Study of Colour Model for Segmenting Mycobacterium Tuberculosis in Sputum Images

A Kurniawardhani, R Kurniawan, I Muhimmah, and S Kusumadewi
Informatics Department, Islamic University of Indonesia, Yogyakarta, Indonesia.

Abstract. One of method to diagnose Tuberculosis (TB) disease is sputum test. The presence and number of Mycobacterium tuberculosis (MTB) in sputum are identified. The presence of MTB can be seen under light microscope. Before investigating through stained light microscope, the sputum samples are stained using Ziehl–Neelsen (ZN) stain technique. Because there is no standard procedure in staining, the appearance of sputum samples may vary either in background colour or contrast level. It increases the difficulty in segmentation stage of automatic MTB identification. Thus, this study investigated the colour models to look for colour channels of colour model that can segment MTB well in different stained conditions. The colour models will be investigated are each channel in RGB, HSV, CIELAB, YCbCr, and C-Y colour model and the clustering algorithm used is k-Means. The sputum image dataset used in this study is obtained from community health clinic in a district in Indonesia. The size of each image was set to 1600x1200 pixels which is having variation in number of MTB, background colour, and contrast level. The experiment result indicates that in all image conditions, blue, hue, Cr, and Ry colour channel can be used to segment MTB in one cluster well.

1. Introduction
Tuberculosis (TB) is included the top 10 disease that cause death. In 2015, there were 1.8 million people in the world died from TB, in other words, there are over 4,900 people died every day [1]. Indonesia is one of the top three high burden country for TB in the world. In 2015, there were 1.02 million TB incidence in Indonesia [2]. These large number of TB suffers can be caused of TB, one of communicable disease, is spread easily through splashes of sputum in the air that contain Mycobacterium tuberculosis (MTB), while the sufferer cough, spit, or sneeze. Moreover, to identify that a person has TB takes a long time. For decreasing the number of TB suffers, especially in low- and middle-income countries, affordable, reliable, and fast enough instrument is necessary.

More than 95% of deaths caused by TB occur in low- and middle-income countries [1]. Diagnostic test commonly used in low- and middle-income countries is examination of sputum samples [1]. Sputum samples were examined through a microscope to see the presence of MTB, name of the bacteria that causes TB. Because inexpensive and easy to maintain, the light field microscope is used commonly in low- and middle-income countries. Before sputum samples are seen under light microscope, the sputum sample is stained using Ziehl–Neelsen (ZN) stain technique.

To develop affordable, reliable, and fast enough instrument, a lot of studies are conducted. There are many studies that start by developing automatic MTB detection in sputum image using light field microscope and ZN staining [3-12]. Those studies were attempting for assisting pathologists in assessing sputum sample to diagnosis TB. The studies have major challenges, namely, sputum images have inhomogeneity intensity of background, number of MTB, and contrast between one image with another images, as shown in Figure 1. When the ZN staining process is done properly, MTB will appear in red colour while other organisms and background will in blue colour [12], and the samples will have adequate contrast, as shown in Figure 1.a. Improper staining procedure and reagent preparation can
produce under-stained or over-stained images [13-14], as shown in Figure 1.b and Figure 1.c, respectively. Under-stained images make images have low contrast, while over-stained images make images have red background and dark red, red, or bright red MTB.

Figure 1. The sputum images. (a) Normal images. (b) Under-stained image. (c) Over-stained image.

