CBIR USING SALIENCY MAPPING AND SIFT ALGORITHM

Mr. D. R. Dhotre, Dr. G. R. Bamnote, Aparna R. Gadhiya, Gaurav R. Pathak

Abstract— With the growing computer technologies and the advance in speed of World Wide Web, there has been increase in the complexity of multimedia information. More users are attracted to text based search. This produces lot of garbage. Content based image retrieval (CBIR) system has been developed as an efficient image retrieval tool, whereby the user can provide their query to the system to allow it to retrieval the user's desired image from the database. CBIR system consists of feature extractor and the derived features are led to SVM. The disadvantage with this is that human perception is not fulfilled successfully. When looking at some image people are usually attracted by some particular objects within the image. Other subjects are uninteresting for them. Detecting these salient regions is called saliency detection. The proposed approach in this paper combines the feature extraction algorithm; SIFT with the Saliency Detection technique in order to provide relevant image output. The approach also considers texture/energy level of an image as a feature. The combination of these three concepts evaluates to refine the CBIR.

Index Terms— Content Based Image Retrieval (CBIR), Support Vector Machine (SVM), Scale Invariant Feature Transform (SIFT), Saliency Detection, Difference of Gaussian (DOG).

1 INTRODUCTION

There have been many research efforts to improve the retrieval efficiency of CBIR. The various approaches made are still limited and do not fulfill the human perception. In order to reduce the semantic gap and acquire efficiency in CBIR, we propose the utilization of Saliency Map in combination with feature extraction algorithm SIFT and we also detect texture/energy level using wavelet transform. In this approach Saliency Map represent the salient regions of an image while SIFT provide salient key points and wavelet transform provide the energy level as a feature of an image. The feature vector derived out of this is then used to compare with the feature already stored in the database. The SVM classifier takes these features as input and classifies the set of images into relevant and irrelevant set [4].

2 BACKGROUND

A. Content Based Image Retrieval

With advances in the multimedia technologies and the advent of the Internet, Content-Based Image Retrieval (CBIR) has been an active research topic since the early 1990's. Most of the early researches have been focused on low-level vision alone. However, after years of research, the retrieval accuracy is still far from users' expectations. It is mainly because of the large gap between high-level concepts and low-level features [2].CBIR system takes query images as input. Further various feature extraction techniques are applied to it so that prominent feature vector is obtained which is led to Support Vector machine and user gets most relevant image as a output. Content Based Image Retrieval (CBIR) is a prominent area in image processing due to its diverse applications in internet, multimedia, medical image archives, and crime prevention. Improved demand for image databases has increased the need to store and retrieve digital images. Extraction of visual features, viz., color, texture, and shape is an important component of CBIR.

3 INTRODUCTION TO FEATURE EXTRACTION

Feature extraction is the heart of the content based image retrieval. As we know that raw image data that cannot used straightly in most computer vision tasks. Mainly two reason behind this first of all, the high dimensionality of the image makes it hard to use the whole image. Further reason is a lot of the information embedded in the image is redundant. Therefore instead of using the whole image, only an expressive representation of the most significant information should extract. The process of finding the expressive representation is known as feature extraction and the resulting representation is called the feature vector.

A. Feature Extraction

Feature extraction is the basis of content based image retrieval. Typically two types of visual feature in CBIR:

- Primitive features which include color, texture and shape.
- Domain specific which are application specific and may include, for example human faces and finger prints.

Primitive features are those which can be used for searching like color, shape, texture and feature which are used for par-

- *D. R. Dhotre is currently working as an Asst. Professor in Comp.Science Department at SSGMCE, Sheagon.*
- *Dr. G. R. Bamnote is currently working as an Professor in Comp.Science Department at PRMIT&R, Badnera.*
- *Aparna Gadhiya is currently pursuing masters degree program in Comp.Science Department at SSGMCE, Shegaon.*
- *Gaurav Pathak is currently pursuing masters degree program in Comp.Science Department at SSGMCE, Shegaon.*

ticular domain and have knowledge about them. For example, we are searching for face of girl which belongs to human category, so here domain is human. Another one is we are searching for elephant which belong to animal category. These features are domain specific.

