Asian Journal of Information Technology 14 (4): 147-153, 2015
ISSN: 1682-3915

Segmentation of Cervical Image Using Unsupervised Clustering Algorithms with L*u*v Color Transformation

[1]Anantha Sivaprakasam Sivaprakasam and [2]Naganathan Ealai Rengasari
[1]Department of MCA, P.S.R. Engineering College, Sivakasi, India
[2]Department of Computer Science and Engineering, Hindustan University, Chennai, India

Abstract: This study deals with the fast and efficient algorithm for segmenting the cervical image using unsupervised methods. Here, the image segmentation method is based on basic region growing method. In this study, both fast k-means with optional weighting and careful initialization and Fuzzy c-means Clustering algorithm are used to deal with segmentation of the cervical image. Both algorithm segments the image in accordance with the colour for each cluster and its neighborhood. In this method, first the cervical image is smoothed, enhanced and converted into L*u*v color space. The L*u*v color space image is segmented using fast k-means algorithms with optional weighting and careful seeding and fuzzy c-means algorithm. Finally, the performance analysis of the three segmentation algorithms is carried out. Experimental results show that fast k-means segmentation with careful seeding methods are fast as compare to Fast K-means with weight and fuzzy c-means method, three algorithm gives better segmented images with finer details and accurate location but FCM takes more time.

Key words: Fast K-means, fuzzy c-mean, L*u*v color transformation, image segmentation, clustering

INTRODUCTION

Image segmentation is an important technology for image processing. It is used in many applications, especially in medical applications. Segmentation refers to the operation of partitioning an image into component parts or into separate objects. In the medical field, cervical cancer is a preventable disease if it is detected at the initial stage and it can be easily detected by a routine screening test. Automate Pap smear screening interacting with the human technologist can be a good solution to reduce the errors in screening slides. However, the automation of the process is challenging task because of processing the huge amount of data to be processor. Detecting the abnormal cell in a Pap smear can be a tedious problem. The number of methods was developed to detect the abnormal cell in both gray scale and color Pap smear image. In general, the quality of the image is affected by the performance of the colour image segmentation. The colour based image segmentation is carried out using K-mean algorithm (Chitade and Katiyar, 2010). In this only color features were considered and it is applicable for mapping the change in land use. Using the colour features the Google map is segmented using k-means algorithm (Hegadi and Sangaoli, 2011). In above said two algorithms, the colours are not frequently used in image segmentation. The natural image is segmented using the colour and texture features (Kothainachiar and Banu, 2007). This method is computationally not efficient at the color feature extraction. Mahanta *et al.* (2012) presents an approach for analysis of Pap smear images of cervical region based on cell nuclei distribution and shape and size analysis. Here color features are not taken. In medical field, the automatic segmentation of cervical cell is needed for pathologist. Both k-means and Fuzzy c-means algorithm are used for segmenting the satellite image using the colour features but not applicable for medical image (Singha and Hemachandran, 2011). A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis is presented. In this study, K-means algorithm and Fuzzy C-mean algorithm were used and comparison were carried out. Madhuri and Sandeep (2012) described perceptual color image segmentation through K-means. Higher time complexity of the k-means techniques is a drawback. This technique uses only the color feature for segmentation and spatial correlation is not used this work is in progress to reduce the time complexity and includes spatial correlation through the image of deformable models.

In this study, fast and efficient segmentation algorithm is used. In this, Fast k-means algorithm with weight and careful seeding and Fuzzy c-means algorithm are used to segment the cervical image. Then, performance analysis of all methods is carried out to find out the best algorithm.

Corresponding Author: Anantha Sivaprakasam Sivaprakasam, Department of MCA, P.S.R. Engineering College, Sivakasi, India

MATERIALS AND METHODS

This is fast and efficient method to segment the cervical image using fast k-means with weight and careful initialization and Fuzzy c-means algorithm with defuzzification. In this, first the image is smoothed using the mean filter, enhanced using de-correlation stretching techniques for enhancing the colour separation and regions are grouped into a set of five classes using Fast k-means algorithm with weight and careful initialization and Fuzzy c-means algorithm with defuzzification. Then, each pixel is labeled and creates the images that segment the image by colour and segment the nuclei into a separate image. Finally, the result of the both algorithm is analyzed for performance analysis.

