ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 1, Issue 8, December 2016

Pixel N-grams: Size, Location and Resolution Invariance for Shape Classification

[1][2]Pradnya Kulkarni , [1]Andrew Stranieri, [1]Julien Ugon
[1] Federation University, Australia, [2] MIT Polytechnic, India

***Abstract:*-- X-ray screening for breast cancer is an important public health initiative in the management of a leading cause of death for women. However, screening is expensive if mammograms are required to be manually assessed by radiologists. Moreover it is subjected to perception and interpretation errors. Automated mammogram classification is promising however relies on the identification of image features that enhance classification accuracies. Features that represent the shape of a tumour have been found to be useful for breast cancer detection using mammographic images. Existing shape feature computation methods are computationally expensive and strongly dependent upon algorithms that can segment an image to localize a region of interest likely to contain the tumour. In this paper, we apply a novel feature extraction technique called Pixel N-grams inspired from character N-gram model in text categorization for classification of shape images. Experiments on a dataset constructed for the purpose, demonstrate that the Pixel N-gram features achieve promising shape classification results irrespective of the size and location of the shape in an image without segmentation. These features also achieve excellent classification accuracy for images of varying resolution. Further, Pixel N-gram features are computationally less complex to generate paving the way for mammogram classification on low resource computers.**

***Index Terms:*-- Breast cancer, Classification, Mammograms, Pixel, N-grams, Shape.**

I. INTRODUCTION

Early detection of breast cancer is widely practiced by screening mammograms for abnormalities [1]. However, the detection of lesions is a difficult and time consuming task for radiologists. Computer algorithms for the interpretation of mammographic images can enhance the productivity of radiologists and reduce the cost of breast screening initiatives. Abnormalities/lesions visible in mammograms are mainly characterized by shape and texture [2]. Automated breast cancer detection can be cast as an image classification problem using features based on shape and texture.

Shape feature extraction techniques including Zernike moments [2] and texture feature extraction techniques such as histogram, co-occurrence matrix [3] and local binary patterns [4] have been used for mammographic classification. These features are known as low level image features and are global image features that generally fail to distinguish between local variations within an image.

To capture the local variations in an image a Bag-of-Visual-Words (BoVW) model has been advanced[5]. This model originated from the Bag-of-Words (BoW) model in the text retrieval domain. Recent studies on image classification have shown that use of Bag-of-Visual-Words (BoVW) model for image representation provide better classification results than the existing low level features such as colour, texture and shape [6]. However, the limitations of this approach are: firstly the spatial relationships between the visual words are not represented; secondly, the algorithms are complex and thirdly, BoVW are prone to noisy words.

In recent years, representation known as the character N-gram model has become more commonly used in text classification than BoW models for languages such as Chinese, which lack specific word boundaries [7]. Character N-grams are sequences of N consecutive letters in a sentence. For example, the 3-grams in the phrase “this dog” are “thi, his, is_d, s_d, _do, dog”; the four grams are “this, his_,is_d, s_do, _dog”. Inspired from the character N-grams in text a novel image representation model called ‘Pixel N-grams’ is proposed for mammographic classification [8].

In the Pixel N-gram model, the Gray scale intensity values of N consecutive pixels in an image are considered so that the image can be represented using frequency of occurrence of each Pixel N-gram present in an image. The advantage of the Pixel N-gram approach over BoVW approach is that firstly, the algorithms are computationally simple; secondly, spatial relationships are taken into account and thirdly, information is not lost in the vocabulary construction process. The preliminary results of Pixel N-grams approach for mammographic classification

International Journal of Science, Engineering and Management (IJSEM)
Vol 1, Issue 8, December 2016

are promising [9]. It has been demonstrated in our earlier work [10] that the Pixel N-grams perform well at classifying texture images.