1) Color: Color is one of the most reliable visual features that are also easier to apply in image retrieval systems. Color is independent of image size and orientation, because, it is robust to background complication. First a color space is used to represent color images. Typically, RGB space where the gray level intensity is represented as the sum of red, green and blue gray level intensities. Swain and Ballard proposed histogram Support Vector Machine Intersection, an L1 metric as the similarity measure for color histogram. Color histogram is the most common method for extracting the color features of colored images. Color histograms are widely used for CBIR systems in the image retrieval area.

2) Texture: Texture is that innate property of all surfaces that describes visual patters, and that contain important information about the structural arrangement of the surface including clouds, trees, bricks, hair, and fabric and its relationship to the surrounding environment. Various texture representations have been investigated in both pattern recognition and computer vision.

3) Shape: Shape is the characteristic surface configuration that outlines an object giving it a definite distinctive form. In image retrieval, depending on the applications, some require the shape representation to be invariant to translation, rotation and scaling, whiles others do not. In general shape representation can be divided into two categories:
a) Boundary based which uses only the outer boundary of the shape.
b) Region-based which uses the entire shape regions.

4 SCALE INVARIANT FEATURE TRANSFORM

Scale Invariant Feature Transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. The algorithm was published by David Lowe in 1999. Applications include object recognition, robotic mapping and navigation image stitching, video tracking, 3D modeling, gesture recognition individual identification of wild life and match moving.

For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects [5]. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and illumination. Such points usually lie on high contrast regions of the image, such as object edges.

SIFT key points of objects are first extracted from a set of reference images and stored in a database. An object is recognized in a new image by individually comparing each
feature from the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors. From the full set of matches, subsets of key points that agree on the object and its location, scale, and orientation in the new image are identified to filter out good matches [10].

5 SALIENCY MAP

A. Definition

The purpose of the saliency map is to represent the conspicuity or "saliency" – at every location in the visual field by a scalar quantity and to guide the selection of attended locations, based on the spatial distribution of saliency [7].
Saliency map has its root in Feature Integration Theory and appears first in the class of algorithmic models above. It includes the following elements:-

1) An early representation composed of a set of feature maps, computed in parallel, permitting separate representations of several stimulus characteristics.

2) A topographic saliency map where each location encodes the combination of properties across all feature maps as a conspicuity measure.

3) A selective mapping into a central non-topographic representation, through the topographic saliency map, of the properties of a single visual location.

4) A winner-take-all (WTA) network implementing the selection process based on one major rule: conspicuity of location (minor rules of proximity or similarity preference are also suggested).

5) Inhibition of this selected location that causes an automatic shift to the next most conspicuous location. Feature maps code conspicuity within a particular feature dimension.

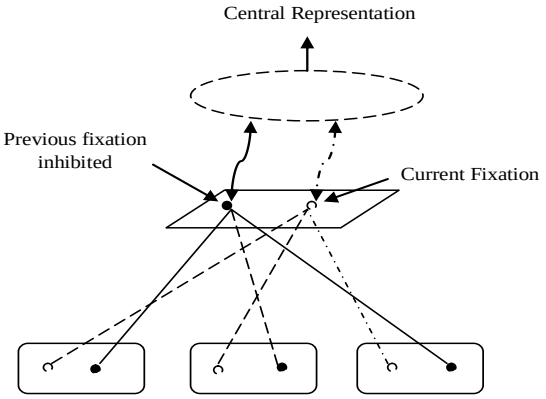

Fig. 1 Basic approach for saliency detection

B. Bottom-Up Approach

The core of visual saliency is a bottom-up, stimulus-driven signal that announces "this location is sufficiently different from its surroundings to be worthy of your attention". This *bottom-up* deployment of attention towards salient locations can be strongly modulated or even sometimes overridden by *top-down,* user-driven factors. Thus, a lone red object in a green field will be salient and will attract attention in a bottom-up manner [7].