Smoothing: Normally an image is often corrupted by noise in its acquisition and transmission. Some time, due to poor intensity, illumination and other factors, the cervical image may be corrupted by more than one noise. To remove the noise from image, smoothing operation is carried out. In general, mean or average filter (Patidar *et al*., 2010) comes under the linear filter category that removes certain types of noises. For example, an averaging filter is used to remove grain noise from photography. It is a simple sliding window spatial filter. This filter replaces the center value in the window with the average of all pixels values in the window. The window or kernel is normally square but can be of any shape. The example of mean filtering of a single 3×3 window of value is given below. Unfiltered window values:

5	10	15
20	3	25
30	35	40

Average = 5+10+15+20+3+25+30+35+40 = 183/9 = 20.33 = rounded to 20. Mean filtered:

*	*	*
*	20	*
*	*	*

De-correlation: De-correlation stretching (Chitade and Katiyar, 2010) is an image processing algorithm which is used to enhance or stretch the color separation. De-correlation stretch means "Principal components can be stretched and transformed back into RGB colors. The resultant image improves visual interpretation and makes feature discrimination easier. It can be based on the co-variance matrix or the correlation matrix. The number of colour bands in the image is taken three. To implement the de-correlation, decorrstretch function is used in Matlab.

L*u*v color space: In color imagery, the image pixel can be represented in a different number of color spaces such as RGB, HSV, HSI, etc. The CIELAB color space provides perceptually uniform space such as L*u*v or L*a*b. That means the Euclidean distance between two color points in the human vision system. L*u*v* (Priya *et al*., 2011) system is an excellent decoupler of intensity (represented by lightness L*) and color (represented by u* for red minus green and v* for green minus blue).

Fast k-means algorithm: This is Fast k-means algorithm with optional weighting and careful initialization method. This method partitions the vectors X into K clusters by applying the well known batch k-means algorithm. Rows of X correspond to points, columns correspond to variable. The output matrix C contains the cluster centroids. The k-element output column vector D contains the residual cluster distortions as measured by total squared distance of cluster members from the centroid.

The main concept of k means algorithm is to calculate the distance between the data point and the centers using the following equation:

$$d(x_k, x_c) = [(x_k - x_{ci})^T (x_k - x_{ci})^T]^{1/2} \quad (1)$$

where, d is the distance between the data point x_k at the cluster k and the initial centers are x_{ci}. The points in one cluster are defined as: x_l for l = 1, 2, n considered as one cluster and n is the total number of points in that cluster. The x_{ci} is chosen randomly either from the dataset or arbitrarily. In this method, we have used the random selection of the centers from the dataset to avoid wasting one more calculation. Any k-means clustering method depends on the number of cluster set at the beginning. There is no guarantee that the centers will move or converge to the mean point of the average of the cluster. This is the drawback of the k-means algorithm. Also there is no guarantee that the convergence will happen to the local mean. Assume that A_n is the set of i clusters to minimize the criteria J(P) so that x_{ci} converges to x_i (the cluster centers):

$$A_n = \{x_{c1}, x_{c2}, x_{c3}, ..., x_{ci}\} \quad (2)$$

Where:

$$J(x_{c1}, x_{c2}, x_{c3}, ..., x_{ci}; P) = P(\min_i |x_k - x_i|2) \quad (3)$$

If the S_n represents the entire dataset, then the objective is to find out a subset S_s of S_n such that

$P(S_s) \leq P(S_n)$. We assume that the data with one center is a stationary random sequence satisfying the following cumulative distribution sequence:

$$F_{x_n, x_{n+1}, \ldots, x_N}(x_n, x_{n+1}, \ldots, x_N) = F_{x_{n+k}, x_{n+1+k}, \ldots, x_{N+k}}(x_n, x_{n+1}, \ldots, x_N) \quad (4)$$

