INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume2, Issue6)
Available online at: www.Ijariit.com

An Enhanced Image Descriptor Algorithm for Image Retrieval using SIFT - PCA

Dr. R. Malini*
malini.raju2@gmail.com

Abstract— Due to the growth of internet and social media sites there has been a proliferation of images which makes image retrieval a challenging job. The proposed an Enhanced Image Descriptor Algorithm for Image Retrieval using SIFT-PCA involves three steps: feature detection (Harris Detector), feature description (SIFT), dimension reduction (PCA). The features of all images in database are retrieved and stored in feature database in the form of histogram. When a query image is given, the generated histogram of query image is matched with feature database of all image histograms and those images with minimum distance are retrieved. Euclidean distance is used as the distance measure. For analyzing the performance of the algorithm, average precision and recall are used. Precision is the measure of ability of a system to present all the relevant items and Recall is defined as the ability to present only relevant items.

Keywords— Euclidean distance, Harris Detector, Principal Component Analysis, SIFT, Feature Descriptor.

I. INTRODUCTION

Invention of the digital camera has given the common person the privilege to capture his world in pictures, and conveniently share them with others. One can today generate volumes of images with content as diverse as family get-togethers and national park visits. Low-cost storage and easy Web hosting has fuelled the metamorphosis of common man from a passive consumer of photography in the past to a current-day active producer. Today, searchable image data exists with extremely diverse visual and semantic content, spanning geographically disparate locations, and is rapidly growing in size. All these factors have created innumerable possibilities and hence considerations for real-world image search system designers.

As far as technological advances are concerned, growth in content-based image retrieval has been unquestionably rapid. In recent years, there has been significant effort put into understanding the real world implications, applications, and constraints of the technology. Yet, real-world application of the technology is currently limited. Content based image retrieval is a set of techniques for retrieving semantically-relevant images from an image database based on automatically-derived image features Content-based image retrieval relies on the characterization of primitive features such as color, shape, and texture that can be automatically extracted from the images themselves. The main goal of CBIR is efficiency during image indexing and retrieval, thereby reducing the need for human intervention. The computer must be able to retrieve images from a database without any human assumption on specific domain.

The disadvantages of Text based Image Retrieval [1] is that the precision is usually unsatisfactory because of the semantic gap between low-level visual features and high-level semantic concepts. Secondly, the efficiency usually low due to the high dimensionality of features. Thirdly, the query form of text based image retrieval is unnatural owing to the possible absence of appropriate example images. In order to overcome these drawbacks, Content Based Image Retrieval system came into existence.

II. LITERATURE SURVEY

In 2011, Ce Liu Jenny Yuen and Antonio Torralba proposed Nonparametric Scene Parsing Label Transfer via Dense Scene Alignment [2], which is a nonparametric approach for object recognition, and scene parsing using dense scene alignment. Given an input image, its best matches are retrieved from a large database with annotated images using coarse-to-fine SIFT flow algorithm that aligns the structures within two images. Based on the dense scene correspondence obtained from the SIFT flow, this system warps the existing annotations, and integrates multiple cues in a Markov random field framework to segment and recognize the query image.

In 2011, Mining Social Images with Distance Metric Learning for Automated Image Tagging [3] was proposed by Pengcheng Wu et all as a machine-learning framework for mining social images and investigate its application to automated image tagging. The paper puts forward a novel Unified Distance Metric Learning (UDML) scheme, which not only exploits both visual and textual contents of social images, but also effectively unifies both inductive and transductive metric learning techniques in a systematic learning framework. An emerging retrieval-based annotation paradigm was developed for automated photo tagging by mining massive social images freely available on the web. Unlike traditional web images, social images often contain tags and rich user-generated contents, which offer a new opportunity to resolve some long-standing challenges in multimedia, for instance the semantic gap. The idea of the retrieval-based paradigm is to first retrieve a set of k most similar images for a test photo from the social image repository, and then to assign the test photo with a set of t most relevant tags associated with the set of k retrieved social images.

