Neighbourhood Information Feature Efficacy in Unsupervised Segmentation Algorithms for Breast Ultrasound Images

Dylan Drover
Systems Design Engineering
University of Waterloo
Waterloo, Canada
Email: djdrover@uwaterloo.ca

***Abstract*—The following paper will explore the efficacy of a novel feature, the neighbourhood fuzzy histogram membership (NFHM), on unsupervised image segmentation of breast ultrasound images. Three unsupervised learning algorithms, k-means, fuzzy c-means and the self-organizing map (SOM) neural network have be augmented with the NFHM feature to determine its efficacy in segmenting breast ultrasound (BUS) images. The algorithms were tested using only pixel intensities for segmentation and were then trained with pixel intensities as well the the new feature. Improvements in similarity of segmented images to the provided ground truth images were 17.99% for k-means, 13.79% for fuzzy c-means and 12.91% for the SOM when NFHM was incorporated. The overall success rates of the algorithms were relatively low, however the novel feature was shown to drastically improve their success rate when segmenting BUS.**

***Index Terms*—breast ultrasound, k-means, fuzzy c-means, self-organizing maps, image segmentation, computer aided diagnosis**

I. Introduction

It is estimated that in 2013 there were be approximately 23800 women diagnosed with breast cancer [1] in Canada alone. Of this number, an expected 5000 cases will be fatal [1]. This places breast cancer as the second leading cause of cancer death in women. It is estimated that 10%-30% of breast cancer cases go undiagnosed [2], [3], [4]. As a diagnostic tool, breast ultrasounds (BUS) are a non-invasive, low risk method that are considered a second best option to digital mammography [5] [6]. Tissue within the breast is complex and includes numerous different types of structures including mammary gland lobules, Cooper's ligaments, lymph nodes and fats. Thus, the images generated from BUS are complex and often noisy. This creates a unique challenge when trying to develop augmenting features for unsupervised learning methods.

A major component of breast cancer diagnosis is determining the characteristics of a growth. The relative length, width and shape of an abnormality must be identified and marked by a medical professional (*e.g,* oncologist, radiologist or radiographer). These attributes will help the medical professional to determine whether the growth is benign or malignant. Limiting misdiagnosis and accurate detection are paramount when examining medical images such as breast ultrasounds. With an increase in the number of patients per medical professional [1] the pressure on the individual is increased and can lead to fatigue and inaccurate results. It is difficult to obtain a definite ground truth.

To expedite diagnosis of breast cancer and relieve pressure on the medical system, it is sensible to employ intelligent algorithms to segment images. Other machine learning techniques will then be able to classify augmented regions of interest from the segmented image. Automatic segmentation would also help by serving as a potential touch-stone for different medical professionals who are making diagnoses based on noisy ultrasound images. To achieve this automatic segmentation it is advantageous to use features that enhance the differences between separate areas which allows for easier identification of suspicious areas. Other works have used various features from both the spatial and frequency domains [7] with success in improving the performance of various segmentation algorithms. In the following paper, a novel feature which encodes a pixel's similarity to its neighbourhood is incorporated into three well known and successful clustering algorithms, self-organizing maps, fuzzy c-means and k-means clustering.

The aforementioned choices for the clustering algorithms were motivated by their unsupervised nature. This allows for the learning to take place without the need of labelled data. The choices have proven track records, with fuzzy c-means and self-organizing maps being used to segment BUS images [7], [8], [9], [10]. It has been shown that using self organizing maps for image segmentation is very suitable method for discerning between colours [11] within an image. Fuzzy c-means has also shown success in the area of colour image segmentation [12].

The problems presented by image segmentation, especially using BUS images, is that it proves to be very difficult to get the algorithms to obtain the salient areas a good percentage of the time. The following paper will thus be focusing on developing and testing a novel image feature that contains membership information about a pixel's surroundings. This feature will be used to augment simple implementations of the aforementioned three clustering algorithms.