This article reports on the application of the Pixel N-grams model to the detection of predefined shapes in an image, regardless of the size or location of the shapes. This is motivated by the observations that lesions in a mammogram can be of different sizes and located at different locations in an image. Moreover, the mammographic images taken at different hospitals may have varying resolutions. A model that classifies mammograms containing lesions of various shapes and sizes regardless of image resolution, from those that do not contain lesions will potentially lead to classification systems that are clinically useful in practice. Mammography is a proven method for early detection of breast cancer however the manual analysis of screening mammograms is expensive and is subjected to perception and interpretation errors [11]. Algorithms that accurately and sensitively classify mammograms for abnormalities while executing with minimal computational resources are challenging.

This paper reports on the extent to which classification using pixel N-grams is independent of size and location of a shape in an image. Further, verification of the resolution independency of these features is also important. For simplicity, a database of binary images consisting of three basic shapes (circle, square, triangle) of different sizes, locations and resolutions has been constructed for the experiments. The rest of the paper is organized as follows. Section II describes relevant literature in the area of shape classification and mammographic classification, Pixel N-grams features are briefly explained in section III, the experimental setup is elaborated in section IV, the results are discussed in section V and section VI and section VII concludes the paper.

II. LITERATURE REVIEW

Computing a shape feature consistent with human perception is very difficult. The literature shows that the shape feature computation can be classified broadly into main two categories: 1) Contour based methods and 2) Region based methods [12]. Contour based methods are based on the shape boundary points whereas the region based methods are based on a shape's interior points.

Another way to classify shape features is according to the processing approaches. The following are some of the commonly used approaches:

- Shape signature by 1-D function. e.g. curvature function, area function etc.
- Polynomial approximation (merging or splitting).
- Spatial interrelation – e.g. principal axis, bounding box, chain code
- Moments- e.g. Zernike moments
- Shape transform- e.g. Fourier, wavelet, R-transform

Shape signatures are computationally simple, however are sensitive to noise [13]; whereas, polygon approximation eliminates noise and leads to simplification of shapes. Polygon approximation is normally used as a preprocessing step [14]. Spatial interrelation describe the region or contour by using geometric features such as curvature, area, location, and length. This model provides compact and meaningful features. Bounding box and chain code are some of the examples of this model. Bounding box is invariant to scaling, rotation and translation and is also robust to noise [15]. A boundary or region can also be described using moments.

The concept of moment is originated from physics. Although, moment features are translation and rotation invariant, noise sensitivity and information redundancy are two major drawbacks [16]. Though Zernike moments [17], which are robust to noise are an exception, however at the cost of computational complexity. Use of signal processing algorithms such as Fourier transform, wavelet transform, R-transform have also been used to represent shapes in an image. Using these transforms a shape description with different accuracy and efficiency can be achieved by choosing the number of transform coefficients. Other advantages of signal processing methods for shape description include high robustness to noise and the great coherence with human perception [18]. However, these transformations are also computationally resource intensive.

Abnormalities in mammograms are mainly characterised by shape and texture of a lesion [2]. Zernike moments [17] can be seen to represent the shape of the calcification in [2]. Spherical wavelet transforms is used for classification of masses into benign and malignant categories in [19] and found to work better than the discrete wavelet transform features [19]. Comparison of wavelet and curvelet features for classification of mammograms is found in [20] and the results suggest that the curvelet transform outperformed the wavelet transform. Similarly, the application of Fourier transform for mammographic classification was reported in [21]. Statistical techniques such as histogram [22] and Grey level Co-occurrence Matrix (GLCM) features [23] have been used for mass classification in mammograms. However, the disadvantage of GLCM is that they are computationally expensive. Local binary patterns are used in combination with Haralick's texture features and Haar wavelet features and found to be very effective for mammograms classification [24]. The main problem with the above stated shape descriptors is that they are computationally complex and need good segmentation algorithms to localise the lesion site. Furthermore, all these texture and shape features are known as low level image features and represent global image features. These features fail to distinguish between local pattern variations across an image.

A recent approach inspired from Bag-of-Words models in the text retrieval context has been put forward for image classification [25] [26]. This technique provides the visual analogy of a word by vector quantizing the local descriptors. The BoVW approach has been found to be more successful than low level image features such as colour, texture and shape [27]. However, it faces three main limitations: firstly, spatial relations between visual words are ignored; secondly, computational complexity is very high and thirdly, noisy words are resulted due to coarseness of vocabulary construction process.