Then, the above sequence has on mean:

$$E(X) = c \quad (5)$$

The process of clustering is equivalent to minimizing the within-cluster sum of squares for the fast stage:

$$\min_s \sum_{i=1}^{c_i} \sum_{x_j \in S_i} ||x_j - \mu_{fa}|| \quad (6)$$

$$\min_s \sum_{i=1}^{c_i} \sum_{x_j \in S_i} ||x_j - c_{sl}|| \quad (7)$$

where, c are the centers of the clusters which are equals to the centers of the previous stage. The within cluster sum of squares is divided into two parts corresponding to the fast and slow stages of the clustering:

$$WSCC = \int_0^{c_i} \min(||x-c||, ||x-\overline{C}|| + \int_{c_i}^{1} \min(||x-c||, ||x-\overline{C}||dx \quad (8)$$

The centers of the slow stage start with c_i.

Algorithm:

Input: S_n, c, per, J_f, J_s
Output: S_n with clusters

S_n = % per of S_n
Select X_c from S_n randomly
While $J_f \leq \mu_{fa}$
For i←1 to n_s
Calculate the modified distance
$d(x_k, x_c) = [(x_k-x_{ci})^T (x_k-x_{ci})]$
Find minimum of d
Assign the cluster number to point X_i
End for
Calculate J_f,
End while
Calculate the average of the calculated clusters to find new centers X_c
Use the whole dataset S_n
While $J_s \leq \mu_{sl}$
For i←1 to n
Calculate the modified distance
$d(x_k, x_c) = [(x_k-x_{ci})^T (x_k-x_{ci})]$
Find minimum of d
Assign the cluster number to point X_i
End for
Calculate J_s,
End while

Fast k-means with careful seeding: The careful seeding procedure chooses the first centroid at random from X and each successive centriod from the remaining points according to the categorical distribution with selection probabilities proportional to the point's minimum squared Euclidean distance from the already chosen centriod. This tends to spread the points out more evenly and if the data is made of k well separated clusters is likely to choose an initial centroid from each cluster. This can speed convergence and reduce the likelihood of getting a bad solution. Fast K-means with careful seeding is also called as K-means++ (Arthur and Vassilvitskii, 2007). This algorithm begins with arbitrary set of cluster centers. At any given time, let D(x) denote the shortest distance from a data point x to the closest center already chosen. The algorithm is:

1a) Choose an initial center c_i uniformly at random x
1b) Choose the next center c_i selecting $c_i = x' \in X$ with probability $D(x')2/\Sigma_x \in X^{D(X)2}$
1c) Repeat step 1b until we have chosen a total of K centers
1d) Proceed as with standard K-means algorithm

We call the weighting used in step 1b simple "D2 weighting.

Fast k-means with weighted option: In k-means algorithm, every data point has an equal importance in locating the centroid of the cluster. This property is not worked out in case of density-biased sample clustering for which each data point represents varied density in the original data. Therefore, the clustering algorithm has to consider a weight associated with each data in the computation of cluster centers. This algorithm is called fast k-means with weighted options (Zhang, 2000).

Algorithm: Fast K-means with weighted option:

Input: A set of n data points and Number of clusters (K)
Output: Centriods of the K clusters

1. Initialize the K cluster centers
2. Repeat

   Assign each data point to its nearest cluster center according to the membership function:

$$m(c_j|x_i) = \frac{||x_i - c_j||^{-p-2}}{\sum_{j=1:k} |x_i - c_j||^{-p-2}}$$

3. For each center c_j, recomputed the cluster center c_j using current cluster memberships and weights:

$$c_j = \frac{\sum_{i=1:n} m(c_j|x_i) w(x_i) x_i}{\sum_{i=1:n} m(c_j|x_i) w(x_i)}$$

   where, $w(x_i)$ is the weighted associated with each data point

4. Until there is no reassignment of data points to new cluster centers. The membership function in this algorithm resembles that of the k-harmonic means algorithm (Zhang, 2000). Zhang (2000) also introduces the weight function $w(x_i)$ that represents the density of the original data points.