The staining issue as previously mentioned increase the difficulty in MTB segmentation process. To overcome that problem, different colour models are considered used in segmentation process. In [3-8], segmentation process is done in RGB (Red Green Blue) colour model. RGB colour model that is adopted from physiology of the human eye is renowned colour model which is ubiquitous used in digital instrument. It, however, cannot mimic the human colour visual system appropriately. In those studies, MTB can be segment well because most of sputum images used in those studies have pretty fine stained. In [9-10], segmentation process is done in HSV (Hue Saturation Value) colour model, which segmenting MTB based on the degree value in Hue channel. HSV that is adopted from how choose colour for paint or ink can defined colour in similar way with how humans perceive colour. Nevertheless, it overlooks the detail of colour appearance. In [11], segmentation process is done in CIELAB (Commission Internationale de l'Eclairage Lab) colour model. CIELAB which is ‘device-independent’ has gamut wider than human vision. In [12], segmentation process is done not only in CIELAB (Commission Internationale de l'Eclairage Lab) colour model, but also combined it in YCbCr colour space. Meanwhile, [13,14] do segmentation process in RGB and C-Y colour model. Because of many variation colour model used in segmentation proses, this study will investigate the appropriate colour model used in segmentation process that can apply in different stained conditions. The colour models will be investigated are each channel in RGB, HSV, CIELAB, YCbCr, and C-Y colour model.

2. Image Dataset

Figure 2. The examples of sputum image database.
Sputum microscopic images of Tuberculosis (TB) sufferer are collected from three community health clinic (CHC) (called Puskemas in Indonesia) in Temanggung Regency, Central Java, Indonesia. The three CHC are Bansari, Ngadirejo, and Traji Puskesmas. Temanggung Regency was chosen because that area has a fairly high TB patient so that has many digital collections of sputum microscopic images of TB patients. From those three CHC, the 234 sputum images of TB patients were obtained. All images contained Mycobacterium tuberculosis (MTB) with varying amounts in each image. Each image has different colouring and lighting conditions against to another images. The sputum image is *. TIF format and 24-bit colour image. The size of each image was set to 1600x1200 pixels. The examples of sputum images used are shown in Figure 2. Those sputum image database with specification of Indonesian people can be accessed by public for research and education purpose at http://psimed-zinsmid.informatics.ui.ac.id/.

3. Colour Model

Colour model is numerical value representation of colour to model visible spectrum of colour. Meanwhile colour space is a variant of colour space that specifies a gamut (range) of colours [16].

3.1. RGB (Red Green Blue)

The Red Green Blue (RGB) colour model is commonly colour model used in digital instruments, such as scanner, camera, digital TV, computer, and mobile phone. Red Green Blue means that it colour model have three primary colours, namely, red, green, and blue. To create another colour besides of the three basic colours is done by combining three basic colours in accordance with the desired colour composition. The principle used in RGB is adopted from physiology of the human eye [15].

3.2. HSV (Hue Saturation Value)

The Hue Saturation Value (HSV) colour model is adapted from how people choose colour for paint or ink. Because of that, it claimed that HSV can defined colour in similar way with how humans perceive colour. Hue describes the colour or tint information, saturation describes how white the colour is, and value describes how dark the colour is [15]. To get HSV colour, RGB image can be convert to HSV using equation (1-3).

\[
H = \cos^{-1} \left( \frac{1}{2} \left( \frac{(R - G) + (R - B)}{(R - G)^2 + (R - B) + (G - B)} \right) \right) \quad (1)
\]

\[
S = 1 - \frac{3}{(R + G + B)} \min(R,G,B) \quad (2)
\]

\[
V = \frac{R + G + B}{3} \quad (3)
\]

3.3. CIELAB (Commission Internationale de l'Eclairage Lab)

CIELAB consist of three-dimension colour representation, namely, \(L\) for luminance or lightness information, \(a\) for colour opponent information between red and green, and \(b\) for colour opponent information between blue and yellow.

3.4. YCbCr

The YCbCr colour space is commonly used for digital video. \(Y\) describes the luminance or lightness information, while \(Cb\) and \(Cr\) describes chrominance information. \(Cb\) represent the colour difference between blue-component and a reference value, while and \(Cr\) represent the colour difference between red component and a reference value. To get HSV colour, YCbCr image can be convert to YCbCr using equation 4).
3.5. C-Y colour
The C-Y consist of three-dimension colour representation $Y$, $R_y$, and $B_y$. $Y$ describes the luminance or lightness information, while $R_y$ and $B_y$ describes chrominance information. $Cb$ represent the colour difference between blue-component and a reference value, while and $Cr$ represent the colour difference between red component and a reference value. To get C-Y consist, RGB image can be convert to YCbCr using equation (5).

