# FEATURES SELECTION FOR IMAGE CLUSTERING USING LIMITED DEVICE

Monika Bhatnagar*

Dr. Prashant Kumar Singh**

## Abstract:

Image classification is becoming a prominent field with the passage of time. Image classification on limited devices such as cellular phone is also coming in demand as the technology is advancing. So for constructing classifier the use of clustering is coming into picture for its ability to use less space which is a prime factor for limited devices. Thus for constructing the classifiers several features have to be considered. This paper shows what features can be considered in order to create an efficient classifier on limited device such as cellular phone.

***Keywords-*** *Clustering,limied device, feature extraction*

[*] Research Scholar, Dr. K.N. Modi University, Newai, Rajasthan,India.

[**] ISC Software Private Limited, Bhopal,MP, India.

## I. INTRODUCTION:

We see use of gazettes at every step of life. Everywhere we turn around we can find one gazette or the other to make our life comfortable. Cellular phones are one such gazette. Not only that the cellular phones allow us to remain connected to our loved ones and professional connections, they also let us keep memories and information with us in form of the multimedia data like text, video, images and voices. Most of the work that has been done focuses the common people's general needs. We can do several tasks with our cellular phones, you want to connect there is phone, you want to share pictures and videos there is cellular phone, you want to convey some message to somebody there is cellular phone, you want to have any kind of data stored in some had held device, there is cellular phone. But cellular phone does not fulfill some specific needs of some common people. For example, if a farmer wants to decide the harvesting time of his crop, cellular phone is not going to help him which he might be carrying. A facility can thus be provided to the farmers that if they are having a cellular phone with the facility of camera which now a days almost all phones have, he can just click the picture of the crop and let the cellular phone decide whether the crop is ready for harvesting.

With the emergence of data mining, a large amount of possibilities have been developed to provide some standard solutions to our day to day life. In today's world we can see data mining tasks being performed almost everywhere where the concept of data lies. Therefore first thing that we can consider fulfilling our goal is data mining. With this we enter into the promising field of image processing and classification. Image processing and classification becomes a very ambitious task on cellular phones because of limited hardware and connectivity. Also we are choosing tomato for our example of crop.

To fulfill our need we need to analyze and understand various technical aspects related to the problem. To begin with our work we have already decided to make use of data mining. Data mining is one of the steps of Knowledge Discovery in Database process[1]. The data mining process is used to extract knowledge from a data set such that it can be easily understood by anybody. The various terms related to data mining are database , data management, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structure, visualization and online updating. The various classes of tasks of data mining are Association Rule Mining, Classification, Clustering, Regression, Anomaly Detection and

Summarization. Each class of data mining tasks provides one or the other solutions to the day to day problems.

In our case we need to provide facility in the cellular phone to make comparison between the pictures of the crop which is ready to harvest with the one we need to know whether it is ready to harvest. This leads us to the need of classifying the pictures. We can make use of several classification tasks of Data mining such as Association Rule Mining, Classification Rule Mining and Clustering. Since classifiers are very complex they are not preferred to be implemented on the cellular phones [2] because of computational overload on them. Therefore, we are choosing Clustering task of Data mining to create the classifier.

Clustering is the data mining technique which is often considered as been similar to classification[1]. The similarity between clustering and classification is that both the techniques bring together similar data together. But there is a big difference between them. Clustering can be understood as bringing similar objects into a group called cluster[3]. But the groups are not predefined. In case of Classification the groups are predefined. Which means in case of Classification what should be part of the group is pre-decided and in case of Clustering the grouping is accomplished by finding similarities between data according to characteristics found in the actual data.

