A novel method for query based image retrieval using prototype based clustering

R. Tamilkodi*

Saveetha University,
Godavari Institute of Engineering and Technology,
NH-16, Chaitanya Knowledge City,
Rajahmundry, Andhra Pradesh, 533296, India
Email: tamil_kodiin@yahoo.co.in
*Corresponding author

G. Roseline Nesa Kumari

CSE,
Saveetha School of Engineering,
Chennai, Tami Nadu, 600077, India
Email: rosemaruthu@gmail.com

S. Maruthuperumal

CSE,
NBKR Inst. of Science and Tech.,
Nellore, AP, 524413, India
Email: maruthumail@gmail.com

Abstract: Content-based image retrieval is a process applied for searching of images from a large database in which searching is done based on the content of the image. The content of the image can be of colour, texture, and shape. This paper concentrates the primitive feature of texture in which the extraction of image content is considered. The implementation of such a system requires the extraction and storing of the image features to be compared with the features of the query image with this flow, the implementation process is more dynamic since all features have already been stored somewhere. This paper proposed a new method called prototype based cluster (PBC) where the features with similar values or properties are grouped together to form clusters and comparison is made between these clusters with the database images and the relevant images are retrieved and stored. This method will show good performance when compared with existing ones. The experimental result shows the effectiveness of our proposed method PBC applied on the query image.

Keywords: clusters; retrieval; query image; performance; features; extraction.

Reference to this paper should be made as follows: Tamilkodi, R., Kumari, G.R.N. and Maruthuperumal, S. (2018) 'A novel method for query based image retrieval using prototype based clustering', *Int. J. Image Mining*, Vol. 3, No. 1, pp.79–89.

Biographical notes: R. Tamilkodi received her MCA from Bharathidasan University, Trichy, in 2002, and MTech (IT) from Karnataka University, Karnataka in 2010. She is working towards her PhD at Saveetha University. She is with Godavari Institute of Engineering and Technology, AP as an Associate Professor. She has 11 years of experience in teaching undergraduate students and postgraduate students. Her research interests are in the areas of image processing, content-based image retrieval and cloud computing. She is a Life Member of ISCA and Red Cross Society.

G. Roseline Nesa Kumari received her PhD in Computer Science and Engineering at Dr. MGR University, Chennai, and received her ME from Sathyabama University, Chennai in 2005. She is having 13 years of teaching experience. At present, she is working as a Professor in the Department of Computer Science and Engineering at Saveetha University, Chennai. She published sixty research publications in various international, national conferences and journals. She is a Life Member of Indian Science Congress Association (ISCA), IAENG, IET, CSI (Life Member), Red Cross. Her research interest includes image processing, steganography, network security and information security.

S. Maruthuperumal received his BE in Computer Science and Engineering from Bharathidasan University and ME in Computer Science and Engineering from Sathyabama University, Chennai. He received PhD from Dr. MGR University, Chennai. He is having 15 years of teaching experience. At present, he is working as a Professor at NBKR Institute of Science and Technology, Vidyanagar, Nellore, AP. He published 57 research publications in various international, national conferences and journal. His research interest includes image processing, digital watermarking, steganography and security. He is a Life Member of ISCA, IAENG and Institutional Member of CSI.

1 Introduction

Imaging has played an important role in our life. Our ancestors used the walls of their caves to paint some pictures telling us some information about their life. The emergence of the World Wide Web enables users to access data from any place and provides the exploitation of digital images in any fields. Naturally, when the amount of data becomes larger and larger it will be useless and unless are effective methods to access. From the past and until now images from different fields are stored in groups as image databases for each field thousands of images have common features. If we want to search for a certain image, the primitive methods like search by text or image description are not accurate and time-consuming.

To search for an image in a huge image database it is not efficient to use text or image descriptors to overcome this problem, a new technique called content-based image retrieval is used to search and retrieve an image from the database. Typical content based image retrieval (CBIR) uses the contents of an image to represent and access CBIR systems extract features from images in the database based on the value of the image pixels. These features are smaller than the image size and stored in a database called feature database.

The main advantage of using CBIR System is that the system uses image features instead of using the image itself. So; CBIR is a cheap, fast and efficient over image

search methods. Research in CBIR is a hot topic nowadays. It is important to realise that image retrieval does not entail solving the general image understating problems. The retrieval system presents similar images. The user should define what the similarity between images has to be. So when we want to develop an efficient algorithm for CBIR, some problems have to be solved.

