Leukemia Prediction from Microscopic Images of Human Blood Cell Using HOG Feature Descriptor and Logistic Regression

Submitted By

Hossain Abedy

ID: 2014-1-60-080

Faysal Ahmed

ID: 2014-1-60-108

Supervised By

Md.Shamsujjoha

Senior Lecturer

Department of Computer Science & Engineering

East West University

A Research Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING EAST WEST UNIVERSITY

August 2018

ABSTRACT

Leukemia is a cancer of blood which originates in bone marrow causing disruption in the production of human blood cells. Earlier detection of leukemia is crucial due to its fatality. Detection of leukemia involves microscopic observation of human blood cells. Application of various image processing methods along with existing machine learning algorithm is a fast and convenient way to detect leukemia. These methods require extracting features from microscopic images of blood cells and applies machine learning algorithm to train and test a classifier based model which can predict leukemia with an acceptable accuracy. However, scarcity of publicly available image dataset and the inconsistency of the information provided from them make it more challenging while developing a model which can predict leukemia accurately. Besides the size of small datasets and the computational cost, memory evaluation and the required accuracy are also concerning issue. Keeping these drawbacks in mind, we proposed an efficient classifier based model which extract features from image dataset with and classify them accordingly. In this book, we have introduced an edge feature with HOG feature descriptor and Logistic Regression Classifier based model which can detect leukemia with an accuracy of almost 96%.

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Md. Shamsujjoha, Senior Lecturer & Assistant Proctor, Department of Computer Science and Engineering, East West University**. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma. Any material reproduced in this project has been properly acknowledged.

Submitted by:

_____

Hossain Abedy
ID: 2014-1-60-080
Department of Computer Science and Engineering
East West University

_____

Faysal Ahmed
ID: 2014-1-60-108
Department of Computer Science and Engineering
East West University

_____

Supervised by:

_____

Md. Shamsujjoha
Senior Lecturer
Department of Computer Science and Engineering
East West University
Bangladesh

LETTER OF ACCEPTANCE

This research book entitled **"Leukemia Prediction from Microscopic Images of Human Blood Cell Using HOG Feature Descriptor and Logistic Regression"** submitted by Hossain Abdey (2014-1-60-080) and Faysal Ahmed (2014-1-60-052) to the Department of Computer Science and Engineering, East West University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 10th August, 2017.

Supervisor

Md. Shamsujjoha
Senior Lecturer
Department of Computer Science and Engineering,
East West University, Dhaka, Bangladesh

Chairperson

Dr. Md. Mozammel Huq Azad Khan
Chairperson and Professor
Department of Computer Science and Engineering,
East West University, Dhaka, Bangladesh

ACKNOWLEDGEMENT

We would like to express our gratitude to our parents. Their diligent support and perpetual inspiration towards study since the early stage of our education, has placed us at the edge of our graduation degree. I believe that whatever we have achieved and whatever we are going to gain are owing to our parents.

We would like to pay homage to our supervisor Md. Shamsujjoha, His cordial directions have kept us on the right track from the very first day of supervision. Whenever, we came up with complicated issues, he guided us the simple way to resolve the issues. Besides, we are grateful to all our course directors for providing us with contemporary insights from the field of system development and implementation.

Our special thanks to all our friends, colleagues for their continuous inspiration and guidelines throughout my study period in East West University.

Moreover, we heartily thank to my family members for their financial supports for our study. Without their support, our study in this university could have been a dream, nothing more. We are profoundly grateful to our Creator that we have been in touch with and guided by such great individuals in the world.

TABLEOFCONTENTS Page

Chapter Page

Chapter 1

Introduction

1.1 Overview

In this book we have introduced an efficient classifier based model which can detect leukemia infected blood cell with an accuracy of 96% with less computational cost. For that purpose we have analyzed previously worked methods for feature extraction and image segmentation techniques and discussed an efficient method which can detect blast and non-blast cell by detecting edges and shapes of each cells. The extracted features were computed into reduced dimensions and then applied machine learning algorithm such as logistic regression for classification.

1.2 Motivation

The main motivation of this book is to explore the concepts of automated leukemia detection using various image processing techniques and the application of various machine learning algorithm for the classification of leukemia infected cells. It is already making a revolutionary change in the diagnosis system of leukemia. Most of the traditional treatments were used to depends on the operator's skills to perform. If the implementation of computer aided diagnosis could be introduced all over, then patients will have a better chance of detecting leukemia in early stage and the proper diagnosis of leukemia could be made possible with lower cost, less time consuming and less resources required.

1.3 Objectives

The main objective is to develop a model which can perform classification and detect leukemia with a prediction accuracy of over 90%. Building an efficient models requires implementation of existing image processing techniques along with classifier. Here we will perform the following tasks.

- Explain what is leukemia and how it spreads.
- Explain how leukemia was treated and the introduction of CAD leukemia detection.
- Explain the algorithms and methods used in our model.
- Explain the implementation of our model and the results outcomes.

1.4 Contribution

In this book, we have discussed a method for classification of images containing blast and non-blast cell using HOG feature descriptor and Logistic Regression based classifier. First we have acquired a valid leukemia image dataset which consist 108 images which 59 non-blast cell and 49 blast cells. We have also applied image pre-processing techniques like canny edge detector for the blast shape, Gaussian Filter for removing unwanted noise, kernel operator sobel kernel for image filtering and PCA (Principal component analysis) for dimension reduction of the feature vector. We applied these methods on the ALL_IDB1 image dataset for training, testing and validating our proposed model. For the development we have used anaconda spider and python as codes structured language and we have also analyzed the performance of our model for different test runs on the dataset and got an average accuracy of almost 96% showing room for improvement.