III. PROPOSED ENHANCED IMAGE DESCRIPTOR ALGORITHM FOR IMAGE RETRIEVAL USING SIFT - PCA

In an Image Retrieval system, the very important factor to be considered is the proper extraction of image features. To define an image, the visual features of image have to be extracted and are described using a descriptor [4]. When a query is given, the algorithm will check the database of images to find an appropriate match.

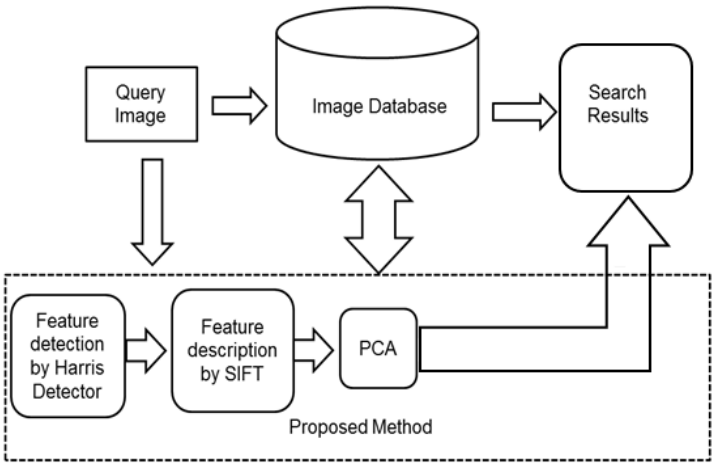


Fig. 1 Proposed Content Based Retrieval System

For feature extraction, Scale Invariant Feature Transform (SIFT) is used which extracts the local features of the image. Since the SIFT algorithm produces a large quantity (128 dimensional feature descriptor) of data, Principal Component Analysis (PCA)[5] is used for mapping these features to a low dimensional space where the volume of data decreases considerably. Now, when a query image is given, the feature description followed by PCA is performed and is compared with database of histogram and the images with minimum distance are retrieved. Euclidean distance is used as the distance measure.

The SIFT approach, for image feature generation, takes an image and transforms it into a "large collection of local feature vectors"[6][7]. Each of these feature vectors is invariant to any scaling, rotation or translation of the image. Following are the major stages of computation used to generate the set of image features:

1. Scale-space extrema detection: The first stage of computation searches over all scales and image locations. It is implemented efficiently by using a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation.

2. Key point localization: At each candidate location, a detailed model is fit to determine location and scale. Key points are selected based on measures of their stability.

3. Orientation assignment: One or more orientations are assigned to each key point location based on local image gradient directions. Operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations.

4. Key point descriptor: The local image gradients are measured at the selected scale in the region around each key point. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination. This approach

has been named the Scale Invariant Feature Transform (SIFT), as it transforms image data into scale-invariant coordinates relative to local features.

A. Detection of Scale -Space Extrema

First step is to detect key points using a cascade filtering approach that uses efficient algorithms to identify candidate locations. Key point detection is to identify locations and scales that can be repeatedly assigned under differing views of the same object. Detecting locations that are invariant to scale change of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as scale space (the scale-space kernel is the Gaussian function). Therefore, the scale space of an image is defined as a function, L(x, y,σ) that is produced from the convolution of a variable-scale Gaussian, G(x, y, σ), with an input image, I(x, y):

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

Where $*$ is the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2)$$

Fig.2 Gaussian Blurred Images

To efficiently detect stable key point locations in scale space, use the scale-space extrema in the difference-of-Gaussian function convolved with the image, D(x, y,σ), which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k:

$$D(x, y, \sigma) = (G(x, y, k) - G(x, y, \sigma)) * I(x, y) \quad (3)$$

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (4)$$

There are a number of reasons for choosing this function. First, it is a particularly efficient function to compute, as the smoothed images, L, need to be computed in any case for scale space feature description, and D can therefore be computed by simple image subtraction.