The first problem is to select the image features that will represent the image, some of the features can be visual features (colour, texture, shape). The second problem is the computational steps to extract features of an image and to deal with large image databases. We have to keep in our mind that large image databases are used for testing and retrieving. In this area of research selection of features to represent an image in the database is still an unresolved issue. Existing approaches are inefficient and inaccurate to the query image. We propose a new efficient technique called PBC to solve these problems by using clustering algorithms. Image clustering based on the PBC reduces the size of matching data in the retrieval.

The rest of the paper is organised as then following: Section 2 brief about the reviews based on some related work and CBIR systems. Section 3 presents the algorithm description and its main role as a pre-processing step. In Section 4, we introduce the proposed (PBC) system for CBIR system implementation and experimental results are given in Section 5. Section 6 summarises our proposed system and some proposed future work.

2 Reviews on related work

Clustering is a form of unsupervised classification that aims at grouping data points based on similarity. Clustering has widespread applications in image processing. Colour-based clustering techniques have proved useful in image segmentation (Yang et al., 2012). The K-means algorithm is quite popular for this purpose. Clustering based on the visual content of images (Dey and Ashour, 2016) is an area that has been extensively researched for several for this finds application in image retrieval. Image clustering can also be used to optimise the performance of a number of CBIR system while the performance of a number of clustering algorithms in image retrieval has been analysed in previous works (Liu et al., 2007; Amory et al., 2012; Li and Shapiro, 2002). A hybrid method has been proposed for image clustering based on combining the particle swarm optimisation (PSO) with k-means clustering algorithms (Younus et al., 2014). We apply our proposed algorithm to CBIR and compare its performance with that of the k-means clustering. K-means have been widely used for its simple and rapid calculation. The World Wide Web consists of billions of web images, and the results of a query to a search engine can return thousands of images. Clustering can be used to group these search results into a small number of clusters, each of which captures a particular aspect of the query. Younus et al. (2014) discussed his work towards the primitive features of colour, shape, Texture. Concentrates more on all the methods and made the final resultant by applying PSO method which named as swarm optimisation algorithm and showed the results for image retrieval. Jhanwar et al. (2004) focus on the cooccurrence matrices where a kind of analysis is the computation of grey level cooccurrence matrices. The matrix can be computed by separating the intensity values present in the image into a small number of different levels. But the present suggested method is purely based on clustering and extracting the features from various levels of indexing gives exact retrieval of query

image from a large collection comparison is made with Younus et al. (2014) and Jhanwar et al. (2004) which is discussed in Table 1.

Table 1 Precision and recall values of methods compared

	Jhanwar et al. (2004)		*Younus et al. (2014)*		*Proposed method (PBC)*	
Image	*Avg. precision*	*Avg. recall*	*Avg. precision*	*Avg. recall*	*Avg. precision*	*Avg. recall*
Food	0.745	0.198	0.824	0.245	0.843	0.285
Horse	0.725	0.176	0.765	0.195	0.789	0.211
Dinosaur	0.789	0.186	0.765	0.192	0.802	0.201
Elephant	0.801	0.241	0.841	0.257	0.858	0.276
Building	0.721	0.187	0.736	0.19	0.751	0.199
House	0.765	0.179	0.762	0.211	0.808	0.242
Transport	0.814	0.218	0.808	0.237	0.857	0.279
Roses	0.798	0.212	0.821	0.215	0.831	0.235
Mountain	0.854	0.245	0.845	0.236	0.883	0.271
Sea	0.811	0.247	0.841	0.219	0.847	0.258
Avg.(%)	0.7823	0.2089	0.8008	0.2197	0.8269	0.2457

3 Proposed method PBC

Prototype based clustering techniques create a one level partitioning of the data objects. There are a number of such techniques but the most prominent are K-means and prototype based clustering. Prototype based clustering defines in terms of a finding mean which we can call it as centroid or in other words the mean of a group of points and is typically applied to objects in a continuous-dimensional space.