Now since we have decided for the classification technique for the image classification we are now required to look for the operating systems of the cellular phones. Now days there are several operating systems available with different cellular phones of various companies like Android from Google, BlackBerry RIM, iOS from Apple Inc., Symbian OS from Symbian Foundation, Windows Phone from Microsoft and webOS from HP. Most of the Mobile operating systems are closed source. Android being an open source operating system we prefer it to show the utility of the work done by us. Android is an open source, Linux-derived OS backed by Google. The Android operating system is preferable for benchmarking due to its recent growth in popularity with varying hardware manufactures e.g. HTC, Motorola, and Samsung. The Android operating system is supported and a part of the Open Handset Alliance. This alliance positions key manufacturers, cellular providers and the Android operating system in a collaborative environment which has caused large growth since October 2008 when the first Android mobile phone was released.

## II. LITERATURE SURVEY:

With emergence of data mining and classification techniques we see that image classification is emerging as an exciting and ambitious field. A lot work has been done in the field of image classification. Since work has been done in image classification a lot of work has been done to show various techniques of classification. We are proposing a handy solution for farmers. So the first requirement of image classification is to have image classification of any crop. Now since we have chosen tomato for our crop we found that a lot of work has been done on the image classification of tomatoes.

G. Polder, G. W. A. M. van der Heijden, I. T. Young [5] have shown that  spectral images offer more discriminating power than standard RGB images for measuring ripeness stages of tomatoes. An RGB color camera is frequently used for ripeness sorting. However, the results show that considerable errors may occur when classifying small differences in ripeness using RGB images. Spectral images are more suitable for classifying ripeness because they have a higher discriminating power compared to regular color images. Spectral images allow us to become independent of the light source and, after normalization, of object geometry as well. G. Polder, G.W.A.M. van der Heijden and I.T. Young showed that considerable errors can occur when classifying small differences in ripeness stage using RGB images. Spectral images are more suitable for classifying ripeness stage but these images are very large. Using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), the number of spectral bands is reduced while maintaining most of the relevant information as described by the variations. From both the research papers it can be depicted that classification of the tomato as crop is feasible. Also the papers show that ripeness of tomato can be a criterion of classification.

Hassan Asadollahi, Morteza Sabery Kamarposhty, Mir Majid Teymoori have presented in their paper the tomato classification by images processing. Data was used in the paper in related to 90 images of tomato[7]. They extracted 10 features from 90 images by machine vision technique. Then results were evaluated that accepted in considered images. Also they classified the data using different classification methods and then compared them. This paper shows us that to classify an image we can choose between several classifiers such as NaiveBayes, Multi Layer Perceptron, RBF Network, Neural Binary Tree, Random Tree, Random Forest, Instance Base K-

Nearest Neighbor , K-Star(K*). So any classifier can be used according to the need and required of the hardware and operating system.

Now since the need is to provide a handy solution to the farmer with the use of cellular phone it was needed to analyse the use and possibilities of classifying images on cellular phone. Classificaion of images on cellular phone is not very popular task till date though this field is emerging very quickly. Some work has been done in the research field related to image classification on cellular phone.

Giuseppe Amato, Paolo Bolettieri, Gabriele Costa, Francesco La Torre, Fabio Martinelli have presented a prototype for parental control that detects images with adult content received on a mobile device[2]. More specifically, the application that we developed is able to intercept images received through various communication channels (bluetooth, MMS) on mobile devices based on the Symbian operating systems. Once intercepted, the images are analysed by the component of the system that automatically classify images with explicit sexual content. At the current stage the application that intercepts images runs on the mobile device, the classifier runs on a remote server. This paper in general tells us that image classification is possible on cellular phone. The problem with this paper is that the image classification is done on the server side of the cellular phone. So the drawback with this idea is that in case there is any sort of connection failure between the client and server the classification result might not be available at the time of need[2]. So to have better results classification should be performed on the device itself.