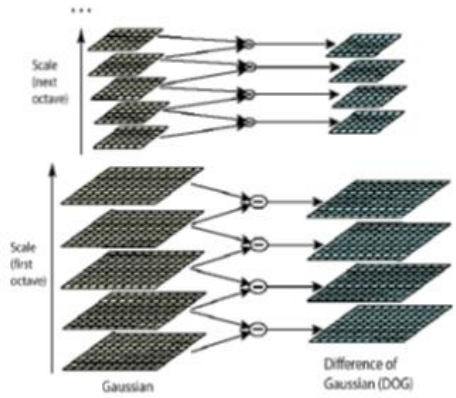


Fig.3 Generation of Difference of Gaussian

The maxima and minima of $\sigma^2 \nabla^2 G$ produce the most stable image features compared to a range of other possible image functions, such as the gradient, Hessian, or Harris corner function. The relationship between D and $\sigma^2 \nabla^2 G$ can be understood from the heat diffusion equation.

$$\frac{\partial D}{\partial x} = \sigma^2 \nabla^2 G \quad (5)$$

This shows that when the difference-of-Gaussian function has scales differing by a constant factor it already incorporates the σ2 scale normalization required for the scale-invariant Laplacian. In order to detect the local maxima and minima of D(x, y, σ), each sample point is compared to its eight neighbours in the current image and nine neighbours in the scale above and below. It is selected only if it is larger than all of these neighbours or smaller than all of them. Each image was then subject to a range of transformations, including rotation, scaling, affine stretch, change in brightness and contrast, and addition of image noise. Because the changes were synthetic, it was possible to precisely predict where each feature in an original image should appear in the transformed image, allowing for measurement of correct repeatability and positional accuracy for each feature.

B. Accurate Key point Localization

Once a key point candidate has been found by comparing a pixel to its neighbours, the next step is to perform a detailed fit to the nearby data for 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.

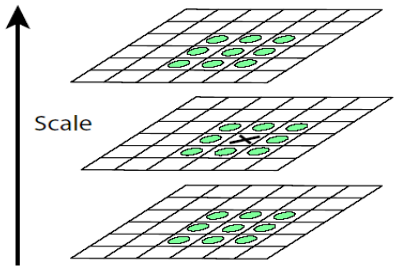


Fig.4 Key point localization

To locate maxima or minima the Taylor expansion (up to the quadratic terms) of the scale-space function, D(x,y,σ) shifted so that the origin is at the sample point:

$$D(X) = D + \frac{\partial D^T}{\partial X}X + \frac{1}{2}X^T\frac{\partial^2 D}{\partial X^2}X \quad (6)$$

Where D and its derivatives are evaluated at the sample point and X = (x, y, σ)T is the offset from this point. The location of the extremum, $\hat{X}$, is determined by taking the derivative of this function with respect to x and setting it to zero, giving

$$\hat{X} = -\frac{\partial^2 D^{-1}}{\partial X^2}\frac{\partial D}{\partial X} \quad (7)$$

The function value at the extremum, D($\hat{X}$), is useful for rejecting unstable extrema with low contrast.

For stability, it is not sufficient to reject key points with low contrast. The difference-of- Gaussian function will have a strong response along edges, even if the location along the edge is poorly determined and therefore unstable to small amounts of noise. A poorly defined peak in the difference-of-Gaussian function will have a large principal curvature across the edge but a small one in the perpendicular direction. The principal curvatures can be computed from a 2x2 Hessian matrix, H, computed at the location and scale of the key point.

$$H = \sum_{x,y}\omega(x,y)\begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \quad (8)$$

The derivatives are estimated by taking differences of neighboring sample points. The eigenvalues of H are proportional to the principal curvatures of D. Let α be the eigenvalue with the largest magnitude and β be the smaller one. Then, we can compute the sum of the eigenvalues from the trace of H and their product from the determinant:

Trace of Matrix, $Tr(M) = I_x^2 + I_y^2 = \alpha + \beta$ (9)

Determinant, $Det(M) = I_x^2I_y^2 - (I_xI_y)^2 = \alpha\beta$ (10)

Let r be the ratio between the largest magnitude eigenvalue and the smaller one, so that $\alpha\beta = r$. Then,

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \quad (11)$$

which depends only on the ratio of the eigenvalues rather than their individual values. The quantity $\frac{2(r+1)}{r}$ is at a minimum when the two eigenvalues are equal and it increases with r.