First initialise the number of clusters. Each point is assigned to the initialised clusters based on how close it is to the cluster (Closeness is identified by calculating the distance between the point and the cluster value). This procedure is repeated till all the points are assigned to the initialised clusters. The group of point present constitutes a cluster. Again the procedure is repeated with these groups of clusters and the above-said process is repeated until we are not getting repeated points in clusters i.e., further there is no variation in points present in the cluster. The process is stopped. For each cluster, we can generate the output image. Figure 1 shows the block diagram of the proposed method.

3.1 Steps involved in proposed method

1 Initialise the number of clusters.(Say C_1,C_2,C_3, …, C_K).

2 Assign points to the clusters by using distance formula (If $R = (R_1, R_2, R_3, \ldots, R_n)$ and $S = (S_1, S_2, S_3, \ldots, S_n)$ are two points, then the distance from R to S, or from S to R is given by

$$D(R, S)=\sqrt{\sum_{I=1}^{n} R_i - S_i}.$$

3 Group the points then calculate the new clusters.

4 The above procedure is repeated until there is no difference in points and their clusters.

5 Calculate the mean by using equation (1) of all clusters and store the value.

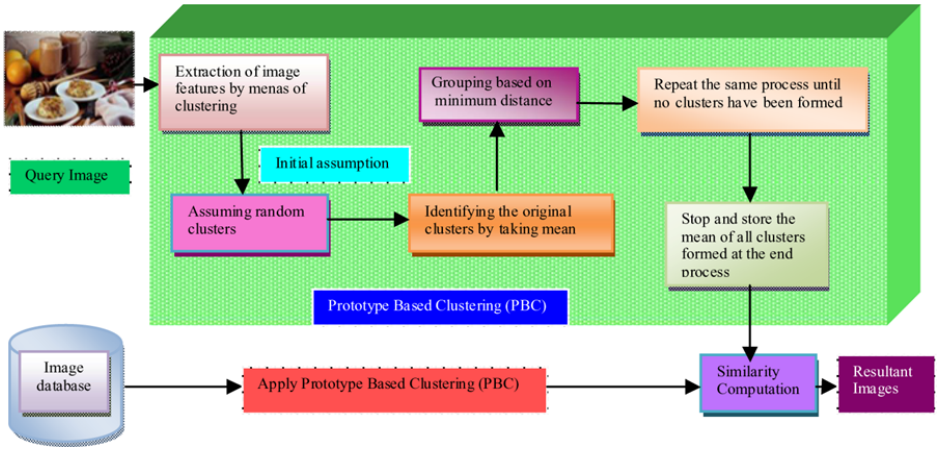
$$\text{Mean} = \frac{\text{Sum of data points}}{\text{Number of data points}}. \quad (1)$$

6 The above procedure is repeated for the database image and the resultant mean is stored.

7 Similarity measurement is applied to the stored mean for the query as well as database image.

8 If mean equals for the particular cluster specified by the user then the corresponding images are retrieved.

Figure 1 Block diagram of planned PBC (see online version for colours)



4 Experimental results

The experimental results shown in Figure 2(a)–(d) here describe the clustering output for our proposed method PBC. In this, we have taken a query image in which our proposed method PBC is applied. We made an initial assumption as 3. The image can be fed into this proposed method PBC and the corresponding cluster outputs are as shown in Table 2. Once the output image has been generated we are finding the mean for all three clusters. Then one can able to compare the mean of any cluster with the database images. By repeating the procedure which is explained in the proposed method one can able to get the resultant mean for the database image. Here our proposed method PBC is specific in mentioning the number of clusters, as well as the comparison, is made between

any one of the clusters. Take for example if we want to compare the first cluster mean of the query image then the database image is also compared with the first cluster means it has got.

Figure 2 (a) Image clustered into 3; (b) output image generated for cluster 1; (c) output image generated for cluster 2 and (d) output image generated for cluster 3 (see online version for colours)