To overcome some of the drawbacks of the existing techniques, Pixel N-grams has been proposed for mammographic lesion classification [8]. This technique is inspired from the character N-gram concept in text retrieval. In this technique an image is represented with the help of histogram of occurrence of pixel N-grams in an image. The main advantage of this technique is that the algorithms are simple and computationally inexpensive. Other advantages include: spatial relationships between various pixels are considered and information is not lost in the vocabulary construction process. The preliminary mammographic classification results using Pixel N-grams are quite promising [9]. The work in [28] demonstrates that these features show improved classification performance as compared to co-occurrence matrix features and are almost seven times faster to compute. The Pixel N-grams also show improved texture classification results as compared to the state of the art BoVW approach [10]. A more exhaustive review on the computer aided detection of breast cancer can be found in [29]. Pixel N-gram technique is described briefly in section III below.

III. PIXEL N-GRAMS

Pixel N-grams are inspired from the character N-gram concept in the text categorization or retrieval domain. The detailed explanation of the technique can be found in [9]. In text retrieval context, essentially, a character N-gram is a sequence of N consecutive characters in a sentence. Character N-grams have been very useful and efficient for text categorization in languages such as Chinese [30], which do not have specific word boundaries. Similarly, images are formed by bunch of pixels and it is very difficult to find visual word boundaries for an image. Character N-grams are language independent, robust to grammatical errors and do not require any text pre-processing (tokenizer, lemmatizer) or other NLP tools [30]. Furthermore, the character N-gram algorithms are very simple and require minimal computational resources.

Therefore, we infer that the application of a character N-gram model for image representation (Pixel N-grams) would be effective for image classification applications. However, use of the character N-gram model for images is not straightforward. While text documents have a single spatial direction, images are two dimensional and the sequence of pixels can be obtained in different orientations (vertical, horizontal or at diagonals). Deciding the right direction to interpret a given N-gram is also a challenging task. Additional thing to consider in case of an image is the effect of rotation. Pixel N-grams are the sequence of N consecutive gray level pixels in an image. As opposed to text, an N-gram with the same order but different orientation may refer to the same pattern in an image.

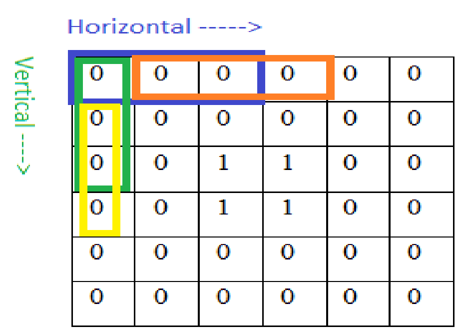

Fig 1: Sliding window for 3-gram calculation

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 1, Issue 8, December 2016

Fig 1 shows the sliding window for calculating 3-gram features in a horizontal and vertical direction for a binary image of square shape. Similarly, the N-grams can be considered in diagonal direction as well. For a binary image the total number of possible 3-grams in one direction is 23 = 8.

IV. EXPERIMENTAL SETUP

A. Dataset and Materials

A dataset of binary images consisting of three basic shapes (circle, triangle, square) was prepared for the experiments. The shapes were constructed as solid white on black background so that the impact of size and location only on classification accuracy could be examined. Three shapes, square, triangle and circle were selected because these are geometrically diverse and are basic. Classification results with simple, basic shapes of different sizes and locations can act to provide a baseline for more complex shapes.

For each shape, 80 images of different sizes and shape locations were created. For the first experiment all images were of size 512 × 512 pixels. For the second experiment, dataset consists of images of shapes (circle, square, triangle) with 10 different resolutions. Fig 2 shows some of the sample images from this dataset. All the classification experiments were conducted using Weka 3.6 data mining software developed at University of Waikato [31]. The machine used for experiments has i5-4210U CPU @2.90GHz PC with windows 10 (64 bit) operating system.