C. Orientation Assignment

By assigning a consistent orientation to each keypoint based on local image properties, the keypoint descriptor can be represented relative to this orientation and therefore achieve invariance to image rotation. The scale of the keypoint is used to select the Gaussian smoothed image, L, with the closest scale, so that all computations are performed in a scale-invariant manner. For each image sample, L(x, y), at this scale, the gradient magnitude, m(x, y), and orientation, θ(x, y), is pre-computed using pixel differences:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad (12)$$

$$\theta(x,y) = \tan\left((L(x,y+1) - L(x,y-1))/L(x+1,y) - L(x-1,y)\right) \quad (13)$$

An orientation histogram is formed from the gradient orientations of sample points within a region around the keypoint. The orientation histogram has 36 bins covering the 360 degree range of orientations. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window with a σ that is 1.5 times that of the scale of the keypoint. Peaks in the orientation histogram correspond to dominant directions of local gradients.

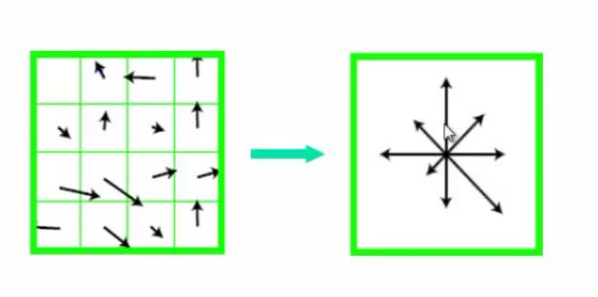


Fig.5 Key point Histogram

The highest peak in the histogram is detected, and then any other local peak that is within 80% of the highest peak is used to also create a key point with that orientation. Therefore, for locations with multiple peaks of similar magnitude, there will be multiple key points created at the same location and scale but different orientations. Only about 15% of points are assigned multiple orientations, but these contribute significantly to the stability of matching. Finally, a parabola is fit to the 3-histogram values closest to each peak to interpolate the peak position for better accuracy.

D. Image Descriptor

The previous operations have assigned an image location, scale, and orientation to each keypoint. These parameters impose a repeatable local 2D coordinate system in which to describe the local image region, and therefore provide invariance to these parameters. The next step is to compute a descriptor for the local image region that is highly distinctive yet is as invariant as possible to remaining variations, such as change in illumination or 3D viewpoint.

Figure 6 illustrates the computation of the key point descriptor. First the image gradient magnitudes and orientations are sampled around the key point location, using the scale of the key point to select the level of Gaussian blur for the image. In order to achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the key point orientation. For efficiency, the gradients are precomputed for all levels of the pyramid. These are illustrated with small arrows at each sample location on the left side of Figure 6.

A Gaussian weighting function with σ equal to one-half the width of the descriptor window is used to assign a weight to the magnitude of each sample point. This is illustrated with a circular window on the left side of Figure 6, although, of course, the weight falls off smoothly. The purpose of this Gaussian window is to avoid sudden changes in the descriptor with small changes in the position of the window, and to give less emphasis to gradients that are far from the centre of the descriptor, as these are most affected by mis registration errors. It is important to avoid all boundary affects in which the descriptor abruptly changes as a sample shifts smoothly from being within one histogram to another or from one orientation to another. Therefore, tri-linear interpolation is used to distribute the value of each gradient sample into adjacent histogram bins.