Chapter 2

Background Study

2.1 What is Leukemia?

Acute Lymphocytic Leukemia, also known as Acute Lymphoblastic Leukemia is a life threatening hematic disease, which is the result of overproduction and multiplication of immature lymphocytes within white blood cells. It can life threatening if left undiagnosed due to its rapid spread into the bloodstream and other vital organs of human body. Leukemia is common in young children and adults over 50. Early detection and diagnosis of the leukemia is very crucial for the recovery of patients especially in the case of children. The symptoms of ALL are similar also in other disease and for this reason; the diagnosis is very difficult to conduct. One of the steps in the diagnostic procedures encompasses the microscope inspection of peripheral blood.
There are two types of acute leukaemia1, namely acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML).A typical blood microscope image is plotted in Fig. 1. The principal cells present in the peripheral blood are red blood

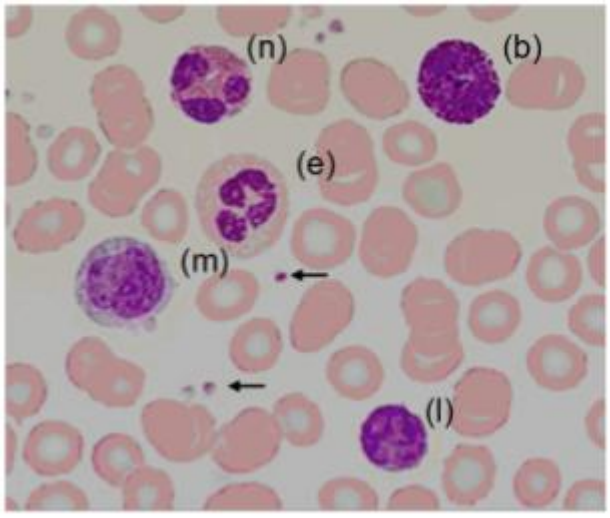


Fig. 1. Blood's white cells marked with colorant: basophil (b), eosinophil (e), lymphocyte (l), monocyte (m), and neutrophil (n). Arrows indicate platelets. Others elements are red cells.

cells, and the white cells (leucocytes). Leucocyte cells containing granules

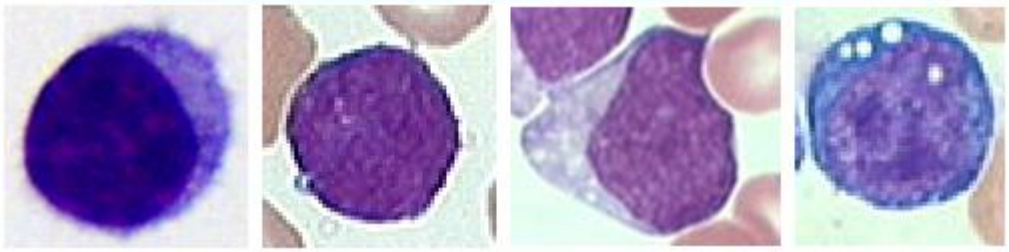
are called granulocytes (composed by neutrophil, basophil, eosinophil). Cells without granules are called agranulocytes (lymphocyte and monocyte). The percentage of leucocytes in human blood typically ranges between the following values: neutrophils 50-70%. eosinophils 1-5%. basophils 0-1%. monocytes 2-10%, lymphocytes 20.45% [16]. The ALL disease is related to the lymphocytesin the bone marrow and into the peripheral blood. The colorant used in the preparation of the blood tends to concentrate only in white cells, in particular in their nuclei that are typically center positioned (the darker elements in Fig. 1). In most of case, the white cells are also bigger than the red cells. The most common leukemia classification by morphological analysis is the FAB method [18], even if nowadays it has been updated with the immunologic classification [19], which it is not image-based. Differently than the FAB method (requiring only a microscope), the immunologic classification needs a more sophisticated setup for the procedure. Usually, an automatic method for the detection of lymphoblasts in microscopically color images can be divided in the sequent steps.

• Segmentation - the cells are separated from the background by using algorithms based on different characteristics of the cells (e.g. shape, color, inner intensity).

• Identification of white cells - the cells are classified in white cells and red cells. The classifiers can search the presence of the nucleus by using color information.

• Identification of lymphocytes - the lymphocytes can be distinguished from the other white cells by analyzing the shape of the nucleus (e.g., a deeply staining nucleus which may be eccentric in location, and a small amount of cytoplasm).

• Identification of candidate lymphoblasts - candidate lymphoblasts can be identified in a set of lymphocytes by the analysis of morphological deformations of the cell.



(a) (b) (c)

(d)

Fig.2. Morphological variability associated to the blast cells according to the FAB classification: (a) healthy lymphocytes cell from non-ALL patients, (b-d) lymphoblasts from ALL patients where (b), (c) and (d) are L1, L2 and L3 respectively.

In particular, lymphocytespresent a regular shape, and a compact nucleus with regular and continuousedges. Instead, lymphoblasts present shape irregularities. Concerning the ALL, the candidate lymphoblasts are analyzed by using the FAB classification as follows.

• L1 - Blasts are small and homogeneous. The nuclei are round and regular with little clefting and inconspicuous nucleoli. Cytoplasm is scanty and usually without vacuoles.

• L2 - blasts are large and heterogeneous. The nuclei are irregular and often clefted. One or more, usually large nucleoli are present. The volume of cytoplasm is variable, but often abundant and may contain vacuoles.

• L3 - blasts are moderate-large in size and homogeneous. The nuclei are regular and round-oval in shape. One or more prominent nucleoli are present. The volume of cytoplasm is moderate and contains prominent vacuoles.

Fig. 2 shows the great variability in shape and pattern of the blast cells according to the FAB classification.

2.2 Typical Leukemia Diagnosis

The main treatment for acute lymphocytic leukemia (ALL) in adults involves the long- term use of chemotherapy (chemo). [20] In the past several years, doctors have begun to use more intensive chemo regimens, which have led to more responses to treatment. But these regimens are also more likely to cause side effects, such as low white blood cell counts. Patients may need to take other drugs to help prevent or treat these side effects.
Typically, leukemia is detected by analyzing microscopic inspection of white blood cells and its components, which performs two main analyses: cell classification and counting of blasts. Acute leukemia is usually diagnosed by a morphological analysis of blood slides by hematologists', which is a complex, time-consuming, and costly process [21]. It also requires considerable training and experience. Furthermore, the results often lack of a standardized performance owing to a variety of factors including insufficient expertise or imperfection of the samples. Some digital diagnosis systems were developed to analysis microscopic blood images for leukemia detection. However, they suffered from a number of limitations; in particular accurate diagnosis of leukemia requires discrimination of one cell type from another, and of cell nucleus from cell cytoplasm. Indeed, separation of leukemia cell nucleus with diverse complex irregular morphology from cytoplasm is a challenging task. Besides this manual approach of detecting leukemia is time consuming and requires experts' supervision, which is also costly and tiring. In most cases, results are subjective and imprecise as the whole process relies on operator's skills which can be subjected for unwanted errors.

