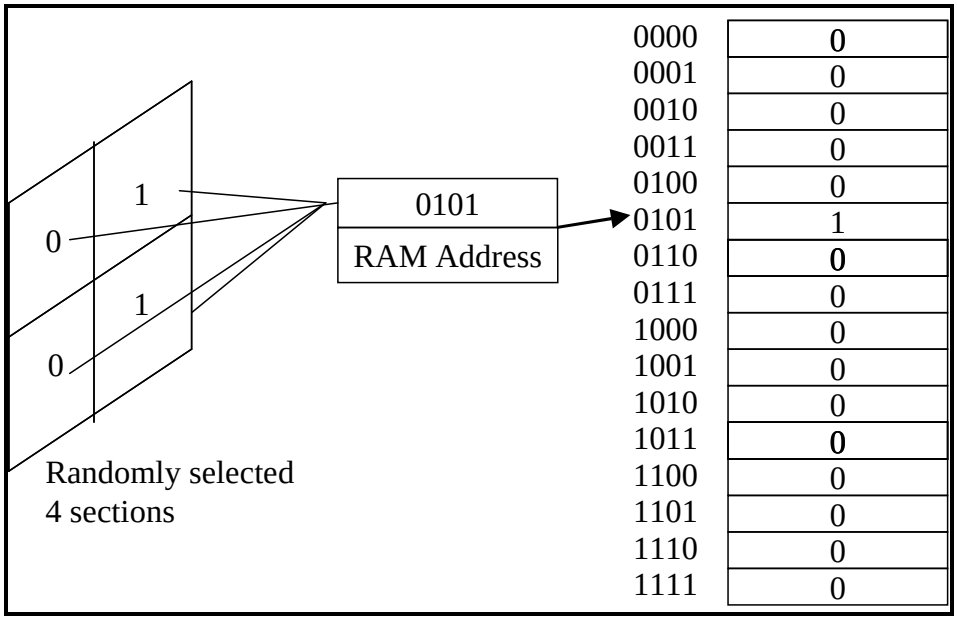

Figure 5 Example of RAM Neural Network training

To simulate the image using a trained network thresholded sample of the image is fed into the network. The sample image is then divided into the random sections, which are same as that of the trained network. The addresses of sample images are compared with the trained network values. If the network is already seen the same section during training it outputs '1' if the network hasn't seen anything like the section it outputs '0'. Output image is constructed with 1's as the background and 0's as the unseen object during the training. After inverting the image the unseen objects or peoples are then appears as a cluster of 1's in that image. 51 reference samples are used for training of RAM based neural network.

5.1.2.3. Object count

A 5x5 section is scanned over the output image by the neural network. For highest counts found in 5x5 section in the image, a count is incremented, and the 3x3 section in the middle is set to zero where as the 16 outside values are halved. This process is

continued until a certain cut-off value is achieved for the image. Optimized cut-off value is found by comparing the result found with the actual result.

5.2. Infrared Analysis

For the development of low-resolution infrared counting system the certain guidelines are laid which should be followed.

The infrared analysis system developed will be used in conjunction system but can be used as a stand-alone system with very slight modifications.

Accuracy	Approximately 5%
Error	Maximum of +/-1 error in 4 to 10 people in a scene
Adaptability	To most indoors building conditions and objects in scene (except extreme weather conditions like $+50^{O}$Celcius)
Lightening Variations	Completely prone to lightening variations in a scene

Table 2 Design guidelines for Infrared Counting system

The infrared analysis system developed will be used in conjunction system but can be used as a stand-alone system with very slight modifications.