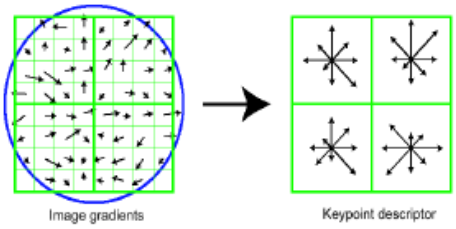


Fig. 6 SIFT Feature Descriptor

Finally, the feature vector is modified to reduce the effects of illumination change. First, the vector is normalized to unit length. A change in image contrast in which each pixel value is multiplied by a constant will multiply gradients by the same constant, so this contrast change will be cancelled by vector normalization. A brightness change in which a constant is added to each image pixel will not affect the gradient values, as they are computed from pixel differences. Therefore, the descriptor is invariant to affine changes in illumination. However, non-linear illumination changes can also occur due to camera saturation or due to illumination changes that affect 3D surfaces with differing orientations by different amounts. These effects can cause a large change in relative magnitudes for some gradients, but are less likely to affect the gradient orientations. Therefore, reduce the influence of large gradient magnitudes by thresholding the values in the unit feature vector to each be no larger than 0.2, and then renormalizing to unit length.

E. Principal Component Analysis

In computational terms the principal components are found by calculating the eigenvectors and eigenvalues of the data covariance matrix. This process is equivalent to finding the axis system in which the co-variance matrix is diagonal. The eigenvector with the

largest eigenvalue is the direction of greatest variation, the one with the second largest eigenvalue is the (orthogonal) direction with the next highest variation and so on.

Let A be a n x n matrix. The eigenvalues of A are defined as the roots of:

Determinant (A-λI)=|(A-λI)=0 (14)

Where, I is the n x n identity matrix. This equation is called the characteristic equation (or characteristic polynomial) and has n roots.

Let λ be an eigenvalue of A. Then there exists a vector x such that:

Ax=λx (15)

The vector x is called an eigenvector of A associated with the eigenvalue λ. There is no unique solution for x in the above equation. It is a direction vector only and can be scaled to any magnitude. To find a numerical solution for x, set one of its elements to an arbitrary value, say 1, which gives a set of simultaneous equations to solve for the other elements. If there is no solution, repeat the process with another element. Now normalise the final values so that x has length one ie; x xT = 1.

If the eigenvectors of the co-variance matrix was calculated, the direction vectors are calculated which are indicated by ϕ1 and ϕ2. Putting the two eigenvectors as columns in the matrix ϕ = [ϕ1, ϕ2], create a transformation matrix which takes our data points from the [x1, x2] axis system to the axis [ϕ1, ϕ2] system with the equation:

$$p_{\phi} = (p_x - \mu_x)\,\phi \quad (16)$$

Where px is any point in the [x1, x2] axis system, μ_(x) = (μx1, μx2) is the data mean, and pϕ is the coordinate of the point in the [ϕ1, ϕ2] axis system.

Algorithm for PCA

Transform an $N \times d$ matrix X into an $N \times m$ matrix Y:

1. Centralize the data (subtract the mean).

2. Calculate the $d \times d$ covariance matrix:

$$C = \frac{1}{N-1} X^T X \quad (17)$$

$$C_{i,j} = \frac{1}{N-1} \sum_{q=1}^{N} X_{q,i} . X_{q,j} \quad (18)$$

$C\ i$, i (diagonal) is the variance of variable i and $C\ i$, j (off-diagonal) is the covariance between variables i and j.