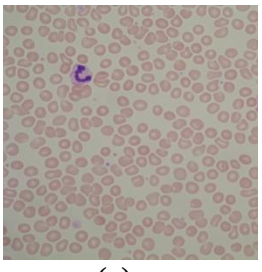
2.3 Automated Approaches to Detect Leukemia

Since manual diagnostic methods are time-consuming, less accurate, and prone to errors due to various human factors like stress, fatigue, and so forth; the introduction of automated approaches to detect leukemia is now a popular choice as the methods extract cellular information from the images of a human blood cell and diagnosis can be applied according to cell's condition. Computer aided detection and diagnosis of leukemia overcomes much of the drawbacks which previously were a major issue for the treatment of leukemia. Digital histopathology has witnessed a lot of improvement in the recent years. With the new technological advancement, much effective methods have been proposed for automated microscopic image analysis. Due to this development, computer-aided detection (CAD) is becoming a reliable method for ALL detection.[22]

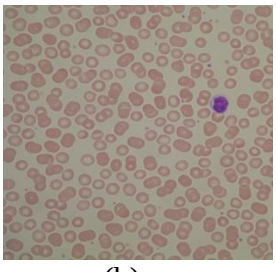
CAD system for ALL detection can be divided into four phases, namely, preprocessing, segmentation, feature extraction, and classification. This section provides a detailed survey of different techniques and methods that have been proposed, developed, and used in the automatic detection of acute lymphocytic leukemia cell.

2.4 ALL_IDB1 image dataset

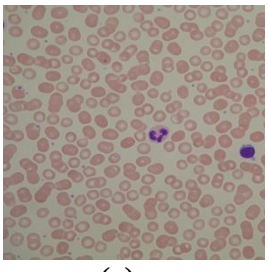
The ALL_IDB1 version 1.0 can be used both for testing segmentation capability of algorithms, as well as the classification systems and image preprocessing methods. This dataset is composed of 108 images collected during September, 2005. It contains about 39000 blood elements, where the lymphocytes have been labeled by expert oncologists. The images are taken with different magnifications of the microscope ranging from 300 to 500.



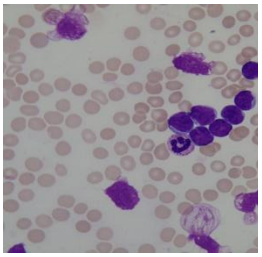
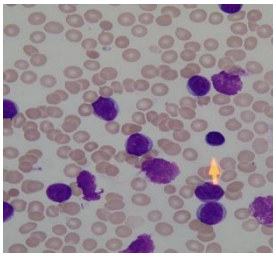
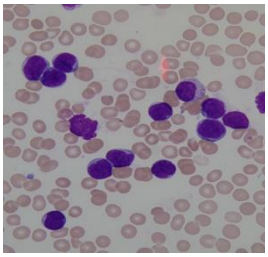
(a)



(b)



(c)

(e) (f) (g)

Fig.3.Non-blast cells (a),(b),(c) and Blast cells (e) (f) (g)

The annotation of ALL-IDB1 is as follows. The ALL-IDB1 image files are named with the notation ImXXX_Y.jpg where XXX is a 3-digit integer counter and Y is a boolean digit equal to 0 is no blast cells are present, and equal to 1 if at least one blast cell is present in the image. Please note that all images labeled with Y=0 are from for healthy individuals, and all images labeled with Y=1 are from ALL patients. Each image file ImXXX_Y.jpg is associated with a text file ImXXX_Y.xyc reporting the coordinates of the centroids of the blast cells,

2.5 Related Works

Recently, a good number of works have put efforts to detect leukemia from the microscopic image. A low-pass filter has been employed by Scotty[1] in order to remove noise from the background and, after that, white blood cells are segmented with different threshold operations and image clustering. Piuri's[2] approach is based on detecting edges for white blood cells segmentation. Otsu proposed a threshold selection method[3]. Cseke [4] proposed a fast segmentation scheme with automatic thresholding where thresholds are selected with a simple recursive method derived from maximizing the interclass variance between dark, gray, and bright regions based on the method proposed by Otsu[3]. Pattern recognition methods are also used which could be further categorized as supervised [5] and unsupervised [6]. Supervised methods includes Support Vector Machine and Artificial Neural Network and unsupervised clustering mainly includes K-mean clustering [7-9], fuzzy C-means [10] etc. B.C.Ko [11] introduced step wise merging rule on mean shift clustering and boundary removal rules with a gradient vector flow (GVF) snake for the segmentation of white blood cells. Transformed color space such as HSV [12,13] and HIS [14] was also introduced by various researchers for image segmentation.

Chapter3

Method Description

3.1 Overview

In the previous chapters, we discussed the nature and the symptoms of leukemia and the diagnosis process of leukemia infected patients. We also discussed the drawbacks of the typical diagnosis process and the effectiveness of computer aided diagnosis models. An automated approach of detecting leukemia is a combination of various image processing techniques and application of classifier based models. There are steps for making an image ready for feature extraction, then extract features from an image, removing unwanted information for reducing computational cost then train a classifier that can predict leukemia infection with an acceptable accuracy.In this chapter we will briefly discuss the images pre-processing and post processing methods and the classifier which we applied in our work.