5.2.1. Stage 1b: Pre-processing

Infrared data taken from the experiment are taken offline into MATLAB. The raw infrared data taken from the experiment is interpolated to find the 'average body heat'. The temperature of person is generally higher than the background except in very hot areas like desert but as this experiment is conducted inside building. Average heat of the background image in this experiment is found to be:

$$\text{averageheat} = \frac{\sum \text{pixels}}{256} \cong 24.3^{o}\,\text{Celsius}$$

While the internal temperature of the IRISYS® infrared camera remains 32.375^{O} Celsius. Thus the overall temperature ranges for the duration of our experiment remain to be within

$$\min \text{bodytemp} = 27^{o}\,\text{C}$$

$$\max \text{bodytemp} = 32^{o}\,\text{C}$$

$$\text{averagebodyheat} \cong 29.5^{o}\,\text{Celsius}$$

Thus the 'average body heat' calculated from the infrared data is then used as the thresholding value for the experiments conducted. This 'average body heat' is varies upon weather conditions and location of the experiment like if the experiment is conducted indoors or outdoors.

Infrared images of 16x16 pixels are processed using following equation.

$$[m]_{(j,k)} = \begin{cases} 1 \text{ if } x \geq \delta \\ 0 \text{ if } x \geq \delta \end{cases}$$

$$where\ j = 16;$$

$$k = 16;$$

Let x is an element of original matrix of 16x16 elements from infrared imager. Where m is an element of the thresholded matrix and δ is the average body heat. The infrared images after thresholding at average body heat give a distinguishable result that can be used for object recognition. But this is true only for small number of people as with the area under consideration gets crowded then the algorithm becomes unreliable.

5.2.2. Stage 2b: Back Propagation Neural Network

For infrared image counting neural networks area are selected, as the images are of small up to 16x16 pixels. After thresholding of infrared images the images are trained on back propagation neural networks.

5.2.2.1. Construction and training

A back propagation neural network is created. The specification for the final network selected is follows:

Inputs	256
Hidden Layer	1
Hidden Layer Neurons	280
Hidden Layer Function	Sigmoid Function
Output Layer Neurons	1
Output Layer Function	PureLin Function
Training Performance goal achieved	0.00642496
Epochs	500
Learning rate	0.005
Training Samples	360

Table 3 Configuration of optimized neural network for Infrared Analysis

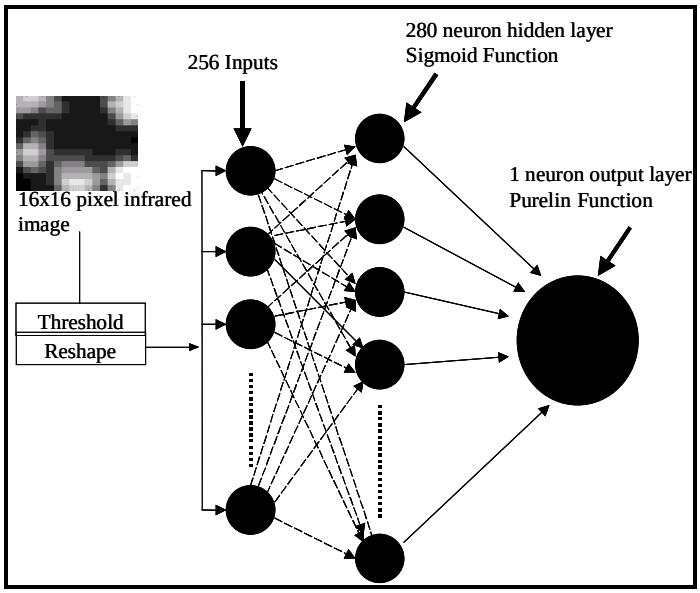

Figure 6 Back propagation Neural Network

Training of back propagation neural network is done by using twenty (20) samples from all eighteen (18) experiments is fed into the network.

6. Result

6.1.1. Infrared neural network simulation and result

The result acquired from the infrared data is plotted in the form of percentage error. The error plot is based on the simulation of selected 200 samples each from 18 experiments. The error increases as the people counting in the scene increases.

