Automatic Image Annotation using SURF Features

Tuhin Shukla
School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, Madhya Pradesh, India

Nishchol Mishra
School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, Madhya Pradesh, India

Sanjeev Sharma, PhD.
School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, Madhya Pradesh, India

ABSTRACT
Automatic image annotation is a challenging field with a far reaching effect. As the world moves towards becoming more and more dependent on digital technologies every day, use of machine to automatically annotate images can be proved as demanding in many fields of image processing. Automatic Image Annotation reduces the gap between low level image features and high level image semantics. Utilization of Speeded Up Robust Features (SURF) in automatic image annotation is very appealing due to the fact that SURF is scale and rotation invariant detector and descriptor and is much faster than any other schemes. Unlike other methods SURF features use the entire image instead of segmented blocks of image. That is why annotation of images by using SURF can be considered as more accurate. In this paper, a SVM based image annotation approach is proposed that uses SURF features of image for annotation purpose. The experiments suggest that the method proposed is much more efficient than other methods.

General Terms
Pattern Recognition, Image Annotation.

Keywords
Image retrieval; Machine learning; Semantic gap; Image annotation; SURF; SVM

1. INTRODUCTION
An Image can be defined as a two dimensional function f(x, y) where x and y are coordinate values in plane and the amplitude of function f at any point in plane is called intensity value of image at that particular point and is always positive [1]. Though it is easy to represent data in the form of an image but to process the same data in pictorial form is very difficult. It involves a lot of processing and use of intelligent and advanced systems. Digital image processing is an advanced area that deals with digital images.

Digital images are now at the forefront in our increasing machine dependent world. Digital photography has become a universal medium of obtaining and sharing information due to convenient and easy access of internet. That is why unique databases of image and videos are available on the web and can be accessed easily. In present scenario visual data is very common on the web and the heap of digital images on the web is increasing minute by minute with very rapid rate. So there is a need to find efficient tools that can search and provide desired visual data on demand with sufficient accuracy. This causes many researchers to pay their attention to develop efficient image retrieval techniques.

In last two decades Image Retrieval (IR) has been a hot research area. Research in image retrieval is divided into two broad categories.

- CBIR- Content Based Image Retrieval is the branch of image retrieval that focuses on the contents of image for searching purpose. In CBIR, low level content features are used for example color and texture. Search results depend on the best possible matching of feature vector extracted from query image. The only problem with techniques under this category is that users are not concerned about such low level features as they can't interpret images based on such features. They are more comfortable with natural languages.
- AIA- Automatic Image Annotation (AIA) is another branch of image retrieval that can be said as more user friendly. Users are much at ease if images are given with semantic keywords but manual indexing is a time consuming process. That is why techniques under this category first annotate images with semantic key words automatically and once images are annotated they can be retrieved much more easily. Generating fully automatic annotation systems is still the subject of extensive research.

Image annotation refers to labelling of images with a set of predefined keywords based on contents of image. This can be helpful for filling the huge gap between low-level features obtained from the image and high-level semantics derived from image. The basic concept of image annotation is to automatically learn about semantic concepts from large number of sample images and use these concepts to label new images [4] and as images are already annotated with labels so they can be easily retrieved on the basis of keywords.

In today's modern world AIA covers a wide range of application areas. AIA is mainly used for maintaining large databases of image and video files and later retrieving images from their image collections. Personal digital image collection cultural heritage collections television archives, satellite imagery, medical imaging and many other related fields make use of automatic image annotation. Traditionally manual annotation has been used for databases having large collections of images. But manual annotation is very time consuming and very costly process. That is why AIA has become a research area having extremely high interest.

Automatic image annotation is a challenging task due to various imaging conditions, complex and hard-to-describe objects, a highly textured background and occlusions [5]. Usually images can be automatically annotated in two different ways. First they can be annotated using learning based techniques that lead to train images categorized manually and label the uncategorised images based on training results. Second type of techniques uses relevance feedback from user to annotate images. Refinement of results is taken place by asking for feedback in several rounds. As the

first approach gives us higher accuracy the second is more effective. But first approach has a limited training image so it is unable to effectively cover all real life aspects; the second approach becomes burden to users as usually more feedback is needed.