3.2 Methods used in the proposed model

- ROI for selecting the region of cells
- Canny Edge detector for detecting the shape and edges of cells
- HOG feature descriptor for feature extraction
- PCA for reducing the dimensions of the extracted feature
- Logistic regression classifier for classification

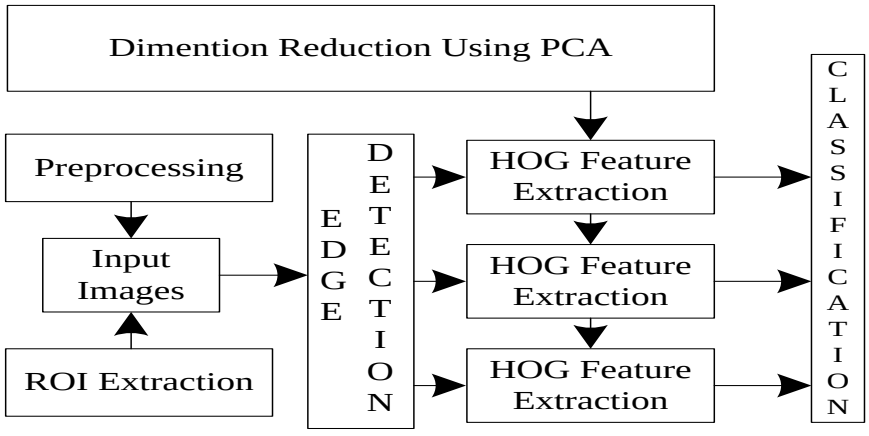


Fig.4.Propossoed Model

3.2.1 Region Of Interest (ROI)

A region of interest (often abbreviated ROI), are samples within a data set identified for a particular purpose.[24] The concept of a ROI is commonly used in many application areas. For example, in medical imaging, the boundaries of a tumor may be defined on an image or in a volume, for the purpose of measuring its size. The endocardial border may be defined on an image, perhaps during different phases of the cardiac cycle, for example, end-systole and end-diastole, for the purpose of assessing cardiac function. In geographical information systems (GIS), a ROI can be taken literally as a polygonal selection from a 2D map. In computer vision and optical character recognition, the ROI defines the borders of an object under consideration. In many applications, symbolic (textual) labels are added to a ROI, to describe its content in a compact manner. Within a ROI may lie individual points of interest (POIs).

3.2.2 Canny edge detector

Canny edge detection is a technique to extract useful structural information from different vision objects and dramatically reduce the amount of data to be processed. [25] It has been widely applied in various computer vision systems. Canny has found that the requirements for the application of edge detection on diverse vision systems are relatively similar. Thus, an edge detection solution to address these requirements can be implemented in a wide range of situations. The general criteria for edge detection include:

Detection of edge with low error rate, which means that the detection should accurately catch as many edges shown in the image as possible
The edge point detected from the operator should accurately localize on the center of the edge.
A given edge in the image should only be marked once, and where possible, image noise should not create false edges.
To satisfy these requirements Canny used the calculus of variations – a technique which finds the function which optimizes a given functional. The optimal function in Canny's detector is described by the sum of four exponential terms, but it can be approximated by the first derivative of a Gaussian.

Among the edge detection methods developed so far, Canny edge detection algorithm is one of the most strictly defined methods that provides good and reliable detection. Owing to its optimality to meet with the three criteria for edge detection and the simplicity of process for

implementation, it became one of the most popular algorithms for edge detection.

The Process of Canny edge detection algorithm can be broken down to 5 different steps:

- Apply Gaussian filter to smooth the image in order to remove the noise
- Find the intensity gradients of the image
- Apply non-maximum suppression to get rid of spurious response to edge detection
- Apply double threshold to determine potential edges
- Track edge by hysteresis: Finalize the detection of edges by suppressing all the other edges that are weak and not connected to strong edges.

3.2.3 HOG feature descriptor

Theory: The essential thought behind the histogram of oriented gradients descriptor is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled. The descriptor is the concatenation of these histograms. For improved accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination and shadowing.[26]

The HOG descriptor has a few key advantages over other descriptors. Since it operates on local cells, it is invariant to geometric and photometric transformations, except for object orientation. Such changes would only appear in larger spatial regions. Moreover, coarse spatial sampling, fine orientation sampling, and strong local photometric normalization permits the individual body movement of pedestrians to be ignored so long as they maintain a roughly upright position. The HOG descriptor is thus particularly suited for human detection in images.

Algorithm implementation:

Gradient computation: The first step of calculation in many feature detectors in image pre-processing is to ensure normalized color and gamma values. However, this step can be omitted in [26] HOG descriptor computation, as the ensuing descriptor normalization essentially achieves the same result. Image pre-processing thus provides little impact on performance. Instead, the first step of calculation is the computation of the gradient values. The most common method is to apply the 1-D centered,

point discrete derivative mask in one or both of the horizontal and vertical directions. Specifically, this method requires filtering the color or intensity data of the image with the following filter kernels:

$[-1,0,1]$ and $[-1,0,1]^T$

Other more complex masks, such as the 3x3 Sobel mask or diagonal masks, but these masks generally performed more poorly in detecting humans in images. They also experimented with Gaussian smoothing before applying the derivative mask, but similarly found that omission of any smoothing performed better in practice.

Orientation binning: The second step of calculation is creating the cell histograms. Each pixel within the cell casts a weighted vote for an orientation-based histogram channel based on the values found in the gradient computation. The cells themselves can either be rectangular or radial in shape, and the histogram channels are evenly spread over 0 to 180 degrees or 0 to 360 degrees, depending on whether the gradient is "unsigned" or "signed". It has been found that unsigned gradients used in conjunction with 9 histogram channels performed best in their human detection experiments. As for the vote weight, pixel contribution can either be the gradient magnitude itself, or some function of the magnitude. In tests, the gradient magnitude itself generally produces the best results. Other options for the vote weight could include the square root or square of the gradient magnitude, or some clipped version of the magnitude.