3. Calculate the eigenvectors of the covariance matrix (orthonormal).

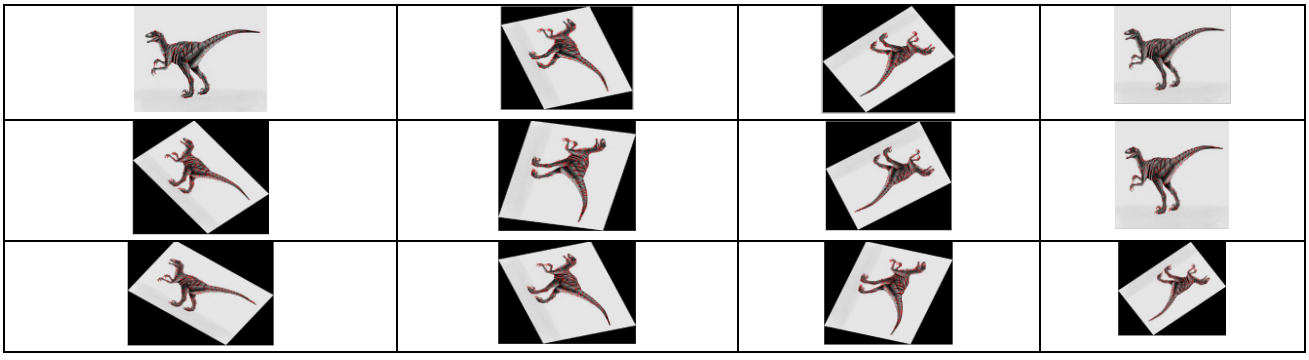
4. Select m eigenvectors that correspond to the largest m eigenvalues to be the new basis

After feature detection, several local patches abstract each image. SIFT converts each patch to 128-dimensional vector. After this step, each image is a collection of vectors of the same dimension (128 for SIFT. This 128-dimensioned feature vector is reduced in dimension using Principal Component Analysis (PCA).

IV. IMPLEMENTATION RESULTS

The implementation of proposed Enhanced Image descriptor algorithm for Image Retrieval using SIFT-PCA is simulated using MatlabR2010. The Wang database is used for the simulation of the project. In Wang database [9], there are 10 sets of image, each containing 100 images in 1000 images.

Harris Detector is the detection stage of the SIFT algorithm. It gives the valid corner points of the objects in the image. The red coloured points are the detected Harris Features. After detection, different orientations of image are checked. And Finally, the feature points which are invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint are described using a 128 long vector called SIFT descriptor.