\[
\begin{bmatrix}
Y \\
R_y \\
B_y
\end{bmatrix} =
\begin{bmatrix}
0.299 & 0.587 & 0.114 \\
-0.1687 & -0.3313 & 0.5 \\
0.5 & -0.4187 & -0.0813
\end{bmatrix}
\begin{bmatrix}
\text{Red} \\
\text{Green} \\
\text{Blue}
\end{bmatrix}
\] (4)

4. Method

4.1. K-means Clustering
K-means clustering is one of well-known clustering method that has been widely used in various fields. K-means clustering is famous for its simplicity and reliability. K-means clustering divides the data into multiple clusters, with the number of clusters corresponding to the k value specified by the user [17]. The algorithm of K-means clustering is:
1. Determine the value of k
2. Determine the cluster centre coordinates as much as k, randomly
3. Calculate the distance of each data point with each cluster centre
4. Determine the membership of each data point against a cluster based on the nearest distance between the data points to one of the cluster centres.
5. Recalculate the cluster centre based on the average value of all data point of cluster members
6. Do step 3-5 continuously until the coordinates of cluster centre does not change.

4.2. Segmentation Process
The segmentation process aims to separate objects and backgrounds. Objects in the sputum image not only Mycobacterium tuberculosis (MTB), but also can be cell remnants, food scraps, fibres, or blood. This segmentation process will look for the colour model channel which can segment most of MTB against the background and all non-MTB objects. The scheme of the segmentation process can be seen in Figure 3.

Figure 3. The scheme of the segmentation process.
paper, namely, RGB, HSV, CIELAB, YCbCr, and C-Y colour model, the scheme that shows in Figure 3 will be repeated five times. Therefore, the 15 segmented images will be gotten after this step, namely, segmented image in Red, Green, Blue, Hue, Saturation, Value, L, a, b, Y, Cb, Cr, Y, Ry, and By colour channel.

The first stage of segmentation process is conversion. The original images gotten from local Community Health Clinic (CHC) are in 24-bit colour of RGB. To investigate those images in HSV, CIELAB, YCbCr, and C-Y colour model, the colour conversion is necessary. After the five images in the five colour models are obtained, each colour model images are separated based on their three colour model channels. Therefore, there are 15 colour channel images is derived from one sputum image.

Clustering process using k-Means algorithm is performance at each colour channel of 15 colour channel images. K-Means Clustering is used to cluster the pixels to be one cluster if the pixels have same colour and separate them to different cluster if the pixels have the different colour. Hence, a cluster will have object that have similar colour. The parameter of k for k-Means clustering used in this study is five, so that there are five clusters produced. Further, the cluster that contains most of MTB objects is selected for the result of segmentation process.

5. Result and Discussion

The experiments are conducted in accordance with stages that shown in Figure 3 and used sputum images that have variation in number of MTB, background colour, and illumination. Firstly, the best k-parameter of k-Means clustering used in this study is investigated. Secondly, how to select the cluster that obtained from k-Means clustering is considered to get the best segmented images from each colour channel. Thirdly, the segmented images are investigated to seek the best representative colour channel in MTB segmenting process. Hardware specifications used in this study are Intel® Core™ i5 Processors and 4 GB RAM. We run the experiment in Microsoft Windows 10 Professional x64. All experiment is performed using Matlab 2013.

For selecting k-parameter for k-Means clustering, the 15 images that have variation in number of MTB, background colour, and illumination are selected randomly. The value of k-parameter is varied from two up to seven. The experiment results define that the average of optimum value for k-parameter for 225 images, 15 images multiplied by 15 colour channel, is five. It means in an images the colour will grouped into five clusters. It is due to in an image there are various objects, such as fibber, cell remnants, food scraps, and blood. Although they have the same colour properties, but they have different colour gradation level because of improper staining process. MTB sometimes have the same colour with other objects around them or even similar to the one of gradation colour level of background. Thus, the level of gradation also needs to be grouped again.