Descriptor blocks: To account for changes in illumination and contrast, the gradient strengths must be locally normalized, which requires grouping the cells together into larger, spatially connected blocks. The HOG descriptor is then the concatenated vector of the components of the normalized cell histograms from all of the block regions. These blocks typically overlap, meaning that each cell contributes more than once to the final descriptor. Two main block geometries exist: rectangular R-HOG blocks and circular C-HOG blocks. R-HOG blocks are generally square grids, represented by three parameters: the number of cells per block, the number of pixels per cell, and the number of channels per cell histogram. In the human detection experiment, the optimal parameters were found to be four 8x8 pixels cells per block (16x16 pixels per block) with 9 histogram channels. Moreover, they found that some minor improvement in performance could be gained by applying a Gaussian spatial window within each block before tabulating histogram votes in order to weight pixels around the edge of the blocks less. The R-HOG blocks appear quite similar to the scale-invariant feature transform (SIFT) descriptors; however, despite their similar formation, R-HOG blocks are computed in dense grids at some single scale without orientation alignment, whereas SIFT descriptors are usually computed at sparse, scale-invariant key image points and are rotated to align orientation. In addition, the R-HOG blocks are used in conjunction to encode spatial form information, while SIFT descriptors are used singly.

Circular HOG blocks (C-HOG) can be found in two variants: those with a single, central cell and those with an angularly divided central cell. In addition, these C-HOG blocks can be described with four parameters: the number of angular and radial bins, the radius of the center bin, and the expansion factor for the radius of additional radial bins. Dalal and Triggs found that the two main variants provided equal performance, and that two radial bins with four angular bins, a center radius of 4 pixels, and an expansion factor of 2 provided the best performance in their experimentation(to achieve a good performance, at last use this configure). Also, Gaussian weighting provided no benefit when used in conjunction with the C-HOG blocks. C-HOG blocks appear similar to shape context descriptors, but differ strongly in that C-HOG blocks contain cells with several orientation channels, while shape contexts only make use of a single edge presence count in their formulation.

Block normalization: Let v be the non-normalized vector containing all histograms in a given block, ||v|| be its k-norm for k=1,2 and e be some small constant (the exact value, hopefully, is unimportant). Then the normalization factor can be one of the following:

L2-norm: $f = \frac{v}{\sqrt{||v||2+e^2}}$

L2-hys: L2-norm followed by clipping (limiting the maximum values of v to 0.2) and renormalizing,

L1-norm: $f = \frac{v}{||v||2+e^2}$

L1-sqrt: $f = \sqrt{\frac{v}{||v||2+e^2}}$

In addition, the scheme L2-hys can be computed by first taking the L2-norm, clipping the result, and then renormalizing. In the experiments, it has been found the L2-hys, L2-norm, and L1-sqrt schemes provide similar performance, while the L1-norm provides slightly less reliable performance; however, all four methods showed very significant improvement over the non-normalized data.

Object recognition: HOG descriptors may be used for object recognition by providing them as features to a machine learning algorithm. HOG descriptors used as features in a support vector machine (SVM); however, HOG descriptors are not tied to a specific machine learning algorithm.

3.2.4 PCA

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables [clarification needed] into a set of values of linearly uncorrelated variables called principal components.[27] If there are n observations with p variables, then the number of distinct principal components is min(n-1,p). This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors [clarification needed] are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

PCA was invented in 1901 by Karl Pearson,[1] as an analogue of the principal axis theorem in mechanics; it was later independently developed and named by Harold Hotelling in the 1930s.[2] Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, proper orthogonal decomposition (POD) in mechanical engineering, singular value decomposition (SVD) of X (Golub and Van Loan, 1983), eigenvalue decomposition (EVD) of XTX in linear algebra, factor analysis (for a discussion of the differences between PCA and factor analysis see Ch. 7 of Jolliffe's Principal Component Analysis[3]), Eckart–Young theorem (Harman, 1960), or empirical orthogonal functions (EOF) in meteorological science, empirical eigenfunction decomposition (Sirovich, 1987), empirical component analysis (Lorenz, 1956), quasiharmonic modes (Brooks et al., 1988), spectral decomposition in noise and vibration, and empirical modal analysis in structural dynamics.

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It's often used to visualize genetic distance and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering[clarification needed] (and normalizing or using Z-scores) the data matrix for each attribute.[4] The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point),

and loadings (the weight by which each standardized original variable should be multiplied to get the component score).

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection of this object when viewed from its most informative viewpoint [citation needed]. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

PCA is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.

PCA is also related to canonical correlation analysis (CCA). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.

3.2.5 Logistic regression

In statistics, the logistic model (or logit model) is a statistical model that is usually taken to apply to a binary dependent variable. In regression analysis, logistic regression or logit regression is estimating the parameters of a logistic model. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables. The two possible dependent variable values are often labeled as "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. The binary logistic regression model can be generalized to more than two levels of the dependent variable: categorical outputs with more than two values are modeled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model.[28]

Logistic regression was developed by statistician David Cox in 1958. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables

(features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other. The coefficients are generally not computed by a closed-form expression, unlike linear least squares; see § Model fitting.

Chapter 4

Methodology

The accuracy of the classifier is significantly depends on the extraction of features from the images. The images of the dataset consist of blood components such as white blood cells and others. Since leukemia infection occurs in white blood cells so our Region Of Interest (ROI) should be the region covered by white blood cells.

4.1Feature Detection

There are many popular algorithms to detect the feature of an image such as edges, corners etc. In this work, we have used the edge feature of the image. To detect the edges of images, the canny edge detection algorithm [16] was used. Canny edge detection is a popular and multistage edge detection algorithm, which works as follows:

•Since edges are affected by noise, the image is smoothed by a Gaussian Filter to reduce noise.

•Then the smoothed image is filtered with kernel operator such as sobel kernel in both horizontal and vertical direction to get first derivative in horizontal (G1) and vertical direction (G2). From G1 and G2, the edge gradient and the direction of each pixel can be calculated as follows,

$$G = \sqrt{G1^2 + G2^2}$$

$$\theta = atan2(G1, G2)$$

•After getting gradient magnitude and direction, a full scan of the image is made to remove the unwanted pixels which may not constitute the edge. For this, at every pixel is checked whether it is a local maximum in its neighborhood in the direction of the gradient.