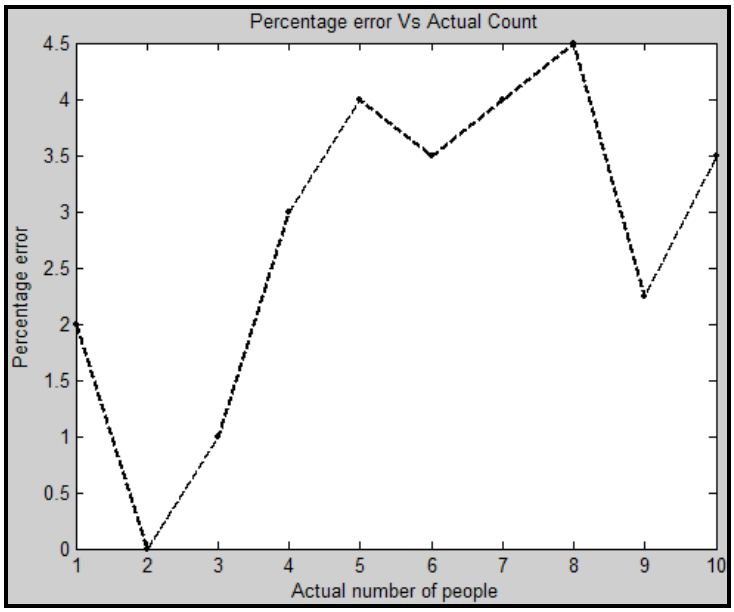

Figure 7 Infrared Image neural network error percentage

6.1.2. Visual RAM based neural network simulation and result

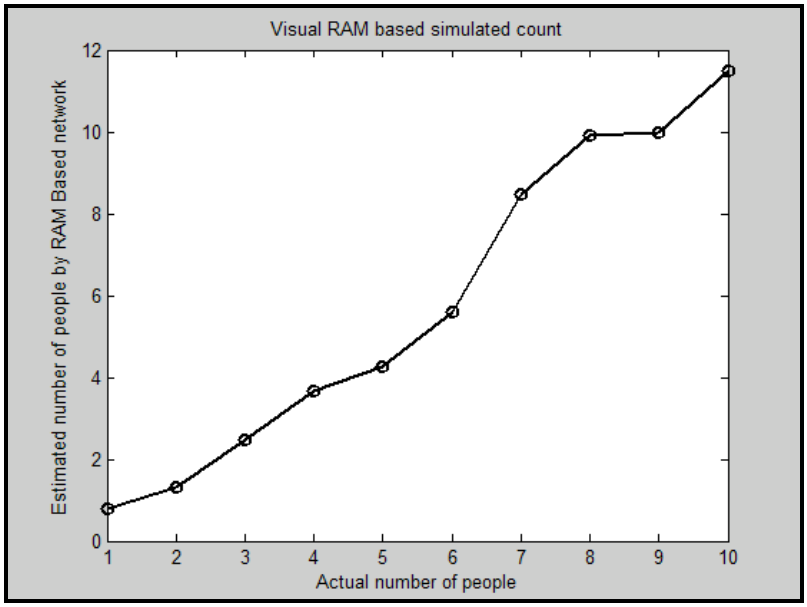

Figure 8 Visual RAM based simulation result comparison

6.1.3. Combined result of visual and infrared system

The results of infrared and visual system combined gives the average of less than 2% error in the system. Even for the high number of people count in that scene.