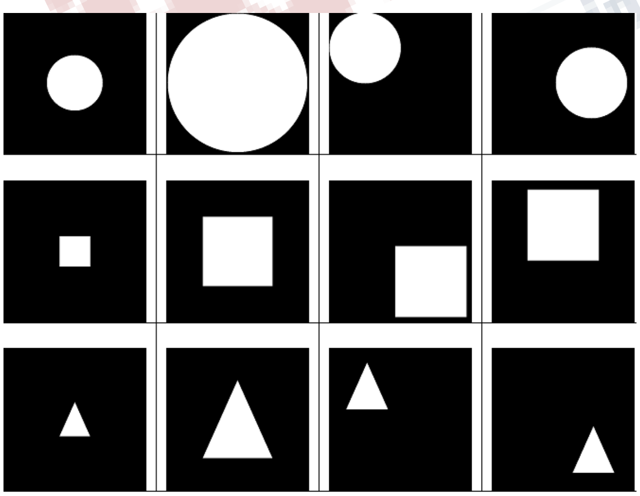

Fig 2: Shape database of binary images

The Multilayer Perceptron (MLP) Classifier in the Weka suite was used for classification and the generalization was estimated using leave one out cross-validation resampling.

B. Experiment 1 Shape and Location Invariance

In this experiment the 3-gram features in horizontal as well as vertical direction as described in the section III are computed for the images with different sized shapes and different locations of shapes in the dataset. The decision to consider 3-gram features is taken empirically by trying various values of N (2, 3, 4, 5). The images are binary and contain only two gray levels (black is considered 0 and white is considered 1). Therefore, we get the following N-grams for the shape images: 000, 001, 010, 100, 110, 011, 111. These frequency counts of each N-gram are then normalized so that they fall in the 0 to 1 range and used as inputs to the MLP classifier. For comparison of results, two other techniques were used for classification under the same setup: 1) histogram features, 2) co-occurrence matrix features (Contrast, Correlation, Energy, Homogeneity).

C. Experiment 2 Resolution Invariance

This experiment is conducted to check if the pixel N-grams features are able to classify different shapes under varying resolution of images. As mentioned in section III earlier, 3-gram features in horizontal and vertical directions are computed for the images of different resolutions in the dataset. Classification is performed using MLP classifier. Again the results are compared against two well-known existing techniques namely: 1) histogram and 2) co-occurrence matrix (Contrast, Correlation, Energy, Homogeneity).

V. RESULTS

Table I shows classification results for the experiment 1 and Table II shows results of the experiment 2. Both the tables list comparison of the proposed technique with the commonly used existing techniques (histogram and co-occurrence matrix features. The evaluation is performed based on three parameters (precision, recall and classification accuracy).

Table I: Shape classification (different size and location)

Features	Classification Accuracy	Precision	Recall
Histogram	87.9 %	89%	87%
Co-occurence Matrix	66.67%	50%	67%
Pixel N-grams	100%	100%	100%

Table II: Classification under varying resolution

Features	Classification Accuracy	Precision	Recall
Histogram	80%	80.2%	80%
Co-occurence Matrix	36.67%	36.4%	36.7%
Pixel N-grams	90%	90.9%	100%

VI. DISCUSSION

Results depicted in Table I demonstrate that the pixel N-gram features are able to clearly distinguish between various basic shapes (100% accuracy, precision and recall). The fact that the change in size and location of the shape in an image has no effect on the classification accuracy is potentially very useful for mammographic lesion classification as well as various other image classification and retrieval applications. Moreover, classification accuracy is compared with the existing techniques such as histogram and co-occurrence matrix features. It is observed that the classification accuracy is considerably better than these two existing techniques (100% as compared to 87.9% and 66.67%). Further, the precision with pixel N-grams is observed to be 100% as compared to 89% and 50%, whereas the recall rate is also found to be 100% as compared to 87% and 67% of the existing techniques.

Experiment 2 demonstrates that pixel N-gram features are good at classifying shapes irrespective of the resolution of the images. Thus these features can be used to build resolution independent classification systems. The classification accuracy resulted with pixel N-gram features is 90% as opposed to 80% of the histogram and 36.37% of the co-occurrence matrix features. Moreover, the precision (90.9% as compared to 80.2% and 36.4%) and recall (100% as compared to 80% and 36.7%) has been observed to be far better than these existing techniques.