The rest of the paper is organized as follows. In Section 2, we give an overview of related work. In Section 3, the proposed approach is given. In Section 4, the experimental results have been provided. The last section 5 gives the final conclusions.

2. RELATED WORK

Zhang et al. have provided a comprehensive study on automatic image annotation techniques [4]. This is a vast survey on AIA methods where they have classified AIA techniques into categories such as SVM, ANN, DT, non-parametric and parametric approach and annotation incorporating metadata.

Chapelle et al [13] showed that classification can be improved based on image histograms using support vector machines (SVM). Before this, it was known that classification approaches generalised poorly on classification tasks if the dimensionality of the feature space was extremely high but this approach showed that SVM can perform this classification easily if the only attributes provided are high-dimensional histograms. Chapelle et al used heavy-tailed RBF kernels. Also, they showed that decreasing a while using a-exponentiation improves performance of linear SVM that they can be used to substitute RBF kernels.

Cusano et al [10] gave an image annotation tool that classifies image region in one of seven classes. This tool has been provided to maintain huge image and video databases. The seven classes that are deemed are sky, skin, vegetation, snow, water, ground and buildings. The tool proposed uses as input tiles of an image by computing subdivisions in form of square which is the size of fixed fraction of total area of an image. A multiclass SVM is used for classification using "one per class" approach.

Shi et al. [9] gave an adaptive content representation scheme. This encompassed two main parts: (i) adaptive visual representation of image contents; and (ii) adaptive two-level segmentation method. To sufficiently represent the contents of image they used texture of image as features centred on matching pursuit algorithm coupled with color histograms. To segment an image into meaningful regions, at diverse levels segmentation is done. At global level segmentation is done using color, texture and position i.e. global features of an image. At local level, segmentation is done using adaptive matching point features of an image. Also at global level, Gaussian Mixture Model is used while at local level, K-means algorithm is used for segmentation.

Lei et al. [5] have given an automatic image annotation technique which incorporates both Hidden Markov Model and Support vector Machines. Using two kinds of HMM and keyword correlation this approach gives a better result. This approach uses a two-staged mapping model. At the first, two hidden markov models are used for classifying color and texture features separately. Then SVM is used to classify results and give the final annotations of the image.

Qi et al [6] have used multiple SVMs for automatic image annotation. The system gives the concept of combining multiple instance learning (MIL) based and global feature based SVMs. In this system, each image is divided into blocks so that MIL method could be used to extract features based upon color and texture of the block. The efficiency is boosted by utilizing an enhanced diversity density method and a fast searching algorithm to accurately extract features. These features called bag features are then given as input to a set of SVMs to annotate. Another set of SVMs are trained using global color and edge histogram based features. This second set allows for removal of any inaccuracy issues related to the first set. From any test image, features both bag as well as global can be constructed so that they can be fed to their respective set of SVMs .The output received from these two harmonizing SVMs is then integrated by an automatic weight estimation method to give final results of annotation.

Goh et al. [8] use one-class, two-class and multiclass SVMs. They have proposed a confidence -based dynamic ensemble (CDE) so that it can be concluded when retraining of classifier is needed and whether new low-level features or training data can be included. A three level classification scheme is proposed. At the base level, SVMs are used for computing the prediction of one semantic label. A confidence factor is given for each prediction by employing algorithm for one-class SVMs which also uses a density distribution of training data. At multiclass level, the confidence factors of all multiple classifiers are cumulated to give only one prediction. Again a multi class level confidence factor is computed for this prediction. At the bag level, CDE cumulates the predictions from multiple bags to give an aggregated prediction. An overall confidence factor is given at this level. If this is high, a semantic is assigned. This approach overcome the disadvantages of traditional static classifiers as it makes adjustments to include semantics leading to discovery of low level features and thus improving accuracy.