For selecting the cluster that obtained from k-Means clustering, the 15 images that have variation in number of MTB, background colour, and illumination are selected randomly too. In order to get the best segmented images from each colour channel, an image in each cluster is investigated and compared to other images that belongs to remaining clusters. In most of experiment, most of MTB can be segmented in a cluster that have the fewest pixel number of members. When the ZN staining process is done properly, MTB will appear in red colour while other organisms and background will in blue colour. Moreover, MTB has tiny size about 0.3-0.6 μm in width and 1–4μm in length, which give a size of 70–140 pixels in the image [10,14]. In some image, however, MTB cannot be clustered alone because the red colour of MTB does not look conspicuous. It is due to the image has low contrast or MTB are blocked by another object. Furthermore, when the ZN staining process is done improperly, background can appear in bright red colour, which is almost the same colour as MTB.

The colour channel lists that can segment MTB well are shown in Table 1. In all image condition, blue, hue, Cr, and Ry colour channel can be used to segment MTB in one cluster well. It is due to blue colour channel have different intensity in MTB and background colour. In background colour, which background colour tends to white either in the red or blue background, blue intensity tends to be higher than it is in MTB colour. On the other side, red intensity tends to have similar intensity for MTB and background colour. MTB colour tends to in red, magenta, or purple colour, that have combination colour between red and blue. In some of images, however, the segmented image in blue colour channel contain another blue object that should be eliminated. Hue describes the colour or tint information. Thus, Hue
can segment the colour purity information which is not affected by brightness information. In most of images, however, the segmented image in hue colour channel contain a tiny object like noise that should be eliminated. Cr and Ry colour channel have the similar segmented result. In Cr and Ry colour channel, most of MTB can be grouped as one cluster but MTB is not separated from background. MTB colour properties in Cr and Ry colour channel tends to have similar colour properties of bright background, either in the red or blue background. The examples of segmented images are shown in Figure 4.

Table 1. List of colour channel that can segment MTB.

| Image | Background Colour | Contrast Level | Number of another object | Number of MTB | Colour Channel |
|-------|-------------------|----------------|--------------------------|----------------|----------------|
| 01.TIF | Blue             | Low            | ±20                      | 8              | Blue, hue, value, Cr, Ry, a, green |
| 03.TIF | Blue             | Low            | >70                      | ±50            | Cr, Ry, blue, value |
| 14.TIF | Bright blue      | Low            | ±20                      | 10             | Blue, green, hue, a, value, luminance, Y, Cr, Ry |
| 08.TIF | White            | High           | ±20                      | 13             | Hue, a, value, green, blue, Cr, Ry |
| 16.TIF | Magenta          | High           | ±10                      | ±15            | a, Cr, Ry, luminance, green, saturation, blue, value, hue |
| 07.TIF | Bright red       | High           | >50                      | 4              | Blue, a, saturation, hue, Cr, Ry, green |
| 17.TIF | Bright red       | Low            | ±20                      | 18             | a, Cr, Ry, saturation, blue, green, hue |
| 18.TIF | Bright red       | Low            | ±20                      | >70            | a, hue, Cr, Ry, blue, green |

The 03.TIF is little bit difficult to segment because it has many pretty big dark blue objects and tiny MTB. Many MTB in that picture are blocked by pretty big dark blue objects. Hence, further study is necessary in order to address some issues, such as image enhancement before doing segmentation because most of the images have low contrast.

Figure 4. The examples of segmented images.

6. Conclusion
This study investigated the colour models to look for colour channels of colour model that can segment MTB well in different stained conditions. The colour channels investigated are each channel in RGB, HSV, CIELAB, YCbCr, and C-Y colour model. K-Means clustering algorithm is used to segment the colour information. The sputum image dataset used in this study has variation in number of MTB, background colour, and contrast level. The experiment result indicates that in all image conditions, blue, hue, Cr, and Ry colour channel can be used to segment MTB in one cluster well.