•This stage decides which are all edges are really edges and which are not. For this, we need two threshold values, minimum value and maximum value. Any edges with intensity gradient more than minimum value are sure to be edges and those below minimum value are sure to be non-edges, so discarded. Those who lie between these two thresholds are classified edges or non-edges based on their connectivity. If they are connected to "sure-edge" pixels, they are considered to be part of edges. This stage also removes small pixels' noises on the assumption that edges are long lines. So what we finally get is strong edges in the image.

4.2 Feature Description

To describe the edge features obtained by applying canny edge detection algorithm, we have used Histogram of Oriented Gradients (HOG) feature descriptor. HOG feature descriptor is mainly used for object detection in computer vision and machine learning. However, HOG can also be used as feature descriptor for representing both shape and texture.

The following steps are executed for calculating HOG features.

- Image normalizing prior to description.
- Computation of gradients in both the x and y directions.
- Obtaining weighted votes in spatial and orientation cells.
- Contrast normalizing overlapping spatial cells.
- Collecting all Histograms of Oriented gradients to form the final feature vector.

Feature Dimension Reduction: Dimensionality reduction is the process of reducing the dimension of a feature, provided that the lesser dimension has similar information like before. One of the most popular dimensionality reduction algorithms is Principle Component Analysis (PCA) [Ref]. PCA projects a higher dimensional data to a lower dimension without much loss of information. The following steps are included while performing PCA.

- Compute covariance matrix.
- Compute eigenvalue and eigenvectors from covariance matrix.
- Select K largest eigenvalues where K is the dimension of the new subspace.
- Compute projection matrix from K selected eigenvalues.
- Transform the dataset through projection matrix to form a new dataset of K dimension.

4.3 Classification

For classification purpose, we used one of the most popular classification algorithms Logistic Regression. Logistic regression is the go-to method for binary classification problem. It works for the logistic function. Logistic function also called sigmoid functions an S-shaped curve that takes any real valued input and maps it into a value between 0 and 1.

Here X is the input and Y is the output and e is the base of natural logarithm.

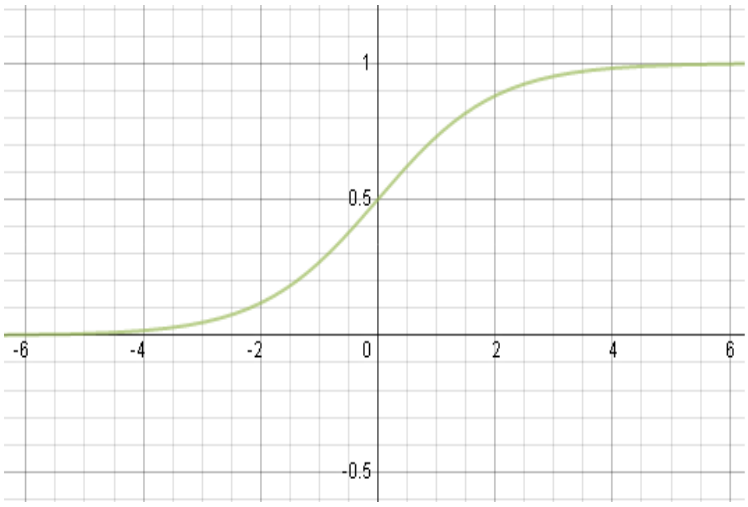


Fig.5. Sigmoid Function

PROPOSED MODEL

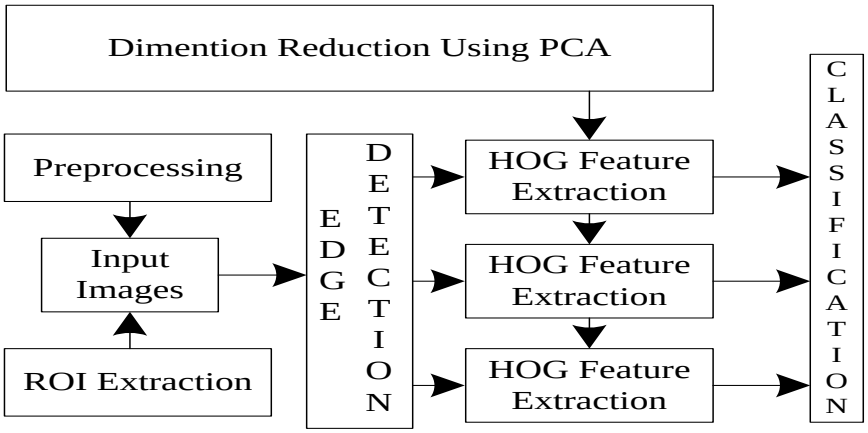
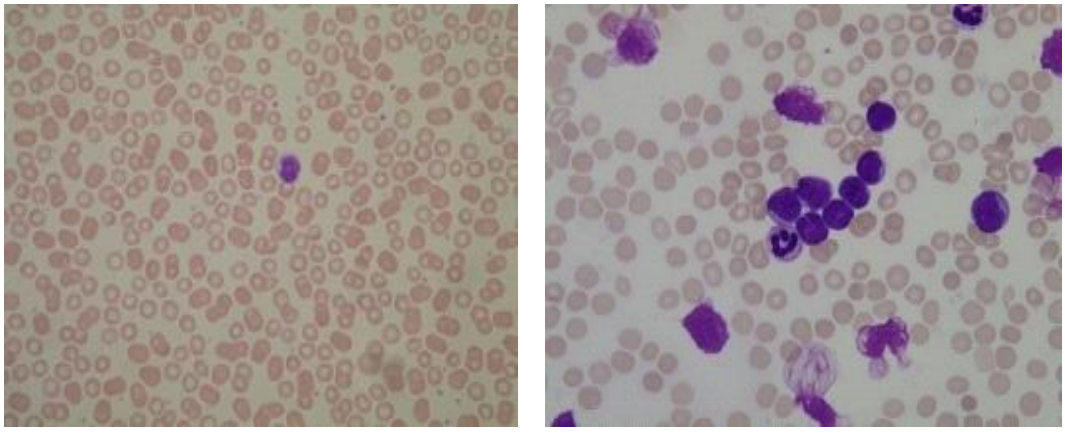


Fig.6. Proposed Model

In our proposed model, firstly edges are detected form the image by using canny edge detection algorithm. After that, the edged image is passed through the HOG feature descriptor and a one dimensional feature vector is collected for every image. Then the dimension of feature vector is reduced by using Principle Component Analysis. The last step includes the classification of two types of images using Logistic Regression classifier.