The pixel N-gram features have been computed without segmentation algorithms so potentially by-passes challenges associated with segmenting images. As has been illustrated previously, N-gram features have been found to be seven times faster to compute than the co-occurrence matrix features [28].

VII. CONCLUSION AND FUTURE WORK

An important characteristic useful for breast cancer detection and classification is the shape of a lesion. In this study, novel Pixel N-gram features have been used for basic shape (circle, square, triangle) classification using a MLP classifier. Experiments clearly demonstrate that the pixel N-gram features can distinguish between various basic shapes irrespective of the size and location of the shape in an image. Additionally, these features can classify various shapes regardless of the resolution of an image. The classification results are significantly better than the existing histogram or co-occurrence matrix features. Above all, the Pixel N-gram features are simple and warrant reduction in computational complexity. The observations are promising but are limited by the use of a database of simple, basic shapes. Breast cancer lesions have texture and shape as two important distinguishing characteristics and are far from simple. Future work aims to incorporate experiments using more complex shapes and using benchmark database. Also, investigation of robustness of Pixel N-gram features to noise would be included in the future work.

REFERENCES

[1] W. I. Suleiman, S. J. Lewis, D. Georgian-Smith, M. G. Evanoff, and M. F. McEntee, "Number of mammography cases read per year is a strong predictor of sensitivity," *Journal of Medical Imaging*, vol. 1, pp. 015503-015503, 2014.

[2] C.-H. Wei, Y. Li, and P. J. Huang, "Mammogram retrieval through machine learning within BI-RADS standards," *Journal of biomedical informatics*, vol. 44, pp. 607-614, 2011.

[3] R. Nithya and B. Santhi, "Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer," *International Journal of Computer Applications*, vol. 28, pp. 21-25, 2011.

[4] X. Lladó, A. Oliver, J. Freixenet, R. Martí, and J. Martí, "A textural approach for mass false positive

reduction in mammography," *Computerized Medical Imaging and Graphics*, vol. 33, pp. 415-422, 2009.

[5] Y. Li, H. Chen, G. K. Rohde, C. Yao, and L. Cheng, "Texton analysis for mass classification in mammograms," *Pattern Recognition Letters*, vol. 52, pp. 87-93, 2015.

[6] G. V. Pedrosa and A. J. Traina, "From bag-of-visual-words to bag-of-visual-phrases using n-grams," in *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on*, 2013, pp. 304-311.

[7] I. Kanaris, K. Kanaris, I. Houvardas, and E. Stamatatos, "Words versus character n-grams for anti-spam filtering," *International Journal on Artificial Intelligence Tools*, vol. 16, pp. 1047-1067, 2007.

[8] P. Kulkarni, A. Stranieri, S. Kulkarni, J. Ugon, and M. Mittal, "Hybrid Technique based on N-gram and Neural Networks for Classification of Mammographic Images " presented at the International Conference on Signal, Image Processing and Pattern Recognition, Sydney, Australia, Feb 2014.

[9] P. Kulkarni, A. Stranieri, S. Kulkarni, J. Ugon, and M. Mittal, "Visual Character N-grams for Classification and Retrieval of Radiological Images," *The International Journal of Multimedia & Its Applications*, vol. 6, p. 35, Mar 2014.

[10] P. Kulkarni, Stranieri, A., Ugon, J., "Texture Image Classification using Pixel N-grams," presented at the IEEE International Conference on Signal and Image Processing, Beijing, China, Aug 2016.

[11] V. Palazzetti, F. Guidi, L. Ottaviani, G. Valeri, S. Baldassarre, and G. Giuseppetti, "Analysis of mammographic diagnostic errors in breast clinic," *La radiologia medica*, pp. 1-6, 2016.

[12] M. Yang, K. Kpalma, and J. Ronsin, "A survey of shape feature extraction techniques," *Pattern recognition*, pp. 43-90, 2008.