Fuzzy c-means algorithm: It is an iterative method for segmenting the image. Fuzzy clustering is one which is mostly used fuzzy approaches in image segmentation. FCM (Singha and Hemachandran, 2011; Priya *et al*., 2011) can be used to build clusters where the class membership of pixel can be interpreted as the degree of belongingness of the pixel to the clusters.

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ represent a set of pixel of the given image. Where n = number of pixels, $V = \{V_1, V_2, V_3, ..., V_n\}$ = set of fuzzy cluster centers, c = number of clusters. The aim of this algorithm is to minimize the objective function D(U, V), i.e. in this case squared error clustering criterion defined as:

$$J(V, U) = \sum_{I=1}^{N} \sum_{J}^{C} \mu_{ijm} \left\| x_i - v_j \right\|^2 \quad (9)$$

where, $\|x_{ij}-v_i\|$ Euclidean distance between x_{ij} and v_j. μ_{ij} membership degree of pixel x_i and v_j cluster centre, μ_{ij} has to satisfy the following conditions:

$$\mu_{ij} \in [0, 1], \forall i = 1,...,n, \forall i = 1, ...,c \quad (10)$$

$$\sum_{j=1}^{c} \mu_{ij} = 1, \forall i = 1,...,n \quad (11)$$

$U = (\mu)_{ijn+2}$ = fuzzy partition matrix. m = fuzziness index used to control the fuzziness of membership of each pixel. Value within the range $m \in [0, 1]$. It is a weighing exponent that satisfies m>1 and controls the degree of fuzziness in the resulting membership function. The FCM algorithm is similar to k-means algorithm:

- Select the number of clusters
- Assign randomly to each point coefficients for being in the clusters
- Repeat until the algorithm has converged
  - Compute the centroid for each cluster
  - For each point, compute its coefficients of being in the clusters

This algorithm is used to minimize the intra-cluster variance of the image.

Proposed image segmentation algorithms:

- Read the colour image
- Smooth the image using the average filter
- Enhance the image using de-correlation technique
- Convert the RGB image into L*u*v color space
- Apply Fuzzy c-means algorithm
- Apply Fast K-means algorithm
- Apply Fast K-means Clustering algorithm with weight option
- Apply Fast K-means Clustering algorithm with careful seeding option
- Label every pixel in the image using the results obtained from each method
- Create the segment image of each method
- Find out the execution time of each method

RESULTS AND DISCUSSION

The fast K-means with careful and weight option and Fuzzy c-means algorithm are implemented using Matlab V7.5. These algorithms are tested on various cervical images. Initially, the given image is smoothened for removing the noise and then image is enhanced using de-correlation technique. Then, the enhanced image is converted into L*u*v color space. Then, the converted image is segmented using Fast K-means algorithm with weighted and careful seeding method and Fuzzy c-means algorithm. Finally, the performance analysis is carried out among different algorithms. From the results, we conclude that (Fig. 1):

Fast K-means algorithm: The speed up of the algorithms is varying according to the accuracy. For the lower range

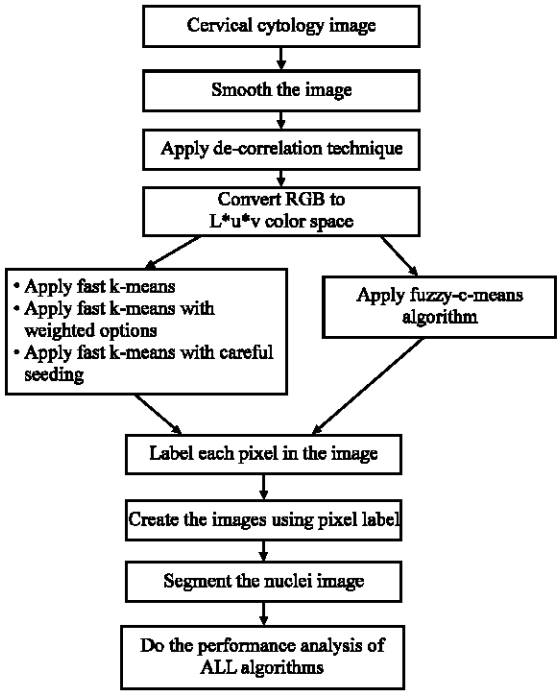

Fig. 1: Scheme of proposed algorithm

of accuracy, the speed up of the clustering is ranges from 1-9 times. This would reduce for the higher accuracy.