C. Top-Down Approach

On the other hand, if you are looking through a child's toy bin for a red plastic dragon, amidst plastic objects of many vivid colors, no one color may be especially salient until your top-down desire to find the red object renders all red objects, whether dragons or not, more salient [7].

6 SALIENCY MAP DETECTION TECHNIQUE

In 1985 the authors Koch and Ullman introduced a concept of a saliency map. It was used to model the human attention and the shifting focus connected with sight and visual stimuli [1]. The saliency map for a given image represents how distinctive the image regions are and in what order the eye and the nervous system process them. In their paper, Koch and Ullman explain that saliency is a measure of difference of the image regions from their surroundings in terms of elementary features such as color, orientation, movement or distance from the eye. Later, Harel et al. combined activation maps derived from graph theory and other maps obtained by Itti's model to form a new graph-based saliency map [3]. Ma and Zhang proposed local contrast analysis to estimate saliency using a fuzzy growth model.

In addition, Liu et al. employed a set of features including multiscale contrast, center- surround histogram and color spatial distribution to describe a salient object, and a Conditional Random Field (CRF) was learned by combining these features to detect salient object. Goferman et al. proposed a context-aware saliency to detect the image regions, which depended on the single scale and multiscale saliency detection.

Lately, Cheng et al. proposed a regional contrast based saliency extraction algorithm, which simultaneously evaluated global contrast differences and spatial coherence. In addition to the contrast based methods mentioned above, saliency map can be computed by image frequency domain analysis. By analyzing the log-spectrum of natural images, Hou and Zhang generated the saliency map based on the spectral residual of the amplitude spectrum of an image's Fourier trans- form. However further authors proposed and proved that it is the phase spectrum instead of amplitude spectrum of Fourier transform is the key to calculate the locations of salient regions. More recently, Achanta et al. applied a frequency tuned method to compute center-surround contrast using color differences from an image, in which saliency values were averaged within image segments produced by Mean Shift pre-segmentation. Then, the authors extended their work by varying the bandwidth of the center-surround filtering near image borders using symmetric surrounds. Generally, compared with methods based on image feature contrast, methods based on frequency domain analysis can be easily implemented since they have lower computational complexity and fewer parameters [6].

7 WAVELET TRANSFORM

Wavelet transform have become one of the most important and powerful tool of signal representation. Nowadays, it has been used in image processing, data compression and signal processing. Due to the fact that human vision is much more sensitive to small variations in color or brightness, that is, human vision is much more sensitive to low frequency signals. Therefore, high frequency components in images can be compressed without distortion. Wavelet transform is one of the best tools for us to determine where the low frequency are and high frequency area is [8].

8 PROPOSED APPROACH

In this paper we attempt to find a solution to meet human perception and get relevant image as a output. The proposed model is a combination of saliency detection technique, SIFT algorithm for feature extraction and wavelet transform that provides texture/energy level of image.

A. *Saliency Detection*

The query image is taken into consideration. Saliency detection follows the following steps:-

Multiscale low level feature extraction is performed for image linear filtering. Features like colors (Red, green, blue, yellow, etc), intensity (on, off), orientation (0, 45, 90, 135), oth-

ers (Motion, Junction, and terminators etc) are taken into consideration. Further 9 spatial scales are created using Gaussian pyramids which low pass filter and subsample the input image progressively yielding horizontal and vertical image reduction factors ranging from 1.1(scale zero) to 1:256 (scale eight) in eight octaves. For each pixel in the pyramid color channels are generates.

R=r-(g+b)/2 (1)

G=g-(r+b)/2 (2)

B=b-(r+g)/2 (3)

Four Gaussian pyramids R( ), G( ), B( ), I( ) are created from these color channels where [0:8] is the scale. Features are mathematically computers using linear "Centre-Surround" operations and spatial competitions. Features maps are created for each feature. This feature maps are combined into conspicuity maps. Across scale addition is used to obtain each map reduced to scale 4 and point by point addition. The obtained conspicuity maps for each feature are then summed up to obtain final saliency map S.

$$S=1/3\ (\overline{N}(I)+\overline{N}(C)+\overline{N}(O)) \quad (5)$$

This procedure is basic approach for saliency detection. The output we get is saliency map. This is the approach of Itti-Koch saliency detection. Many others approaches are based on image feature contrast and frequency domain analysis. The other key point of an image is obtained by using SIFT algorithm.