Dataset: The image dataset consist of 108 images, which has 59 non-blast cell images and 49 blast cell images. All images are in JPG format with 24-bit color depth and resolution of 2592 x 1944.

The task associated with this dataset is to automate the classification of the image set into blast and non-blast. For this purpose, we have divided the whole image dataset into two sets - train set (80% of original dataset) and test set (20% of original dataset).



(a) (b)

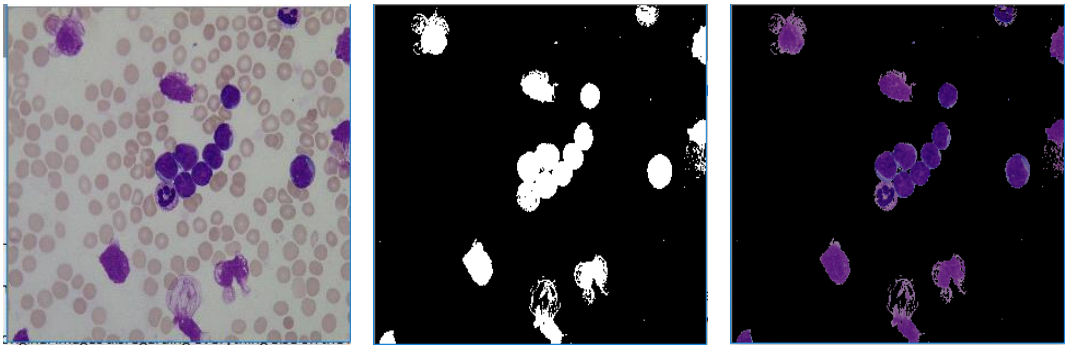
Fig.7. Non-blast Cells(a),Blast Cell(b)

Chapter 5

Experiment & Results

5.1 Stage 1:-Preprocessing and ROI extraction

Image preprocessing is an important part for any image classification task. Since the image dataset contains images of different height and width, we have resized the images to a fixed size of 1000 X 1000pixels. Next, the image is converted into HSV color space and filter the purple color which is the ROI is extracted from the image. At last, the images are converted to gray scale image. Description relating with the figures 4 Original Image (a), ROI Extracted Image in Grayscale(b) ROI extracted images in RGB.



(a) (b) (c)

Fig.8. Original Image(a), ROI Extracted Image in Grayscale(b) ROI extracted images in RGB

5.2 Stage 2: -Canny Edge Detection

The images from step 1 are then passed through canny edge detection algorithm. This produces an image of containing the edges.

Fig.9. Image with canny edges

5.3 Stage 3:-HOG Feature Extraction

Then the HOG feature is calculated from the image produced in step 2. The dimension of the HOG feature vector is 352836.

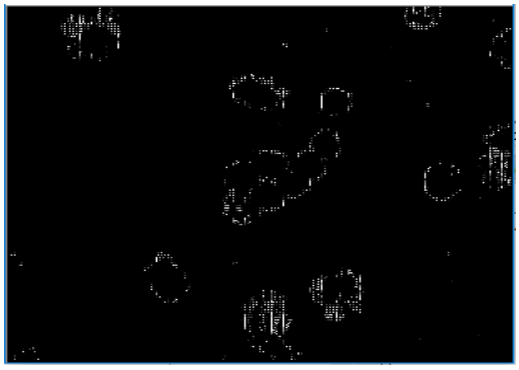


Fig.10. HOG image

5.4 Stage 4: -Dimension reduction by PCA

The dimension of the feature vector is reduced by PCA with an explained variance ratio of 0.95. The reduced dimension of feature vector for 5 run is

Table1: Dimension of the feature vector for five different runs

Run no	***Dimension of Feature***
•	51
•	52
•	53
•	52
•	52

5.5 Stage 5:-Classification by Logistic Regression

Logistic Regression classifier is trained on training set and prediction is made on test set

Table 2: Results of Logistic Regression model for five runs

Validation Accuracy (3-Fold)	Test Accuracy	Confusion Matrix	Precision	Recall
94	91	9 2 0 11	81.82	100
95	100	10 0 0 12	100	100
97	91	8 2 0 12	80	100
94	95	11 1 0 10	91.67	100
98.8	100	12 0 0 10	100	100

From the above table, it can be clearly states that, average cross validation accuracy is almost 96% and the average test accuracy is almost 96%.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this paper we have discussed a method for classification of images containing blast and non- blast cell using HOG feature extractor and Logistic Regression based classifier. We have also applied image pre-processing techniques like canny edge detector for the blast shape, Gaussian Filter for removing unwanted noise, kernel operator sobel kernel for image filtering and PCA (Principal component analysis) for dimension reduction of the feature vector. We applied these methods on the ALL_IDB1 image dataset which contains 108 images of blast and non-blast cells for training, testing validating our proposed model. We have also analyzed the performance of our model for different runs on the dataset and got an average accuracy of almost 96% showing room for improvement.

6.2 Future Work

During this project, every step of building this game was fairly successful. The game-

REFERENCES

[1] Scotti.F,"Automatic morphologic analysis for acute leukemia identification in peripheral blood microscope images," in Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications(DOI : 10.1109/CIMSA.2005.1522835)

[2] Piuri.V and Scotti.F, "Morphological classification of blood leucocytes by microscope images," in Proceedings of the IEEE InternationalConferenceonComputationalIntelligenceforMeasurement Systems and Applications(DOI: 10.1109/CIMSA.2004.1397242)

[3] Otsu.N, "A threshold selection method from gray-level histograms,"Automatica,vol.11,no.285–296,pp.23–27,1975.(DOI: 10.1109/TSMC.1979.4310076)

[4] Cseke.I, "A fast segmentation scheme for white blood cell images,"inProceedingsofthe11thIAPRInternationalConference on Pattern Recognition C:Image, Speech and Signal Analysis,vol. 3,pp.530–533,IEEE,1992.(DOI: 10.1109/ICPR.1992.202041)

[5] Guo.N.N, Zeng.L.B,and Wu.Q.S, "A method based on multispectral imaging technique for white bloodcell segmentation," Computers in Biology and Medicine, vol. 37, no. 1, pp. 70–76, 2007.