Fast k-means with careful seeding: It outperforms K-means algorithm both by achieving a lower potential value in some cases by several orders of magnitude and also by having a faster running time. It produces the accuracy in clustering and always attains the optimal clustering. Local search is converged in less iteration. In the real world dataset, it achieves 10% accuracy improvement and performs better than K-means.

Fast k-means with weighted option: Lower value of a squared objection function reflects a better quality on clustering. It is having the faster running time.

Fuzzy c-means: This method produces large number of segments based only on normal image parameter. If the number of cluster is increased, it will perform more iteration, thus automatically increasing the running time (Fig. 2-5). The result of comparison is given in the Table 1.

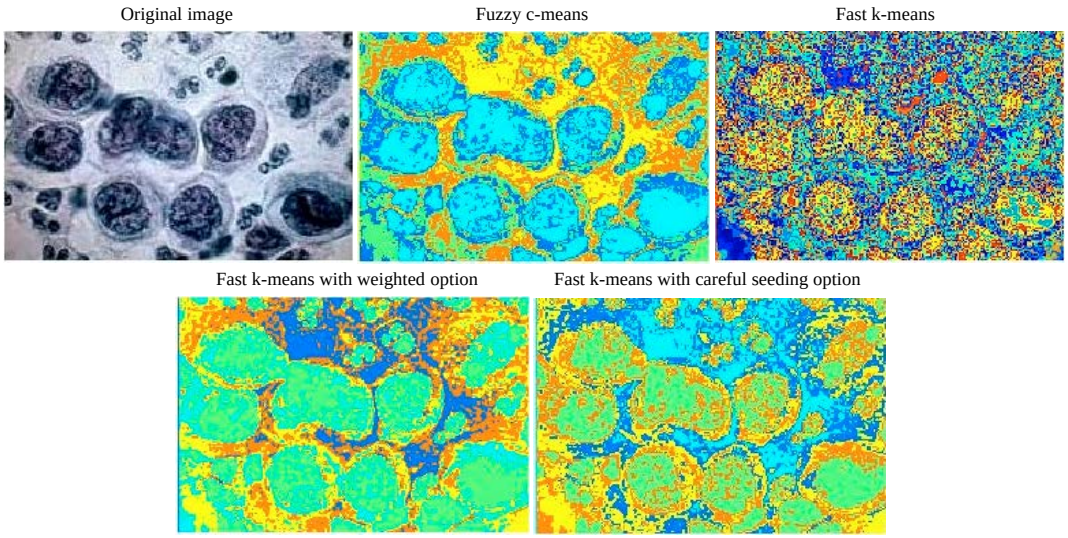

Fig. 2: Input image with segmented result of each algorithm

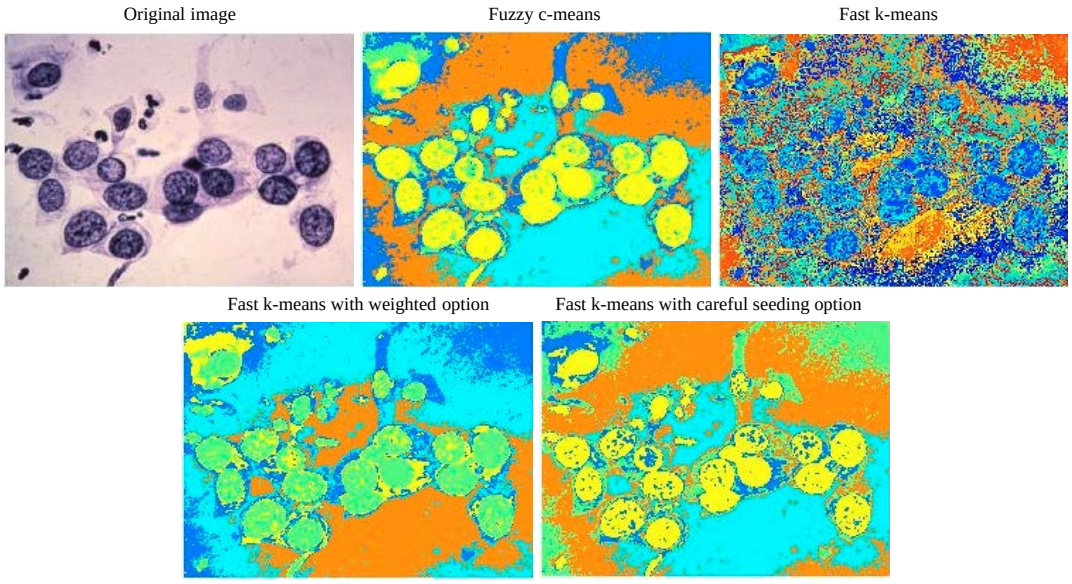

Fig. 3: Img-abnormal4 image with segmented result of each algorithm

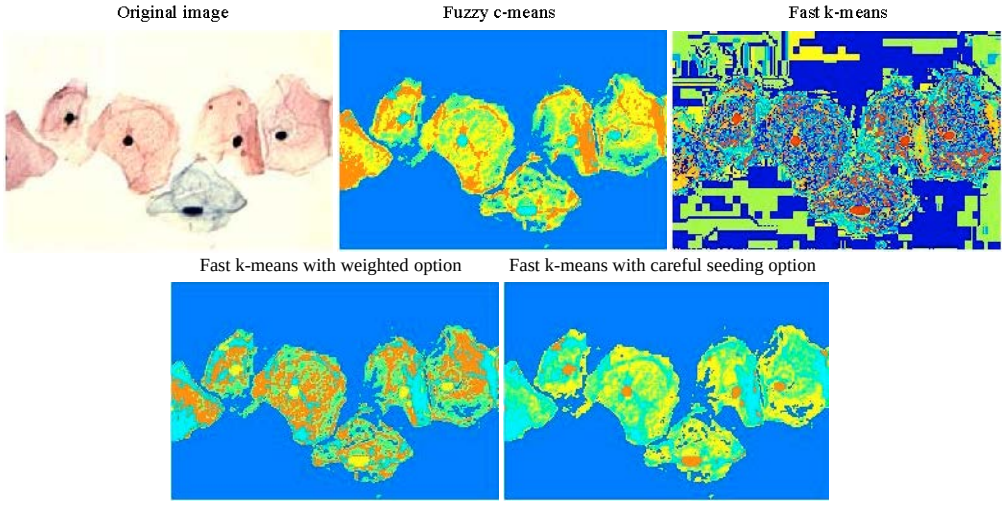

Fig. 4: Img-normal4 image with segmented result of each algorithm

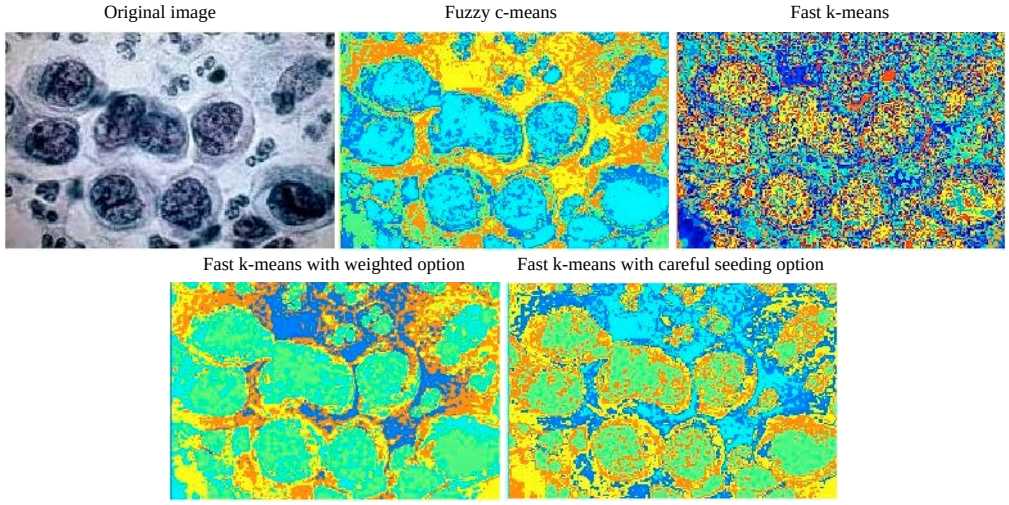

Fig. 5: Original image with segmented result of each algorithm

Table 1: The result of comparison

	Fuzzy-c-means		Fast K-means		Fast K-means with weight option		Fast K-means with careful seeding option	
Image name	No. of clusters	Elapsed time in seconds	No. of clusters	Elapsed time in seconds	No. of clusters	Elapsed time in seconds	No. of clusters	Elapsed time in sec