II. Background Review

A. Selected Algorithms

In the field of machine intelligence there are numerous methods for unsupervised learning. One of the more adaptive tools is the self-organizing map (SOM), which has the ability to extract features and reduce dimensionality within a noisy environment [13]. The SOM has also been shown to act in a similar fashion to some areas of the human visual cortex [13]. It will be shown later that this method provided the best results both with and without the augmenting feature. The SOM works by using a competitive learning technique that rewards output neurons that have weights that are similar to a given input. I in equation (1) is a euclidean distance measure. In equation 2, w_{ij} is the winning neuron's weight that will be updated. Depending on the neighbourhood size, the weights in the winning node's area will also be updated [14].

$$I = ||x - w_c|| = min_{ij}||x - w_{ij}|| \tag{1}$$

$$w_{ij}(k+1) = \begin{cases} w_{ij}(k) + \alpha(k)[x - w_{ij}(k)] & if(i,j) \in N_c(k) \\ w_{ij}(k) & if(i,j) \notin N_c(k) \end{cases} \tag{2}$$

The SOM's effectiveness is not limited to one method of medical imaging. Chang and Teng developed a generalized two stage SOM to segment a variety of medical images [15]. Their method exploited colour differences to discern between different areas in magnetic resonance, x-ray and ultrasound images. They employed a two layer method for extracting regions within an image, first getting general areas, eliminating outliers and finally merging similar regions.

Li and Chi as well as Logeswari and Karnan [16] [17] have delved into compartmentalizing brain regions. Li and Chi explored the use of Markov Random Fields to modify the weights of the traditional SOM to add emphasis on the spacial region in which a pixel was located.

The fuzzy c-means method was chosen because of its previous success in segmenting colour images like in the work of Tan and Isa [12]. FCM has also been used for breast ultrasound images [18]. It and its more primitive ancestor, K-means clustering are natural choices when dealing with unlabelled information in smaller dimensions (due to the issues with euclidean distance in higher dimensions). These algorithms were chosen to show that the neighbourhood fuzzy histogram membership feature can make non-trivial improvements to not just the SOM clustering algorithm but to other learning algorithms as well.

B. Relevant State of the Art Review

Medical image segmentation is a large and active field of research and there are a myriad of methods that are employed to perform segmentation and identification of suspicious areas. Within this realm there are many methods for segmenting breast ultrasound images as well.

Marcomini and Schiabel have used SOMs to perform segmentation of breast ultrasound images. [9]. Their results came from 10 features and provided accuracy of 91.33% with accuracy measured as $Accuracy = (TP + TN)/(TP + TN + FP + FN)$ with TP being true positive, TN being true negative, FP being false positive and FN being false negative. Their algorithm employed more extensive filtering and post processing fixes to provide better results than the ones presented in this paper. However the raw segmented images from their SOM were very similar to the results from the SOM created for this paper.

Another promising technique for BUS segmentation was developed by Shan et al. [8]. Their method had a similarity rate of 83.1%. They achieved this through the use of ROI generation and automatic seeding techniques. This was facilitated through new features including radial distance and phase in max-energy. These were incorporated with a neural net to improve overall similarity from 76.1% using joint probability to 83.1% using their additional features. This paper is an excellent illustration of the use of new features for improving segmentation. Similarly, the following results will show that the addition of the NFHM is also a noteworthy improvement to segmentation algorithms.

III. Methods and Materials

A. Digital Mammography Dataset

The dataset that will be used for training the various segmentation algorithms has been graciously provided by Segasist Technologies [19]. The dataset contains a total of 50 breast ultrasound images with 8-bit grey-scale colour. Each image has an abnormal area which is not specified as malignant or benign. Of the 50 images, a total of 45 of the images have a "ground truth" binary image that corresponds to the suspicious region. These 45 images will be used to train and evaluate the performance of the segmentation performed by the algorithms. Two pairs of corresponding raw images and ground truth diagnostics can be seen in Figure 1 and Figure 2. Figure 1 shows a simple tumour shape and Figure 2 shows an irregular shaped growth.