Vijay Chadrasekhar, David M. Chen, Andy Lin, Gabriel Takacs, Sam S. Tsai, Ngai-Man Cheang, Yuriy Reznik, Radek Greszezuk, Bernd Girod evaluate the performance of MPEG-7 image signatures, Compressed Histogram of gradients descriptor (CHoG) and Scale Invariant Feature Transform (SIFT) descriptor for mobile visual search applications[8]. It was observed that SIFT and CHoG outperform MPEG-7 image signatures greatly in terms of feature-level Receiver Operating Characteristics (ROC) performance and image level matching. This paper gives us the knowledge about image classification on cellular phones and also gives on basics of image classification process. It shows that image classification follows two different steps before the actual classification. Prior to image classification the image has to first go through feature extraction process then to the feature compression process then the output of third step i.e. feature compression process is fed to the classifier for image classification. The output of classification is then compared to the test image to get the desired output.

## III. PROBLEM DEFINITION:

India is an agricultural country. The advancement of technology can be seen everywhere in the entire country. Life of common people is being more and more dependent on the gazettes. Various gazettes available in the market are making day to day life of common people simpler, easier and comfortable. Now booking of tickets for train, buses and movies too do not require us to stand in long queues. All we need to do is use e-tickets. We do not need to send paper letters to our loved ones and keep them waiting for our letters to be delivered. We have the fantastic facility of e-mails. Not only that we now have facilities of small light weight devices which can be carried by us anywhere we want for facilities like e-mails, e-tickets, information gathering through internet etc like laptops, notebooks, tablets, mobile phone and wireless internet facilities. We do not need to go to a telephone kept at a particular location in order to connect to people. We have hand held device i.e. the cellular phone which can be with us everywhere we go. Also it is worth mentioning here that our cellular phones or the mobile phones have become more than a device to connect to the people near and far than us. Our cellular phones have become a device to let us take digital pictures, make videos, create text and information and store them as desired for information or for memories.

All this description is made in order to show how much day to life of common people is being influenced by the various electronic gazettes. But these electronic gazettes are of not much use to a common farmer in this agricultural country. The problem can now be seen in this way is that amongst several decisions that a farmer has to make regarding his crop for example quality of seed which he is going to sow, health of the plant which he is growing, type of the fertilizer he is going to use in his field and over his crop, he also has to decide the harvesting time of his crop. With the advancement of science and biotechnology good seeds and appropriate fertilizer have provided farmers with some help. But when it comes to deciding the harvesting time he is still dependent on the age old method of experience and knowledge provided by the elders. There is no facility with the help of a gazette to provide help to the farmers for this.

Now since mobile phones or the cellular phones are a must possession to almost everybody now days, it can be said that a farmer too has a cellular phone. Now since it is a handy device it can be thought of giving him a solution to his harvesting problem on the cellular phone itself. Most of the mobile phones that are available in the market have good inbuilt digital cameras. So he can take picture of his crop any time he would like. Now taking tomatoes as the example of research it can be explained in this way. When a farmer has to decide that whether a particular tomato is ready for harvesting all he will have to do is to click a photograph of that particular tomato with the digital camera present in his mobile phone. Once the picture is taken a processing will be done on the cellular phone and a result will be provided stating whether a tomato is ready for harvesting or not. During the processing the image taken by the farmer will be compared by the already existing picture which is ready for harvesting and decision is made likewise. In this way this very popular electronic device can be used to provide him with a handy solution.

## IV. PROPOSED SOLUTION:

When the concept of digital images came into picture people wanted to click digital pictures and store them for their memories and other usages. Facilities were provided to them with the upcoming ever progressing digital cameras. Therefore, digital cameras became an essential device to grasp images, store them and retrieve them. Slowly digital cameras started being an inseparable part of cellular phone or the mobile phones too. With the advancement of time people now wanted to not only use images for memories but for knowledge too. So a lot of research started in the field of image retrieval and text based image retrieval [9]. Then with the advancement of research more and more knowledge was required from the images which let to use data mining techniques. When data mining techniques where used on images a new field of data mining emerged known as Image mining. Image mining concerns the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the images. It is more than just an extension of data mining to image domain. Image mining is an interdisciplinary endeavor which draws upon expertise in computer vision, image understanding, data mining, machine learning, database, and artificial intelligence [9].