Original image	5	7.2031	5	13.8281	5	3.3906	5	2.8594
Img-abnormal4	5	71.8281	5	44.1875	5	17.0313	5	13.2344
Img-normal4	5	15.4219	5	60.1516	5	3.7656	5	1.6094
Input	5	4.7031	5	8.0789	5	1.5000	5	0.7500
Remarks	Sensitive to noise and presents of outlier and not converge to optimal solution		Converge to optimal solution. Speed and accuracy are complement to each other Less sensitive to noise		Converge to optimal solution Achieve speed and accuracy Less sensitive to noise		Converge to optimal solution Achieve speed and accuracy Less sensitive to noise	

CONCLUSION

This study is implemented in MatlabV7.5 and tested on various cervical images. First, the given cervical image is smoothed, enhanced and converted into L*u*v color space. The L*u*v color space image is segmented using fast K-means, fast K-means with careful seeding and weight option and Fuzzy c-means algorithms. Finally,

the performance of all algorithms is carried out. The experimental results show that FCM gives good segmentation but takes more time. Fast K-means provides accurate segmentation but it takes more time. In general, obtaining the optimal clustering is based on the initial seeding. The segmentation time for fast k-means with careful seeding is less and also the quality of segmentation is much accurate than other methods. In future, hybrid segmentation algorithm can be developed to produce the accurate result without noise and avoid the over segmentation.

REFERENCES