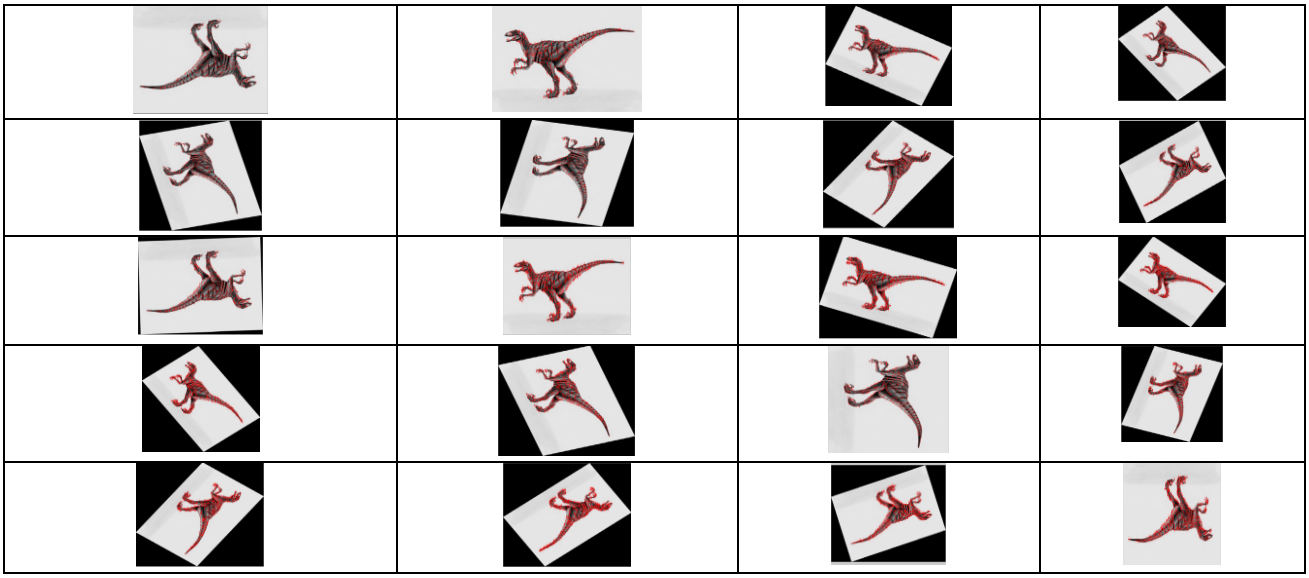


Fig. 7 SIFT output checking for different orientations

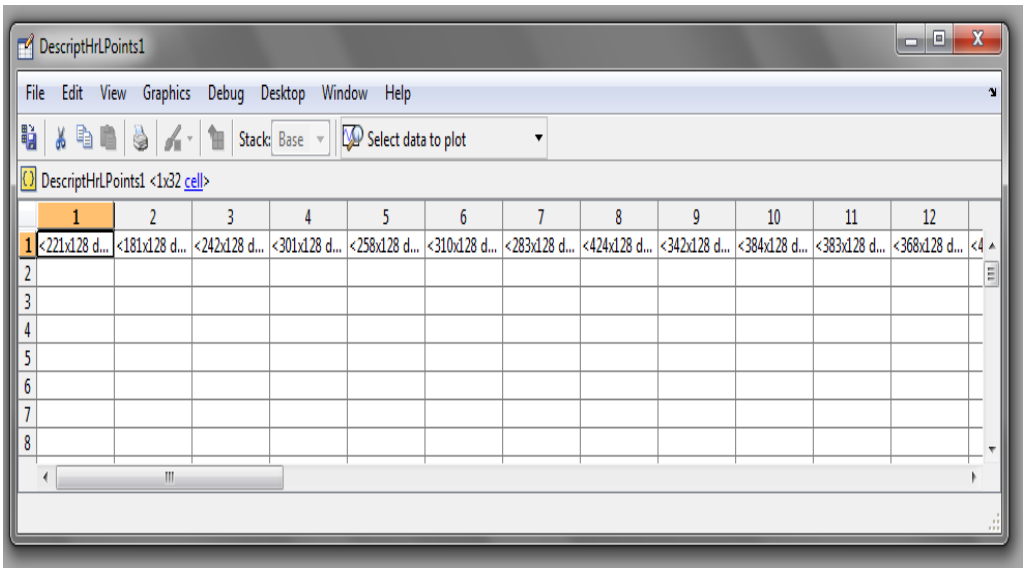


Fig.8 128 long SIFT descriptor

PCA finds a linear projection of high dimensional data into a lower dimensional subspace, ie; it transforms an $N \times d$ matrix X into an $N \times m$ matrix Y. Output of SIFT contains approximately 32x128x200 = 8, 19,200 values which consumes Mega Bytes of memory. In order to reduce the dimension of the feature vector, PCA is used. It uses redundancy property of data for dimension reduction. By doing this, memory required to store the feature descriptor can be reduced considerably. PCA output is more distinctive, more robust to image deformations, and more compact than the standard SIFT representation. It results in increased accuracy and faster matching [8]. Figure 8 below demonstrates the dimension reduction applied to the SIFT descriptor.