When providing handy solution to farmers for deciding the time of harvesting by using cellular phone is considered. Image mining on mobile phones is a new and exciting field with many challenges due to limited hardware and connectivity. Phones with cameras, powerful CPUs, and memory storage devices are becoming increasingly common. Therefore, image mining on mobile phone is becoming an ambitious field. As our requirement says we will be in need of image mining to fulfill our requirement of comparison of pictures of tomatoes.

The block Diagram of the proposed solution of the problem are given below as shown in Figure 1:

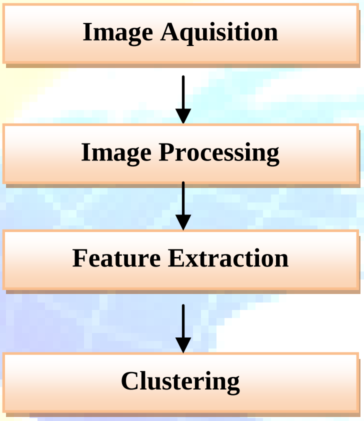

Figure1. Block Diagram of the Proposed Solution

A. Image Acquisition

Image Acquisition refers to the capturing of image data by a particular sensor or data repository such as a digital camera. In the work presented here the images are taken from a farmhouse in Misrod, Bhopal, Madhya Pradesh. The variety of tomato taken for our experiment is Laxmi 5005 sown in the farmhouse in Misrod.

B. Preprocessing

Image mining deals with large collection of image datasets that are high-dimensional and have multiple features, so time and space cost are relative high when analysis them. In order to improve quality and efficiency of the following mining steps, it is vital to discover suitable preprocessing technologies to clean up the un-related data and make useful hidden information more obvious. Traditional image processing technologies are applied to the image data ready to be mined.

C. Feature extraction

One of the key problems is how to express image data, which can usually be represented by features such as texture, color, edge, shape. According to the mining object, extract the basic elements that can present the images, omit features in essential to mining result. In some cases, to get better mining result, it is necessary to converge many features to form multidimensional feature vectors. Color, edge, texture are very important features in image mining and are widely used.

In case of our problem the various features that are considered are average contrast, edge density, histogram and standard deviation. Let us now explain each one of them one by one.

i. Average Contrast

Contrast is the difference in luminance and/or colour that makes an object (or its representation in an image or display) distinguishable. In visual perception of the real world, contrast is determined by the difference in the colour and brightness of the object and other objects within the same field of view. Because the human visual system is more sensitive to contrast than absolute luminance, we can perceive the world similarly regardless of the huge changes in illumination over the day or from place to place. The maximum contrast of an image is the contrast ratio or dynamic range.

Average contrast

= the average intensity within the object relative to

its surroundings

= Saliency / Area

The advantage of average contrast is that it is more 'robust' than Contrast. Adding a single white pixel to a light object may dramatically increase its contrast, but not its average contrast.

ii. Edge density

A window is a fixed-size rectangular region of the image. For a given window, an edge density feature measures the average edge magnitude in a subregion of the window[10]. Let *i*(*x*, *y*) be a window and *e*(*x*, *y*) be the edge magnitude of the window. For a subregion *r* with the left-top corner at (*x*1, *y*1) and the right-bottom corner at (*x*2, *y*2), the edge density feature is defined as

$$f = \frac{1}{a_r} \sum_{x=x_1}^{x_2} \sum_{y=y_1}^{y_2} e(x, y)$$

where *ar* is the region area, *ar* = (*x*2 −*x*1 +1)(*y*2 −*y*1 +1).

iii. Standard Deviation

For standard deviation approach feature vector of 6 components is formed by calculating standard deviation of row and column vector of R, G, and B planes of an image. Then by using the similarity measure that is Euclidean Distance the distance between the query image and database images is calculated. Finally determination of threshold is done so that similar images are retrieved.