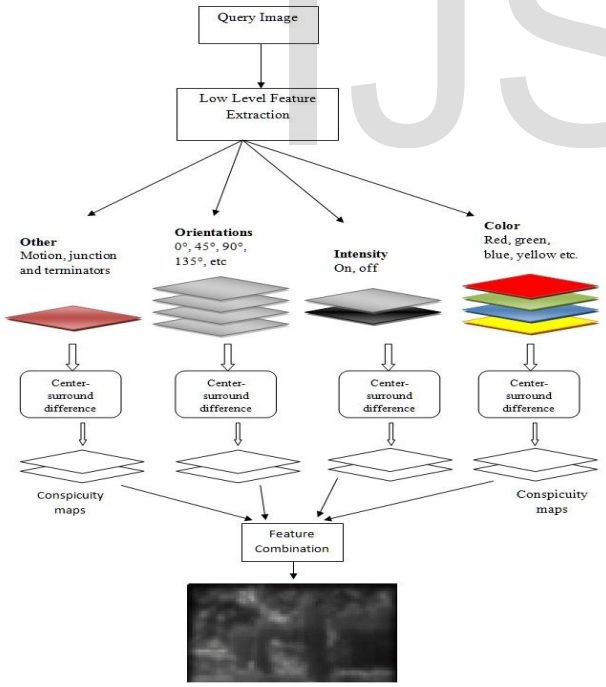

Fig. 2 Architecture representing Itti-Koch Model

B. SIFT Algorithm

Following are the main stages of computation in SIFT algorithm used to generate the set of image features [11].

1) Scale Space Detection: We begin by detecting points of interest, which are termed key points in the SIFT framework. The image is convolved with Gaussian filters at different scales, and then the differences of successive Gaussian-blurred images are taken. Key points are then taken as maxima/minima of the Difference of Gaussians (DoG) that occur at multiple scales. Specifically, a DoG image $D(x,y,\sigma)$ is given by

$$D(x,y,\sigma) = L(x,y,K_i\sigma) - L(x,y,K_i\sigma) \quad (6),$$

where $L(x,y,K\sigma)$ is the convolution of the original image I(x, y) with the Gaussian blur $G(x,y,K\sigma)$ at scale $K\sigma$, i.e.,

$$L(x,y,K\sigma) = G(x,y,K\sigma) * I(x,y) \quad (7)$$

Hence a DoG image between scales $K_i\sigma$ and $K_i\sigma$ is just the difference of the Gaussian-blurred images at scales $K_i\sigma$ and $K_i\sigma$ .For scale space extrema detection in the SIFT algorithm, the image is first convolved with Gaussian-blurs at different scales. The convolved images are grouped by octave (an octave corresponds to doubling the value of σ), and the value of K_i is selected so that we obtain a fixed number of convolved images per octave. Then the Difference-of-Gaussian images are taken from adjacent Gaussian-blurred images per octave. Once DoG images have been obtained, key points are identified as local minima/maxima of the DoG images across scales. This is done by comparing each pixel in the DoG images to its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the neighboring scales. If the pixel value is the maximum or minimum among all compared pixels, it is selected as a candidate keypoint.

2) Keypoint Localization: Scale-space extrema detection produces too many key point candidates, some of which are unstable. The next step in the algorithm is to perform a detailed fit to the nearby data for accurate location, scale, and ratio of principal curvatures. This information allows points to be rejected that have low contrast (and are therefore sensitive to noise) or are poorly localized along an edge.