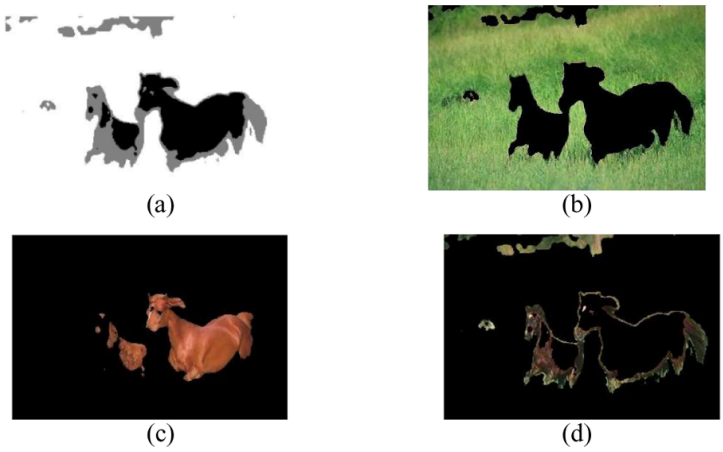


Table 2 Means got from various clusters for the categorised images

Image	*Original mean*	*Indexed mean*	*Mean of Cluster 1*	*Mean of Cluster 2*	*Mean of Cluster 3*
Food	173.8614	206.3135	131.5169	152.0495	134.0906
Horse	160.6258	160.5432	126.6738	141.4642	136.2664
Dinosaur	230.3399	205.1233	159.036	210.2528	147.603
Elephant	198.2735	164.4888	134.2994	149.3761	157.8413
Building	191.4827	216.8341	157.7465	167.8927	152.2955
House	152.2819	188.1724	139.2978	129.1261	136.1259
Transport	191.4827	216.8341	157.7465	167.8927	152.2955
Roses	153.4382	153.1978	126.0175	123.6205	130.6858
Mountain	170.1281	151.4001	124.1347	141.9551	146.9917
Sea	196.9676	184.7245	131.6861	133.6405	174.5845

This experimental section tests our proposed system and traces the changes. This section totally deals with the description of our results and giving clarification for the output we got and also compared with the existing system which is better in retrieving. We have proposed a new system which is PBC works by forming the input image into clusters and on the user's interest depends on the specification of the cluster the mean in compared with the image database. The working procedure for our proposed method is clearly explained in Section 2.

We have taken the image database called WANG for testing our proposed method in which nearly 1000 images were present. For our convenience, we have categorised the images into classes each of which consists of some similar group of images. This method

PBC is applied to the image category and the resultant images are retrieved based on the type of cluster the images retrieved. To know the retrieval efficacy of our proposed method we would like to compute the performance evaluation by knowing the precision and recall. The system is tested with various categories of images from image database called WANG.

Table 2 shows the various cluster means obtained for the random images generated from the WABG image database. Based on the indexing of the clusters present in the image the mean will vary from one cluster to another. In this table, if we look into the values of cluster 1 in dinosaur the mean is 159.036 but while we are looking into the mean of cluster 2 for dinosaur is 210.2528 while comparing with other images this image cluster2 is greater when compared with other cluster image means. Depends on the indexing value of the colour it will differ from one cluster to another.

4.1 Computation of precision and recall

The precision and recall are mainly used to know the performance of the methods in CBIR. Precision measures the retrieval accuracy, while the equation is given as follows:

$$\text{Precision} = \frac{\text{Number of relevant image retrieved}}{\text{Total number of images retrieved}}.$$

Recall measures the retrieval robustness while the equation is given as follows:

$$\text{Revall} = \frac{\text{Number of relevant image retrieved}}{\text{Total number of images retrieved}}.$$

The research is exposed to the number of images set as 10 to compute the average precision and recall of each query image. The proposed PBC method experiments based on sum-of-absolute differences (SAD) Similarity measures is shown in equation (3).

$$\text{SAD}(f_q, f_t) = \sum_{i=0}^{n-1} 1 f_q[i] - f_t[i], \tag{3}$$

where f_q, f_t represent query feature vector and database feature vectors and n is the number of features in each vector.

Table 3 shows the precision of our proposed method PBC gives 82% and Recall gives 24% which shows our method works better.

We are compared our method PBC with other clustering methods also and the result is displayed as shown in Table 1. The output of this method is compared to the performance of the methods of Jhanwar et al. (2004) and Zeyad and Al-Zhour (2014).

Table 1 shows that our proposed system performance is better than other systems for all categories of images except for few images there will be a slight variation. The major reason lies behind this is if we are dividing the image into clusters it will group the pixels values into groups and the values near to the group form a new group. While forming the cluster with less indexing then there will be a chance of getting the output similar to other methods. So have a concentration on grouping the clusters. Our proposed method groups the clusters into three and experimentation is carried out based on this cluster which gives better performance when compared with other methods carried over with clusters.