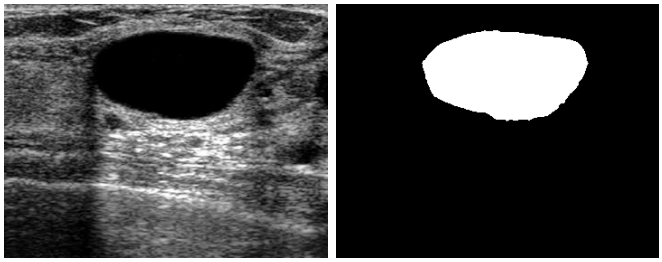


Fig. 1. Raw BUS image and ground truth

B. Algorithm

Some of the problems encountered through attempting image segmentation of breast ultrasound images are that while some images have very definite boundaries of a growth,

Fig. 2. Raw BUS image and ground truth

other images are very noisy and there is no clear delineation between healthy breast tissue and a tumour. The goal then is to provide some features that can be created using only the given image. The colour feature can be accessed through using pixel intensity but other information from the raw image needs to be generated. A new feature was developed and tested to encode neighbourhood information for each pixel.

The newly proposed method of encoding the local membership of a pixel is what will be called a neighbourhood fuzzy-histogram membership (NFHM). The specified pixel is given a local area of interest as can be seen in Figure 3 with the red centre pixel being the specified pixel. This pixel will be given a membership value, similar to de-fuzzification in fuzzy-logic. If this value is high, the pixel can be said to belong to the specified area, whereas if the membership value is low, the pixel is an anomaly.

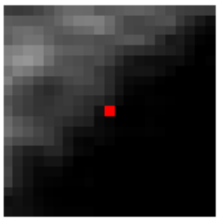
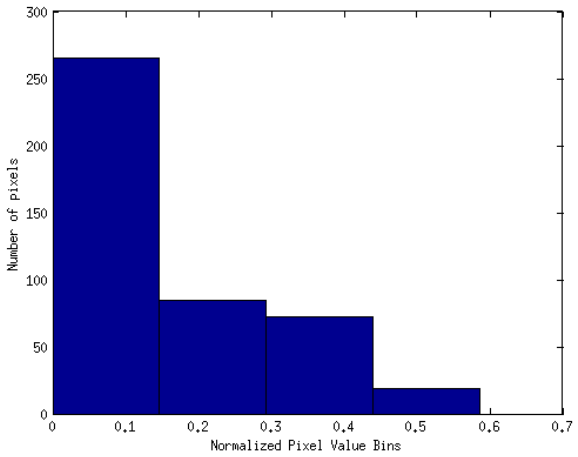


Fig. 3. Red centre pixel with square neighbourhood

A histogram is created for a pixel's neighbourhood, consisting of a specified number of bins for the pixel intensities (0 - 256). The completed histogram is normalized by the total number of pixels in the area. This leaves the bin with the most pixels as the best representation of the area, thus it has a membership closest to one. The centre pixel is not included in the initial histogram but its value is projected into its appropriate bin. The value of the respective bin is then mapped as the pixel's membership within that area. The algorithm will create a membership value for each pixel in the image. This method is suitable to help classify consistent areas as well as edges. Figure 4 shows the histogram from the previous area and the break down of each bin's membership value, the value of our centre pixel (shown as red) is actually 0.0781 (normalized) which sets it in the first bin. Therefore this pixel has a membership of 0.9199.

Using this method a new image based off membership values can be formed that illustrates certain areas of saliency



Bin	1	2	3	4
Membership	0.9199	0.2950	0.2499	0.0660

Fig. 4. Histogram and membership values of area from Figure 3

rather well. An example of a membership images of the images from Figures 1 and 2 can be seen in Figure 5.

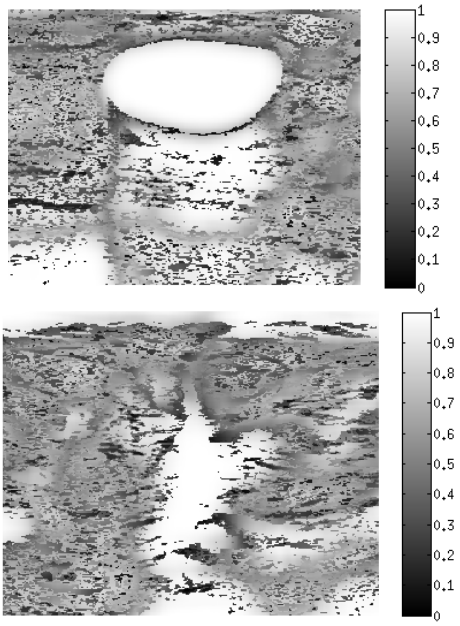


Fig. 5. Membership Images

The modifiable parameters of the NFHM are the neighbourhood size and the number of bins. These were heuristically optimized by testing different values and observing the level of difference they provided between a region of interest and surrounding tissue. As the neighbourhood became smaller, there was more granularity and small disturbances in the image would be picked up. A similar conclusion to the last was drawn in that the increase in the number of bins increases granularity in the membership images (they were noisier). Final values of

a neighbourhood size of 45 and a bin number of 4 were chosen.