3) Orientation Assignment: In this step, each keypoint is assigned one or more orientations based on local image gradient directions. This is the key step in achieving invariance to rotation as the keypoint descriptor can be represented relative to this orientation and therefore achieves invariance to image rotation.

4) Keypoint descriptor: The previous step ensured invariance to image location, scale, rotation. Our aim is to compute a descriptor vector for each keypoint such that the descriptor is highly distinctive and partially invariant to the remaining variations such as illumination, 3D viewpoint, etc. This step is performed on the image closest in scale to the keypoint's scale. First a set of orientation histograms is created on 4x4 pixel neighborhoods with 8 bins each. These histograms are computed from magnitude and orientation values of samples in a 16 x 16 region around the keypoint such that each histogram contains samples from a 4 x 4 sub region of the original neighborhood region. The magnitudes are further weighted by a Gaussian function with equal to one half the width of the descriptor window. The descriptor then becomes a vector of all

the values of these histograms. Since there are 4 x 4 = 16 histograms each with 8 bins the vector has 128 elements. This vector is then normalized to unit length in order to enhance invariance to affine changes in illumination. To reduce the effects of nonlinear illumination a threshold of 0.2 is applied and the vector is again normalized.

Further wavelet transform, characterize texture by the statistical distribution of the image intensity. The feature vector obtained is then combined into single feature vector which is then led to SVM [8].

9 CONCLUSION

There have been many methodologies that proved to be best in context with CBIR. But those approaches does not reduce semantic gap and meet human perception completely. Many experiments fail in reading user's mind. By owing to the approach of combing saliency map detection with SIFT Algorithm can lead to adequate efficiency of image retrieval. Also the addition of energy level of image as a feature supports the model to derive fruitful image that actually meets the human demand and perception. The proposed model is trying to improve speed and accuracy of Content Based Image Retrieval (CBIR).

REFERENCES

[1] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis."IEEE Transactions on pattern Analysis and Machine Intelligence, vol.20, p. 1254-1259, November 1998.

[2] K. Ashok Kumar & Y.V. Bhaskar Reddy "Content Based Image Retrieval Using SVM Algorithm," International Journal of Electrical and Electronics Engineering (IJEEE) ISSN (Print):2231-5284, Vol-1, Iss-3, 2012.

[3] J. Harel, C. Koch and P. Perona, "Graph-Based Visual Saliency," Proceedings of Neural Information Processing Systems (NIPS), 2006.

[4] Lei Zhang, Fuzong Lin, Bo Zhang,"Support Vector Machine Learning for Image Retrieval," IEEE Transactions 0-7803-6725-1/01/$10.00,2001.

[5] Bakar, Hitam, Wan Yussof ,"Content Based Image Retrieval using SIFT for binary and grayscale images", International Conference on Signal and Image Processing Applications (ICSIPA), IEEE, 2013.

[6] Monica Bishops, Institute of Mathematical Machines, Warsaw, "Bottom up Saliency Maps-a review", Electronics, July 2013.

[7] "Saliency Map Tutorial", www.ntu.edu.tw, June 2012.

[8] S. Murala, A. B. Gonde, R. P. Maheshwari, "Color and Texture Features for Image Indexing and Retrieval", International Advanced Computing Conference (IACC), IEEE,2009.

[9] Vanitha. L. and Venmathi.A.R,"Classification of Medical Images Using Support Vector Machine" IPCSIT vol.4 (2011) © (2011).

[10] Mamta Kamath, Disha Punjabi, Tejas Sabnis, Divya Upadhyay, Seema Shrawne, "Improving content based Image Retrieval Using Scale Invariant Feature Transform",International Journal of Engineering and Advanced Technology (IJEAT), ISSN:2249-8958,Volume-1,Issue-5,June 2012.

[11] Kimaya S. Meshram, Ajay M. Agarkar," Content based Image Retrieval System Using SIFT: A Survey SSRG-IJECE-Volume-2 Issue-10, October 2015.

[12] Xuefei Bai, Wenjian Wang, "Saliency-SVM: An automatic approach for image segmentation, Elsevier, 2014.