Figure 3 shows the query image horse and Figure 4 shows the resultant image obtained after applying our proposed method (PBC) for the image database WANG. Table 2 gives the details about the means of all clusters for the random images we have taken from the image database. The image retrieved on the average of 80%. The overall performance of proposed PBC obtained is more than 76% are perfect, competent and ordered. Figures 5–7 shows the graphical representation of the precision and recall values of proposed and compared methods.

Figure 3 Qeury image (see online version for colours)

Figure 4 Retrieved images based on query image (see online version for colours)

Table 3 Performance of the proposed method PBC related to precision and recall

Image	*Avg. precision*	*Avg. recall*
Food	0.843	0.285
Horse	0.789	0.211
Dinosaur	0.802	0.201
Elephant	0.858	0.276
Building	0.751	0.199

Table 3 Performance of the proposed method PBC related to precision and recall (continued)

Image	*Avg. precision*	*Avg. recall*
House	0.808	0.242
Transport	0.857	0.279
Roses	0.831	0.235
Mountain	0.883	0.271
Sea	0.847	0.258
Avg(%)	0.8269	0.2457

Figure 5 Graphical representation has been depicted to show the precision and recall of the proposed method (PBC) (see online version for colours)

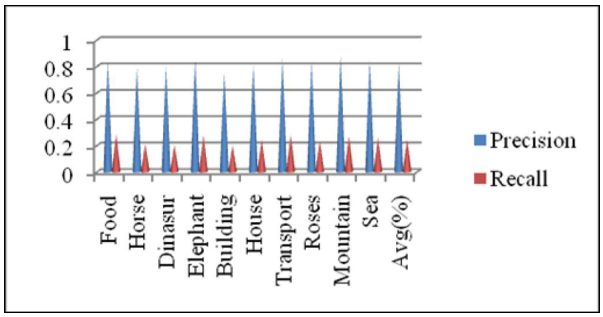


Figure 6 Graphical representation of precision values with other methods (see online version for colours)

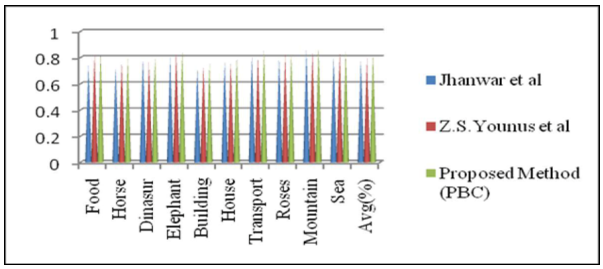
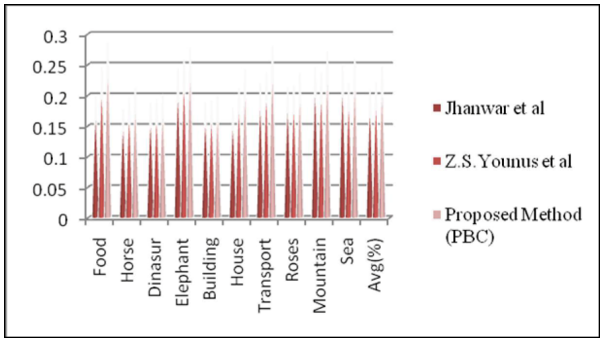


Figure 7 Graphical representation of recall values with other methods (see online version for colours)



5 Conclusion

The proposed method PBC shows effective in the retrieval process. The image is fed into three levels of separation based on the intensity values of the image. Based on the user needs the type of cluster has been selected and the corresponding mean is compared with the image database. These method performances well in retrieving the images related to the query image supplied. A future enhancement may be done with this method by adding an extra primitive feature to get betterment of the results.

References

Amory, A.A., Sammouda, R., Mathkour, H. and Jomaa, R.M. (2012) *A Content Based Image Retrieval Using K-Means Algorithm*, 978-1-4673-2430-4/12/$31.00 ©2012 IEEE.

Dey, N. and Ashour, A. (2016) *Classification and Clustering in Biomedical Signal Processing*, © 2016, 463 pages.

Jhanwar, N., Chaudhuri, S., Seetharaman, G. and Zavidovique, B. (2004) 'Content based image retrieval using motif co occurrence matrix', *Image and Vision Computing*, Vol. 22, pp.1211–1220, Published by Elsevier, B.V. DOI: 10.1016/j. imavis.2004.03.026.