Arthur, D. and S. Vassilvitskii, 2007. K-means++: The advantages of careful seeding. Proceedings of the 18th Annual ACM-SIAM Symposium of Discrete Analysis, January 7-9, 2007, New Orleans, LA., USA., pp: 1027-1035.

Chitade, A. Z. and S.K. Katiyar, 2010. Colour based image segmentation using k-means clustering. Int. J. Eng. Sci. Technol., 2: 5319-5325.

Hegadi, R.S. and R.K. Sangaoli, 2011. Segmentation of google map using colour features. Proceedings of International Conference on Communication Computation and Management and Nanotechnology, September 23-25, 2011, India.

Kothainachiar, S. and R.S.D.W. Banu, 2007. A novel image segmentation based on a combination of color and texture features. GVIP J., 7: 45-51.

Madhuri, E. and V.M. Sandeep, 2012. Perceptual color image segmentation through K-means. Int. J. Eng. Res. Applic., 2: 1312-1314.

Mahanta, L.B., D.C. Nath and C.K. Nath, 2012. Cervix cancer diagnosis from paper smear images using structure based segmentation on shape analysis. J. Emerging Trends Comput. Inform. Sci., 362: 245-259.

Patidar, P., M. Gupta, S. Srivastava and A.K. Nagawat, 2010. Image denoising by various filters for different noise. Int. J. Comput. Applic., 9: 45-50.

Priya, R.K., C. Thangaraj and C. Kesavadas, 2011. Fuzzy C-means method for colour image segmentation with L*u*v colour transformation. Int. J. Comput. Sci. Issu., 1: 123-127.

Singha, M. and K. Hemachandran, 2011. Color image segmentation for satellite images. Int. J. Comput. Sci. Eng., 3: 3756-3762.

Zhang, B., 2000. Generalized k-harmonic means-boosting in unsupervised learning. Technical Report HPL-2000-137, Hewlett-Packard Labs.