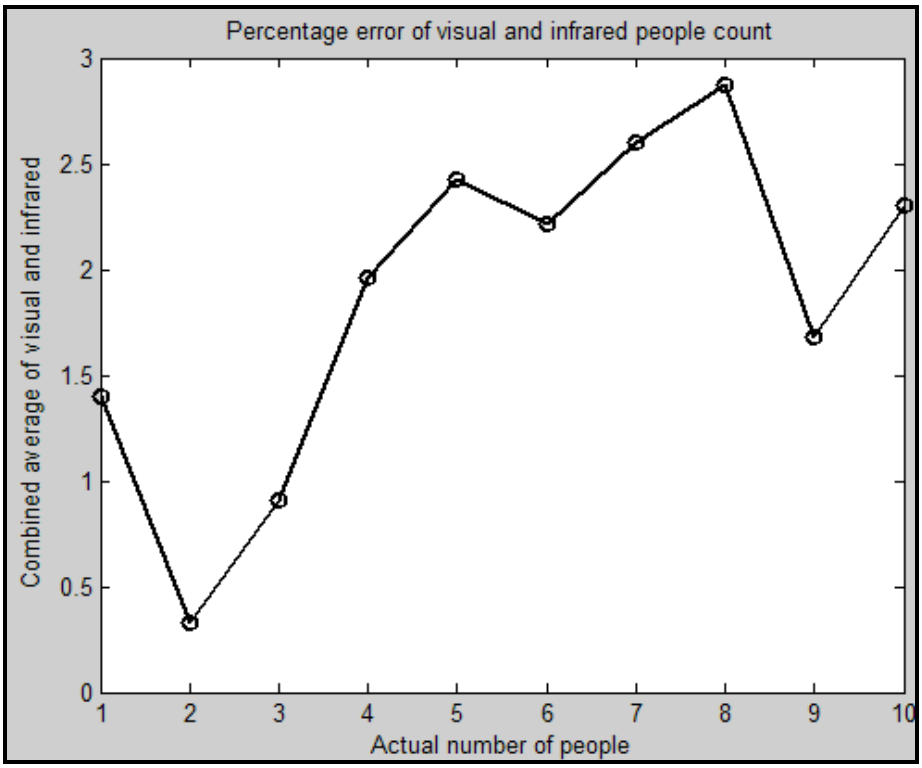

Figure 9 Combined percentage error of visual and infrared signal

References

1. Marotz, J., C. Lubbert, and W. Eisenbei[ss], *Effective object recognition for automated counting of colonies in Petri dishes (automated colony counting)1*. Computer Methods and Programs in Biomedicine, 2001. **66**(2-3): p. 183-198.
2. Morinaka, K., et al., *Human information sensor.* Sensors and Actuators A: Physical, 1998. **66**(1-3): p. 1-8.
3. Chow, T.W.S. and S.-Y. Cho, *Industrial neural vision system for underground railway station platform surveillance.* Advanced Engineering Informatics, 2002. **16**(1): p. 73-83.
4. Schofield, A.J., T.J. Stonham, and P.A. Mehta, *Automated people counting to aid lift control.* Automation in Construction, 1997. **6**(5-6): p. 437-445.
5. Sacchi, C., et al., *Advanced image-processing tools for counting people in tourist site-monitoring applications*1*. Signal Processing, 2001. **81**(5): p. 1017-1040.
6. Burnay, S.G., T.L. Williams, and C.H. Jones, *Applications of Thermal Imaging*. 1998, Bristol, Great Britain: I O P Publishing Ltd.
7. *Non-contact temperature measurement*. in *Tansactions in measurement and control, Vol.1, 3rd Edition, OMEGA*.
8. Holst, G.C., *Common Sense Approach to thermal Imaging*. 2000: SPIE Optical Engineering Press.
9. Miller, J.L., *Principles of Infrared Technology: A practicle guide to the state of the art*. 1994: Van Nstrand Reinhold.
10. *IRISYS : The Affordable Thermal Imager*, InfraRed Integrated Systems Ltd.
11. Al-Habaibeh, A. and R. Parkin, *An autonomous low-cost infrared system for the on-line monitoring of manufacturing processes using novelty detection.* International Journal of Advanced Manufacturing Technology, 2003. **22**(3-4): p. 249-258.
12. Haykin, S., *Neural Networks: A Comprehensive Foundation*, ed. Second. 1999: Prentice Hall.
13. *National Instruments* www.ni.com/labwindows.