The tool sets available from MATLAB were used to implement the SOM, fuzzy c-means and k-means algorithms. Each algorithm was initialised to separate the given inputs into sixteen clusters. Sixteen was chosen because it was one of the lower numbers that allowed the clustering algorithms to converge and halt. It also allowed there to be enough clusters that there would be no overlap between different coloured segments which would cause problems in final segmentation.

With the new feature prepared, the next step was to train the clustering algorithms. Histogram equalization was performed on the raw image to create a new contrast adjusted image from which the NFHM image was created. The equalized image's pixel values were also used as the pixel intensities for the input. Besides this, there was no other modification to the images. The first 20 images from the dataset were used as training data. The three clustering algorithms were initially trained using input vectors of size two, giving a feature space of two dimensions. The first value was a pixel intensity in an image the the other value was the pixel's membership value. From the 20 images 2,797,971 data points were created. The algorithms were trained a second time with only the pixel intensities. This instance of the algorithms would be separate from the first instance. The training set size allowed each algorithm to terminate.

The trained algorithms were then tasked to segment all images in the set. This process produced a "segmented" image, which assigned each pixel value to one of sixteen clusters. The cluster that corresponded to tumours was then extracted and a new "identified", final segmented image was produced. The final segmented image was a binary image with pixels of value 1 identifying the area of the growth and pixels of value 0 representing the background. These respective images can be seen in Figure 6.

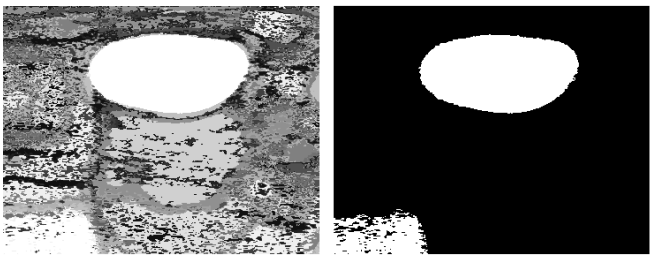


Fig. 6. Segmented result with corresponding identified region

To determine the success of the identification process, the final image was compared against the ground truth image provided in the database. This was done by taking the intersection and union of the images and finding the number of pixels of value 1 that were either common or unique between the two. The size of the array containing the common pixels and the size pf the array containing the unique pixels was used in a ratio to assess the similarity of the two images. The equations used are shown in equations 3, 4 & 5 where SI is similarity, ID is the identified image and G is the ground truth image.

$$a = findOnes(ID \cap G) \tag{3}$$

$$b = findOnes(ID \cup G) \tag{4}$$

$$SI = \frac{length(a)}{length(b)} \tag{5}$$

IV. Results

Two examples of completed segmented images from the augmented self organizing map can be seen in Figure 7 and Figure 8 where image (a) is the initial image, (b) is the segmentation performed by the augmented SOM and (c) is the ground truth image.

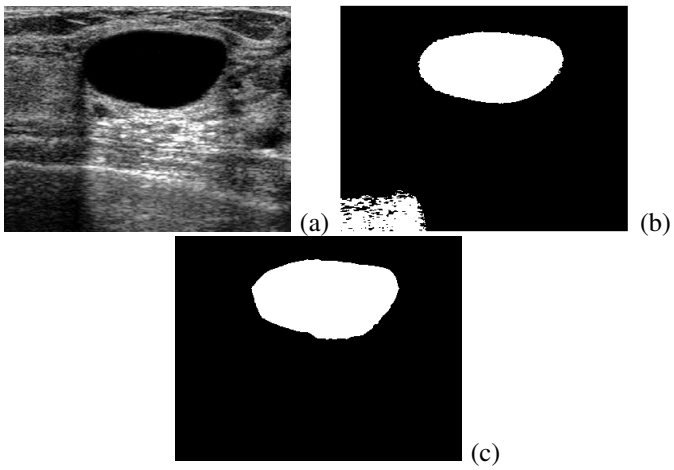


Fig. 7. (a) Original image, (b) identified image using NFHM-SOM, (c) ground truth image

The algorithms performed well for images similar to the ones in Figure 7. This can be attributed to well defined abnormal areas with good contrast difference as well as clear lines as to where a growth begins and ends. The similarity percentage for this image was 84% when using the augmented SOM which was the best overall method.