1. Split image into R, G, and B components.

2. Mean of each row is calculated and a sequence is formed. Similarly a column sequence is also formed.

3. Calculate the variance for all six components obtained

above using:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2$$

where is $x_i$ data vector and *x* is mean given by:

$$\frac{1}{N}\sum_{1}^{N} x_i$$

As this value tends to be very large its square root to get standard deviation is taken.

In this way we obtained the standard deviation for following data vectors Redrow, Redcol, Greenrow, Greencol, Bluerow and Bluecol to form feature vectors of six components as:

V= [V1 V2 V3 V4 V5 V6]

4. Application of similarity measure ‘Euclidean Distance’:

$$D_{QI} = \sqrt{\sum_{i=1}^{6}(FQ_i - FI_i)^2}$$

Where *FQ*is feature vector for query image and *FI* is feature vector for database image.

5. Select those images where the distances are less than threshold value T.

iv. Histogram of RGB

Each pixel in an image has a color which has been produced by some combination of the primary colors red, green, and blue (RGB). Each of these colors can have a brightness value ranging from 0 to 255 for a digital image with a bit depth of 8-bits. A RGB histogram results when the computer scans through each of these RGB brightness values and counts how many are at each level from 0 through 255. The region where most of the brightness values are present is called the "tonal range." Tonal range can vary drastically from image to image, so developing an intuition for how numbers map to actual brightness values is often critical. There is no one "ideal histogram" which all images should try to mimic; histograms should merely be representative of the tonal range in the scene

D.Clustering

The process of grouping a set of images into classes of similar images without prior knowledge is called image clustering[9]. It is an unsupervised learning method, images within a cluster have high similarity in comparison to one another but are very dissimilar to images in other clusters. The process normally comprises of 4 steps: (1) Image representation, feature extraction and selection[11]; (2) Set up similarity metrics suitable for special application; (3) Image clustering; (4) Form clusters as shown in Figure 2. After clustering, field experts are required to examine each cluster and label it with abstract concepts. Nowadays there are many clustering algorithms such as: partitioning methods, hierarchical methods, grid-based methods, model-based methods, etc.

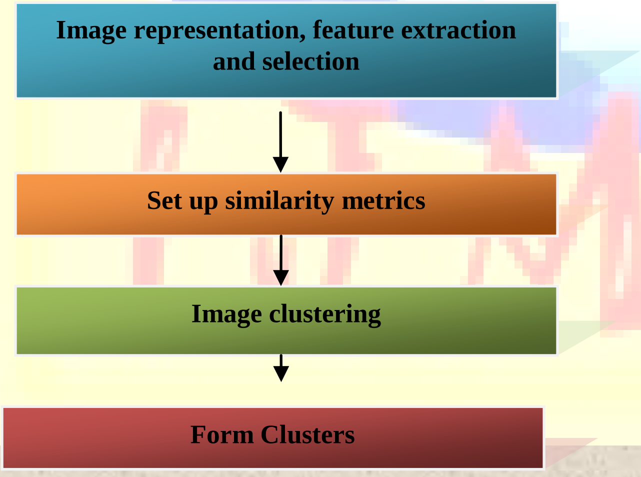
Image representation, feature extraction and selection
Set up similarity metrics
Image clustering
Form Clusters

Figure 2. Steps of Clustering

It is seen is that there will be a need of some cellular phone with an operating system that would allow having the desired solution. There are several operating systems available with the cellular phones of various companies like Symbian, Windows Mobile, Andriod etc. Now some operating systems are closed source while others are open source. Android operating system is an open source operating system. Android being an open source operating system we prefer it to show the utility of the work done by us. Android is an open source, Linux-derived OS backed by Google.