Li, Y. and Shapiro, L.G. (2002) *Consistent Line Clusters for Building Recognition in CBIR*, 1051-465VOL 2, 2002 IEEE.

Liu, Y., Zhang, D., Lu, G. and Ma, W-Y. (2007) 'A survey of content based image retrieval with highlevel semantics', *Pattern, J. Recognit*., Vol. 40, pp.262–282.

Yang, Y-F., Wu, J., Fang, J. and Cui, Z-M. (2012) *Parallel Hierarchical K-Means Clustering-Based Image Index Construction Metho*d, 978-0-7695-4818-0/12 $26.00 © 2012 IEEE.

Younus, Z.S., Mohamad, D., Saba, T., Alkawaz, M.H., Rehman, A., Al-Rodhaan, M. and Al-Dhelaan, A. (2014) 'Content-based image retrieval using PSO and k-means clustering algorithm', *Arab, J. Geosci*., Springer Berlin, Heidelberg, Vol. 8, No. 8, 31 August, pp.6211–6224, Print ISSN: 1866-7511, Online ISSN1866-7538, https://doi.org/10.1007/s12517-014-1584-7

Zeyad, A.A. and Al-Zhour, Z. (2014) 'The general (vector) solutions of such linear (coupled) matrix fractional differential equations by using Kronecker structures', *Applied Mathematics and Computation*, Elsevier Inc., http://dx.doi.org/10.1016/j.amc.2014.01.079 0096-3003/

Bibliography

Chang, R-I., Lin, S-Y., Ho, J-M., Fann, C-W. and Wang, Y-C. (2012) 'A novel content based image retrieval system using K-means/KNN with feature extraction', *ComSIS*, Vol. 9, No. 4, Special Issue, December, pp.1642–1662.

Chen, Y., Wang, J.Z. and Krovetz, R. (2003) 'Content-based image retrieval by clustering', *MIR'*03, 7 November, Berkeley, California, USA.

Chen, Y., Wang, J.Z. and Krovetz, R. (2005) 'CLUE: cluster-based retrieval of images by unsupervised learning', *IEEE Transactions on Image Processing*, Vol. 14, No. 8, August, pp.1057–7149.

Dutta, P.J., Bhattacharyya, D.K., Kalita, J.K. and Dutta, M. (2006) 'Clustering approach to content based image retrieval', *Proceedings of the Geometric Modeling and Imaging–New Trends (GMAI'06)*, 0-7695-2604-7/06 $20.00 © 2006 IEEE.

Ho, J-M., Lin, S-Y., Fann, C-W., Wang, Y-C. and Chang, R-I. (2012) *A Novel Content Based Image Retrieval System Using K-Means with Feature Extraction*, 978-1-4673-0199-2/12/$31.00 ©2012 IEEE.

Jenni, K. and Mandala, S. (2014) *Pre-Processing Image Database for Efficient Content Based Image Retrieval*, 978-1-4799-3080-7114/$31.00 ©2014 IEEE.

Mahajan, S. and PatiI, D. (2014) *Image Retrieval Using Contribution-Based Clustering Algorithm with Different Feature Extraction Techniques*, 978-1-4799-3064-7/14/$31, 00©20 14 IEEE.

Sadiq, M.J., Kaleem, A., Mohammad, A.H. and Wajid, M.A. (2016) 'Content based image retrieval system using K-means and KNN approach by feature extraction', *IJCSCN*, December–January 2015.

Shekhar, S., Singh, A. and Agrawal, S.C. (2015) 'An object-centric image retrieval framework using multi-agent model for retrieving non-redundant web images', *Int. J. Image Mining*, Vol. 1, No. 1, pp.4–22.

Stehling, R.O., Nascimento, M.A. and Falcgo, A.X. (2001) *An Adaptive and Efficient Clustering-Based Approach for Content-Based Image Retrieval in Image Databases*, 1098-8068/01$10.000 2001 IEEE.

Tejasya, T. and Jyothi, S. (2016) 'An improved image searching method K-means clustering', *International Journal of VLSI System Design and Communication Systems*, Vol. 4, No. 5, May, pp.0346–0351, ISSN 2322-0929.

Zhang, J. and Zou, W. (2010) *Content-Based Image Retrieval Using Color and Edge Direction Features*, 978-1-4244-5848-6/10/$26.00 ©2010 IEEE.