In images which had more complex abnormal structures or structures that were not clearly delineated from the background tissue, the results were similar to in Figure 8. It is clear that the algorithm can get a general shape of the growth but its similarity is far from optimal. The best algorithm still gets areas of false positives (as seen in Figure 8) and misses more ambiguous areas of growths. The similarity for the segmentation in Figure 8 was 48.5%.

Upon the completion of the segmentation for each image, the percentage of similarity (SI) as described in equations 3-5, was calculated with respect to its ground truth. This value was stored and the segmentation for the next image was begun. The mean similarity, standard deviation, maximum and minimum were taken from this data. In Table I the results for the non-augmented clustering algorithms can be seen.

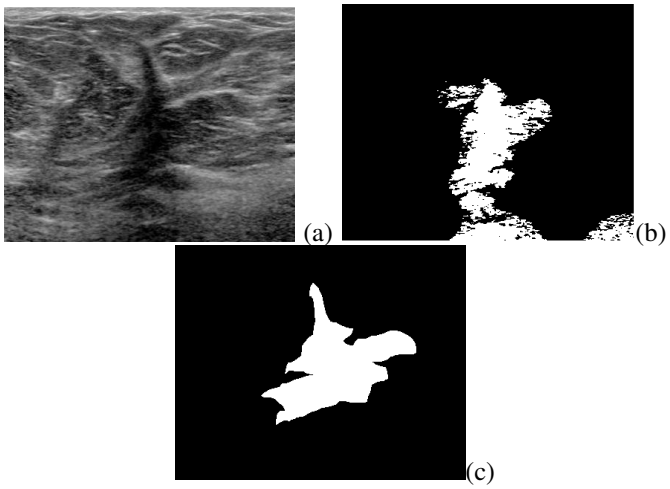


Fig. 8. (a) Original image, (b) identified image using NFHM-SOM, (c) ground truth image

Method	Mean Sim.(%)	Std. Dev.(%)	Min (%)	Max(%)
K-means	17.87	18.98	0	61.94
FCM	21.09	20.94	0	72.68
SOM	30.04	22.82	0	82.14

TABLE I
SEGMENTATION STATISTICS WITHOUT NFHM FEATURE

These results are far from the state of the art results, though the results were taken with a minimal amount of pre and post processing. The main goal of the paper is to gage the effectiveness of the new NFHM feature in improving segmentation results. In Table II the results for the neighbourhood fuzzy histogram membership augmented algorithms can be viewed. There are clear improvements in the mean similarities as well as improvements in the best results. The improvement to the mean similarity percentages can be seen in Table III.

Method	Mean Sim.(%)	Std. Dev.(%)	Min (%)	Max(%)
K-means	35.86	25.12	0	88.60
FCM	34.79	24.39	0	84.67
SOM	42.95	23.77	3.59	90.3

TABLE II
SEGMENTATION STATISTICS WITH NFHM FEATURE

Method	Improvement (%)
K-means	17.99
Fuzzy C-means	13.79
SOM	12.91

TABLE III
IMPROVEMENT TO MEAN SIMILARITIES WITH ADDITION OF NFHM

V. CONCLUSIONS AND FUTURE WORK

From the results that have been described, it is clear that the neighbourhood fuzzy histogram membership has been effective in improving the segmentation abilities of various segmentation algorithms. While the overall segmentation was far from the state of the art methods described in Section 2, the goal of showing NFHM is effective in improving segmentation results across the numerous unsupervised learning methods has been achieved.

Future goals would be to incorporate more features into the algorithms in order to increase similarity and percentage of accuracy to levels similar to or greater than the state of the art methods. This could be achieved by incorporating frequency domain characteristics as they would provide a different feature facet compared to the spatial attributes given by NFHM. Further pre-processing and post-processing could be done on the input data and the output of the algorithms to improve their similarity measures. It would also be desirable to incorporate a supervised learning algorithm such that the segmented images can be classified as either benign or malignant. This could provide a full stack solution that would help medical professionals in their efforts to successfully diagnose breast cancer.