[13] D. Zhang and G. Lu, "A comparative study on shape retrieval using Fourier descriptors with different shape signatures," in *Proc. International Conference on Intelligent Multimedia and Distance Education (ICIMADE01)*, 2001.

[14] K. Selvakumar and B. K. Ray, "Survey on polygonal approximation techniques for digital planar curves," *International Journal ofInfonnation Technology, Modeling and Computing (UITMC) Vol*, vol. 1, 2013.

[15] C. Bauckhage and J. K. Tsotsos, "Bounding box splitting for robust shape classification," in *ICIP (2)*, 2005, pp. 478-481.

[16] M. E. Celebi and Y. A. Aslandogan, "A comparative study of three moment-based shape descriptors," in *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, 2005, pp. 788-793.

[17] G. A. Papakostas, Y. S. Boutalis, D. A. Karras, and B. G. Mertzios, "A new class of Zernike moments for computer vision applications," *Information Sciences*, vol. 177, pp. 2802-2819, 2007.

[18] R. B. Yadav, N. K. Nishchal, A. K. Gupta, and V. K. Rastogi, "Retrieval and classification of shape-based objects using Fourier, generic Fourier, and wavelet-Fourier descriptors technique: A comparative study," *Optics and Lasers in engineering*, vol. 45, pp. 695-708, 2007.

[19] P. Görgel, A. Sertbas, and O. N. Uçan, "Computer-aided classification of breast masses in mammogram images based on spherical wavelet transform and support vector machines," *Expert Systems*, vol. 32, pp. 155-164, 2015.

[20] M. M. Eltoukhy, I. Faye, and B. B. Samir, "Breast cancer diagnosis in digital mammogram using multiscale curvelet transform," *Computerized Medical Imaging and Graphics*, vol. 34, pp. 269-276, 2010.

[21] L. Shen, R. M. Rangayyan, and J. L. Desautels, "Application of shape analysis to mammographic calcifications," *IEEE transactions on medical imaging*, vol. 13, pp. 263-274, 1994.

[22] H. S. Sheshadri and A. Kandaswamy, "Breast tissue classification using statistical feature extraction of mammograms," 医用画像情報学会雑誌, vol. 23, pp. 105-107, 2006.

[23] M. T. Wong, X. He, H. Nguyen, and W.-C. Yeh, "Mass classification in digitized mammograms using

texture features and artificial neural network," in *International Conference on Neural Information Processing*, 2012, pp. 151-158.

[24] S. Joseph and K. Balakrishnan, "Local binary patterns, haar wavelet features and haralick texture features for mammogram image classification using artificial neural networks," in *Advances in Computing and Information Technology*, ed: Springer, 2011, pp. 107-114.

[25] E. Cheng, N. Xie, H. Ling, P. R. Bakic, A. D. Maidment, and V. Megalooikonomou, "Mammographic image classification using histogram intersection," in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, 2010, pp. 197-200.

[26] M. Jiang, S. Zhang, H. Li, and D. N. Metaxas, "Computer-aided diagnosis of mammographic masses using scalable image retrieval," *Biomedical Engineering, IEEE Transactions on*, vol. 62, pp. 783-792, 2015.

[27] C.-F. Tsai, "Bag-of-Words Representation in Image Annotation: A Review," *ISRN Artificial Intelligence*, vol. 2012, pp. 1-19, 2012.

[28] P. Kulkarni, Stranieri, A., Kulkarni, S., Ugon, J., & Mittal, M., "Analysis and Comparison of Co-occurrence Matrix and Pixel N-gram Features for Mammographic Images," presented at the International Conference on Communication and Computing, Banglore, India, 2015.

[29] A. Jalalian, S. B. 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," *Clinical imaging*, vol. 37, pp. 420-426, 2013.

[30] Z. Wei, D. Miao, J.-H. Chauchat, and C. Zhong, "Feature selection on Chinese text classification using character n-grams," in *Rough Sets and Knowledge Technology*, ed: Springer, 2008, pp. 500-507.

[31] W. Weka, "3: data mining software in Java," *University of Waikato, Hamilton, New Zealand (www. cs. waikato. ac. nz/ml/weka)*, 2011.