## V. EXPERIMENTAL ENVIRONMENT:

Since the solution to the problem requires a cellular phone as limited device, the work will be done on Android based Cellular phone. So, first requirement of the implementation is Android Operating system. Therefore, implementation will be done on Android 2.0. The environment used for development will be Java SDK.

## VI. DATASET USED:

The dataset is collected of images captured from a farmhouse in Misrod, Bhopal, Madhya Pradesh which will be used for the experiment. The images are of tomatoes. The variety of tomato is Laxmi 5005 sown in the farmhouse of Misrod.

## VII. CONCLUSIONS:

There are several features that can be considered for image clustering. For creating classifiers using clusters in case of limited device the features that is considered here are average contrast, standard deviation, edge density and histograms.Thus considering these features an efficient classifier can be constructed for classifying crops using limited devicebsuch as cellular phone. Thus it will be used as efficient classifier on cellular phone to let a farmer know when his crop is ready for harvesting. This technical support can be provided to farmers for ease of harvesting.

## REFERENCES:

- Margaret H. Dunham," Data mining Introductory and Advanced Topics", pp 10-13, pp 119-159,7 Edition 2011, Pearson.
- Giuseppe Amato, Paolo Bolettieri, Gabriele Costa, Francesco La Torre, Fabio Martinelli, "Detection of images with adult content for parental control on mobile devices", In Proceedings of the 6th International Conference on Mobile Technology Application Systems Mobility 09 (2009), ISBN 9781605585369:1-5.
- Lokesh Setia, Alexandra Teynor, Alaa Halawani and Hans Burkhardt, " Image Classification using Clustering-Cooccurrence Matrices of Local Relational Features", published in MIR 06 October 26-27,2006, Santa Barbara, California,USA.
- Ed Brunette, Hello Android Introducing Google's Mobile Development Platform, 3 Edition, August 4 2010, The Pragmatic Programmers, ISBN-10: 1934356565
- G. Polder, G. W. A. M. van der Heijden, I. T. Young, " Spectral Image Analysis for Measuring ripeness of Tomatoes", published in American Society of Agricultural Engineers (2002), Vol. 45(4): 1155-1161, ISSN 0001–2351
- G. Polder, G.W.A.M. van der Heijden and I.T. Young, "Hyperspectral Image Analysis for Measuring Ripeness of Tomatoes", published in 2000 ASAE International Meeting, ASAE Paper No. 003089 July 9-12, 2000, Midwest Express Center Milwaukee, Wisconsin
- Hassan Asadollahi, Morteza Sabery Kamarposhty, Mir Majid Teymoori ,"Classification and Evaluation of Tomato Images Using Several Classifier", published in 2009 International Association of Computer Science and Information Technology - Spring Conference, 978-0-7695-3653-8/09 © 2009 IEEE DOI 10.1109/IACSIT-SC.2009.47
- Vijay Chandrasekhar, David M. Chen, Andy Lin, Gabriel Takacs, Sam S. Tsai, Ngai-Man Cheung, Yuriy Reznik, Radek Grzeszczuk, Bernd Girod," Comparison Of Local Feature Descriptors For Mobile Visual Search", published in 17th IEEE International Conference on Image Processing (ICIP) 2010, DOI: 10.1109/ICIP.2010.5649937

- Hu Min, Yang Shuangyuan, “Overview of Image Mining Research”, published in The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010, DOI 978-1-4244-6005-2/10/ ©2010 IEEE
- Phung, SL & Bouzerdoum,” Detecting People in Images: An EdgeDensity Approach”, IEEE International Conference on Acoustics, Speech and Signal Processing, 2007 (ICASSP 2007), Honolulu, Hawaii, USA, 15-20 April 2007, 1, I-1229-I-1232. Copyright 2007 IEEE.
- H. B. Kekre, Kavita Patil,” Standard Deviation of Mean and Variance of Rows and Columns of Images for CBIR”, published in International Journal of Computer and Information Engineering Volume 3:Issue1 2009 ISSN 2010-393X.