Application of the above algorithms could also be extended to other areas of image segmentation. Since BUS images are considered rather noisy and difficult to segment, it would be desirable to test the algorithms on simpler images to observe their performance.

VI. ACKNOWLEDGEMENTS

The author would like to thank Segasist Technologies for providing the breast ultrasound image data set as well as Professor Hamid R. Tizhoosh for providing the formulae for measuring similarity.

REFERENCES

[1] Canadian Cancer Societys Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2013. Toronto, ON: Canadian Cancer Society; 2013.

[2] Bird RE, Wallace T, Yankaskas B. Analysis of cancers missed at screening mammography. Radiology 1992;184(3):613-7.

[3] Kerlikowske K, et al. Performance of screening mammography among women with and without a first-degree relative with breast cancer. Ann Intern Med 2000;133(11):855.

[4] Giger ML. Computer-aided diagnosis in radiology. Academy Radiologist 2002;9(1):1.

[5] E. D. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. K. Baum, S. Acharyya, E. F. Conant, L. L. Fajardo, L. Bassett, C. DOrsi, R. Jong, and M. Rebner, Diagnostic performance of digital versus film mammography for breast-cancer screening., 2005.

[6] A. Jalalian, S. B. T. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review., Clin. Imaging, vol. 37, no. 3, pp. 4206, 2013.

[7] M. Xian, Y. Zhang, and H. D. Cheng, Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains, Pattern Recognit., vol. 48, no. 2, pp. 485497, Feb. 2015.

[8] J. Shan, H. D. Cheng, and Y. Wang, Completely automated segmentation approach for breast ultrasound images using multiple-domain features., Ultrasound Med. Biol., vol. 38, no. 2, pp. 26275, Feb. 2012.

[9] K. D. Marcomini and H. Schiabel, Nodules Segmentation in Breast Ultrasound using the Artificial Neural Network Self- Organizing Map, vol. II, pp. 47, 2012.

[10] W. K. Moon, S.-C. Chang, C.-S. Huang, and R.-F. Chang, Breast tumor classification using fuzzy clustering for breast elastography., Ultrasound Med. Biol., vol. 37, no. 5, pp. 7008, May 2011.

[11] S. Ong, N. Yeo, K. Lee, Y. Venkatesh, and D. Cao, Segmentation of color images using a two-stage self-organizing network, Image Vis. Comput., vol. 20, no. 4, pp. 279289, Apr. 2002.

[12] K. Siang Tan and N. A. Mat Isa, Color image segmentation using histogram thresholding Fuzzy C-means hybrid approach, Pattern Recognit., vol. 44, no. 1, pp. 115, Jan. 2011.
[13] T. Kohonen, The self-organizing map, Proc. IEEE, vol. 78, no. 9, pp. 14641480, 1990.
[14] F. O. Karray and C. De Silva, "Major classes of neural network" in Soft Computing and Intelligent Systems Design. Essex, England: Addison Wesley, 2004, ch.5, sec.4, pp.268-269.
[15] P.-L. Chang and W.-G. Teng, Exploiting the Self-Organizing Map for Medical Image Segmentation, Twent. IEEE Int. Symp. Comput. Med. Syst., pp. 281288, Jun. 2007.
[16] Y. Li and Z. Chi, MR Brain image segmentation based on self-organizing map network, Int. J. Inf. Technol., vol. 11, no. 8, pp. 4553, 2005.
[17] T. Logeswari and M. Karnan, An Improved Implementation of Brain Tumor Detection Using Segmentation Based on Hierarchical Self Organizing Map, Int. J. Comput. Theory Eng., vol. 2, no. 4, pp. 591595, 2010.
[18] L. Zhang, Y. Ren, C. Huang, and F. Liu, A novel automatic tumor detection for breast cancer ultrasound Images, 2011 Eighth Int. Conf. Fuzzy Syst. Knowl. Discov., pp. 401404, Jul. 2011.
[19] OMISA Inc. - Segasist Technologies, Toronto, Ontario, Canada.