[6] Saraswat.M and Arya.K.V, "Automated microscopic image analysis for leukocytes identification:asurvey,"Micron,vol.65, pp.20–33,2014.

[7] Zhang.C,Xiao.X, Lietal.X. ,"White blood cell segmentation bycolor-space-basedk-meansclustering,"Sensors,vol.14,no.9, pp.16128–16147,2014.

[8] Mohapatra.S and Patra.D, "Automated leukemia detection using hausdorff dimension in blood microscopic images," in Proceedings of the International Conference on IEEE Robotics and Communication Technologies (INTERACT '10), pp. 64–68, Chennai,India,December2010.(DOI: 10.1109/INTERACT.2010.5706196)

[9] Salem.N.M, "Segmentation of white blood cells from microscopic images using K-means clustering," in Proceedings of the 31st National Radio Science Conference (NRSC'14),pp.371–376, Cairo,Egypt,April2014.(DOI: 10.1109/NRSC.2014.6835098)

[10] Mondal.P.K,Prodhan.U.K,AlMamunetal M.S,"Segmentation of white blood cells using fuzzy C means segmentation algorithm,"IOSRJornalofComputerEngineering,vol.1,no.16, pp.1–5,2014.

[11] Ko.B.C, Gim J.W, and Nam J.Y,"Automatic white blood cell segmentation using stepwise merging rules and gradient vector flowsnake,"Micron,vol.42,no.7,pp.695–705,2011.

[12] Eldahshan K.A,Youssef M.I, Masameer E.H, and Mustafa.M.A, "Segmentation framework on digital microscope images for acute lymphoblastic leukemia diagnosis based on HSV Color Space," International Journal of Computer Applications,vol.90,no.7,pp.48–51,2014.

[13] ElDahshan K.A, Youssef M.I, Masameer E.H, and Hassan M.A, "Comparison of segmentation framework on digital microscope images for acute lymphoblastic leukemia diagnosis using RGB and HSV color spaces," Journal of Biomedical EngineeringandMedicalImaging,vol.2,no.2,pp.26–34,2015.

[14] Singhal V and Singh P," Co-relation based feature selection for diagnosis of acute lymphoblastic leukemia," in Proceedings of the3rdACMInternationalSymposiumonWomeninComputing and Informatics(WCI'15),pp.5–9,Kochi,India,August2015.

[15] 15.Fatichah C., Tangel M. L., Widyanto M. R., Dong F., Hirota K. Interest-based ordering for fuzzy morphology on white blood cell image segmentation. Journal of Advanced Computational Intelligence and Intelligent Informatics. 2012;16(1):76–86.

[16] W. Rong, Z. Li, W. Zhang and L. Sun, "An improved Canny edge detection algorithm," 2014 IEEE International Conference on Mechatronics and Automation, Tianjin, 2014, pp. 577-582.doi: 10.1109/ICMA.2014.6885761

[17] Labati, Ruggero & Piuri, Vincenzo & Scotti, Fabio. (2011). ALL-IDB: the acute lymphoblastic leukemia image DataBase for image processing. Proceedings / ICIP ... International Conference on Image Processing. 2045-2048. 10.1109/ICIP.2011.6115881.

[18] J. M. Bennett, D. Catovsky, Marie-Theregse Daniel, G. Flandrin, D. A. G. Galton, H. R. Gralnick, and C. Sultan, "Proposals for the classification of the acute leukaemias french-american-british (fab) co-operative group," British Journal of Haematology, vol. 33, no. 4, pp. 451–458, 1976.

[19] A. Biondi, G. Cimino, R. Pieters, and Ching-Hon Pui, "Biological and therapeutic aspects of infant leukemia," Blood, vol. 96, no. 1, pp. 24–33, July 2000.

[20] Amercian Cancer Society https://www.cancer.org/cancer/acute-lymphocytic-leukemia/treating/typical-treatment.html Accessed in August 2 2018

[21] Chin Neoh, Siew,Srisukkham, Worawut,Zhang, Li,Todryk, Stephen,Greystoke, Brigit,Peng Lim, Chee,Alamgir Hossain, Mohammed,Aslam, Nauman,"An Intelligent Decision Support System for Leukaemia Diagnosis using Microscopic Blood Images"Scientific Reports,2015/10/09/online,http://dx.doi.org/10.1038/srep14938.

[22] Computational and Mathematical Methods in Medicine,Volume 2018, Article ID 6125289, 13 pageshttps://doi.org/10.1155/2018/6125289

[23] ALL-IDB https://homes.di.unimi.it/scotti/all/ Accessed in March 2 2018

[24] Ron Brinkmann (1999). The Art and Science of Digital Compositing. Morgan Kaufmann. p. 184. ISBN 978-0-12-133960-9.

[25] Wikipedia contributors. (2018, May 23). Canny edge detector. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:47, August 2, 2018, from https://en.wikipedia.org/w/index.php?title=Canny_edge_detector&oldid=842631206

[26] Wikipedia contributors. (2018, February 27). Histogram of oriented gradients. In Wikipedia, The Free Encyclopedia. Retrieved 19:53, August 2, 2018, from https://en.wikipedia.org/w/index.php?title=Histogram_of_oriented_gradients&oldid=827980546

[27] Wikipedia contributors. (2018, July 31). Principal component analysis. In *Wikipedia, The Free Encyclopedia*. Retrieved 20:12, August 2, 2018, from https://en.wikipedia.org/w/index.php?title=Principal_component_analysis&oldid=852860855

[28] Wikipedia contributors. (2018, July 10). Logistic regression. In *Wikipedia, The Free Encyclopedia*. Retrieved 20:14, August 2, 2018, from https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=849707526