Input Image

Feature points of image

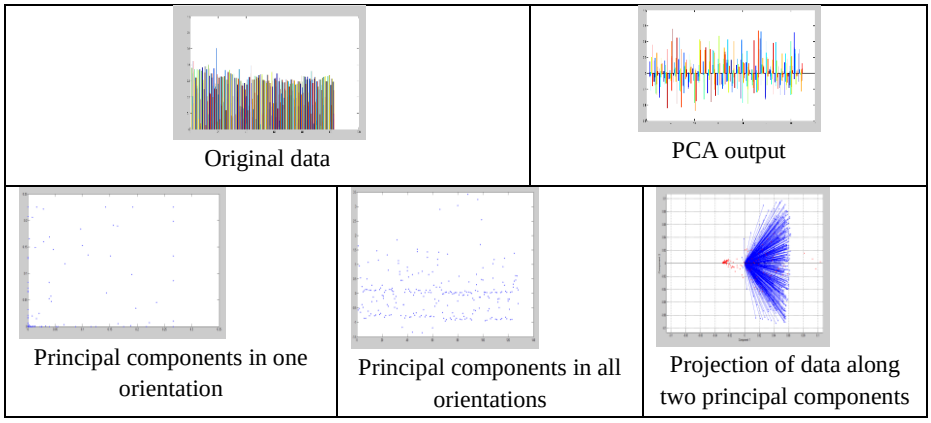


Fig.9 PCA output

Performance of Content based image retrieval algorithm using SIFT-PCA was evaluated using precision and recall. For finding the precision and recall, confusion matrix is used. Confusion matrix is a table, which gives the information about whether the retrieved image belongs to the category, or not. From confusion matrix, using the equations 2 and 3, the performance of the system was analyzed calculating precision and recall. Precision is defined as the measure of a system to present only the relevant items whereas recall is the measure of ability of a system to present all relevant items.

TABLE I
PERFORMANCE MEASURES FOR EACH CATEGORY

Category	Recall	Precision
Beach	70%	100%
Building	60%	96%
Bus	60%	96%
Dinosaurs	70%	100%
Elephant	60%	96%

From the Table 1, average recall is found to be 64% and precision as 97.6%, which is comparatively higher than existing retrieval techniques. That is 64% of total relevant images in the collection is retrieved and with 97.6% precision.

V. CONCLUSION

In this project, An Enhanced Image descriptor algorithm for Image Retrieval using SIFT-PCA was proposed. For the design of an efficient retrieval system, the first stage is to identify the best feature extraction algorithm. The project uses SIFT combined with PCA for feature description. SIFT-PCA provides more distinctive, more robust to image deformations, and more compact feature description than the standard SIFTS representation. It also offers increased accuracy and faster matching. The time and memory required for the SIFT features can be considerably reduced using this algorithm.

When a query image is given, the retrieval algorithm performs the feature detection, description, dimension reduction and histogram formation and matches the query histogram with that of database. The images with minimum Euclidean distance are retrieved. The algorithm offers an average precision of 97.6% and recall rate 64%, which is better than existing algorithms. Also the length of SIFT descriptor is reduced considerably.

REFERENCES

[1] Lei Wu, Rong Jin, and Anil K. Jain, Tag Completion for Image Retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(3),2013.

[2] C. Liu, J. Yuen, and A. Torralba, Nonparametric Scene Parsing via Label Transfer, IEEE Trans. Pattern Analysis and Machine Intelligence, 33(12):2368-82,2011.

[3] P. Wu, S.C.H. Hoi, P. Zhao, and Y. He, Mining Social Images with Distance Metric Learning for Automated Image Tagging, Proc. Fourth ACM Int'l Conf. Web Search and Data Mining, pp.197- 206,2011.

[4] Y. Guo and S. Gu, Multi-Label Classification using Conditional Dependency Networks, Proc. 22nd Int'l Joint Conf. Artificial Intelligence, pp. 1300-1305,2011.

[5] Yang Song, Lu Zhang and C. Lee Giles, Automatic Tag Recommendation Algorithms for Social Recommender Systems, ACM Transactions on Computational Logic,5:1–22,2008.

[6] James Z. Wang, Jia Li, Gio Wiederhold, Simplicity: Semantics-Sensitive Integrated Matching for Picture Libraries, IEEE Transaction on Pattern Analysis and Machine Intelligence, 23(9): 947-963,2001.

[7] David G. Lowe, Distinctive Image Features from Scale-invariant Key points, Computer Science Department, University of British Columbia, lowe@cs.ubc.ca.

[8] Yan Ke1, Rahul Sukthankar, PCA-SIFT: A More Distinctive Representation for Local Image Descriptors, School of Computer Science, Carnegie Mellon University; 2 Intel Research Pittsburgh, http://www.cs.cmu.edu/˜yke/pcasift/.

[9] P. Tirilly, V. Claveau, and P. Gros, Language Modelling for Bag of- Visual Words Image Categorization, Proc. Int'l Conf. Content Based Image and Video Retrieval, pp. 249-258,2008.