References
[1] World Health Organization (WHO) 2017 Fact Sheets–Tuberculosis [Online] Available: http://www.who.int/mediacentre/factsheets/fs104/en/
[2] World Health Organization (WHO) 2017 Global tuberculosis report 2016 [Online] Available: http://www.who.int/tb/publications/global_report/en/
[3] Costa M G F, et al 2008 Automatic identification of mycobacterium tuberculosis with conventional light microscopy Eng. in Medicine and Biology Society (EMBS) pp 382-5.
[4] Sadaphal P, Rao J, Comstock G W, and Beg M F 2008 Image processing techniques for identifying Mycobacterium tuberculosis in Ziehl-Neelsen stains The Int. J. of Tuberculosis
and Lung Disease, 12(5), p 579-582.

[5] Khutlang R, Krishnan S, Whitelaw A, and Douglas T S 2009 Detection of tuberculosis in sputum smear images using two one-class classifiers Biomedical Imaging: From Nano to Macro 2009 (ISBI’09) IEEE International Symposium on pp 1007-10.

[6] Khutlang R, Krishnan S, Dendere R, Whitelaw A, Veropoulos K, Learmonth G, and Douglas T S 2010 Classification of Mycobacterium tuberculosis in images of ZN-stained sputum smears IEEE Transactions on Information Technol. in Biomedicine 14(4) pp 949-57.

[7] Mukti J and Kale K V 2011 Analysis of zn-stained sputum smear enhanced images for identification of mycobacterium tuberculosis bacilli cells Analysis 23(5).

[8] Raof R A A, Mashor M Y, Ahmad R B, and Noor S S M 2011 Image segmentation of Ziehl-Neelsen sputum slide images for tubercle bacilli detection Image Segmentation InTech.

[9] Xu C, Zhou D, Guan T, and Liu Y 2014 A segmentation algorithm for mycobacterium tuberculosis images based on automatic-marker watershed transform Robotics and Biomimetics (ROBIO) IEEE International Conference on pp 94-8.

[10] Makkapati V, Agrawal R, and Acharya R 2009 Segmentation and classification of tuberculosis bacilli from ZN-stained sputum smear images Automation Sci. and Eng. IEEE Int. Conf. on pp 217-20

[11] Rulaningtyas R, Sukmono A B, Mengko T, and Saptawati P 2015 Multi patch approach in K-means clustering method for color image segmentation in pulmonary tuberculosis identification Instrumentation, Communications, Information Technol., and Biomedical Eng. (ICICI-BME) pp 75-8.

[12] Sotaquira M, Rueda L, and Narvaez R 2009 Detection and quantification of bacilli and clusters present in sputum smear samples: a novel algorithm for pulmonary tuberculosis diagnosis Digital Image Processing pp 117-21.

[13] Osman M K, Mashor M Y, and Jaafar H 2010 Detection of mycobacterium tuberculosis in Ziehl-Neelsen stained tissue images using Zernike moments and hybrid multilayered perceptron network Systems Man and Cybernetics (SMC) pp 4049-55.

[14] Osman M K, Mashor M Y, Saad Z, and Jaafar H 2010 Colour image segmentation of tuberculosis bacilli in Ziehl-Neelsen-stained tissue images using moving k-mean clustering procedure Mathematical/Analytical Modelling and Computer Simulation (AMS) pp 215-20.

[15] Gonzalez R C and Woods R E 2007 Digital Image Processing 3 ed (Prentice Hall).

[16] Lal S 2007 Photoshop CS (New Delhi: Laxmi Publication Pvt. Ltd).

[17] MacQueen J June 1967 Some methods for classification and analysis of multivariate observations Proc. of the fifth Berkeley Symp. on Math. Statistics and Probability 1(14) p 281-97.

Acknowledgments
Authors wishing to acknowledge to DRPM Director General of Ministry of Research, Technology and Higher Education (Kemenristek Dikti) who has provided financial support through Prime Development of Higher Education research scheme in 2017 so that the research of "Prototype of TB Detection System based on Sputum Image" can be conducted.