3. PROPOSED WORK

In the algorithm, the key points and descriptors of all training images are extracted. Then these descriptors are clustered into N centroids. The K-means clustering algorithm is used for executing this procedure. The key notion that has been used in this paper is that extracted descriptors are independent and can be used as Bag- of –Words (BoW) in the image. The multiclass SVMs are trained on basis of the BoW.

For an image to be queried, descriptors are extracted. The dictionary formed from BoW is used as basis to map each descriptor to its equivalent visual word. This gives us a subsequent tally for the image to be queried. This result is passed to SVM to classify and annotate the image.

The proposed framework is shown in fig.1.

There are certain modules involved in the algorithm which are:

- Computation of SURF descriptors.
- Compute histograms based on bag-of-words
- Classification using SVM
- Annotation of images

The following sections present a concise description of each step.

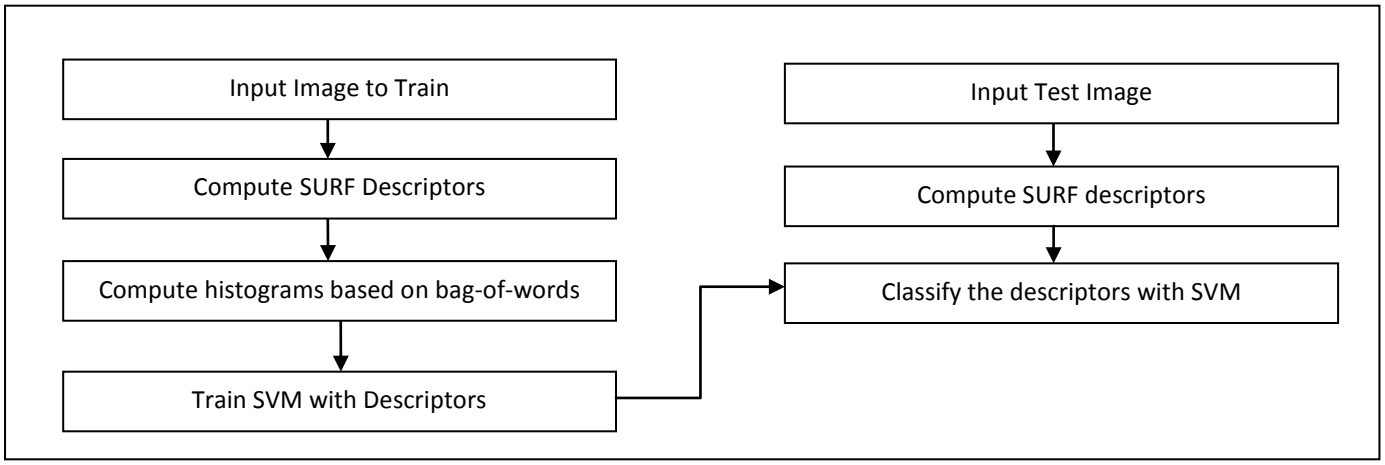


Figure 1: Proposed Framework

3.1 SURF Feature Descriptor

As features in an image can be found very easily now due to distinctive use of descriptors that can be computed on the whole image rather than its segmented parts, it has become easier to be more accurate. SIFT, PCA-SIFT, SURF are some such descriptors. One of the very reasons for their popularity is that they are invariant to image rotation, scaling, changes in illumination.

The reason SURF is preferred over SIFT is due to its concise descriptor length. Whereas the customary SIFT implementation uses a descriptor consisting of 128 floating point values, SURF compresses this descriptor length to 64 floating point values.

3.1.1. Interest Point Localization

The SURF detector is based on the Hessian matrix. Given a point X = (x, y) in an image I, the Hessian matrix H(X, σ) at X at scale σ is defined as follows:

$$H(X, \sigma) = \begin{pmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{pmatrix} \quad (1)$$

Where $L_{xx}(X, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I at point X, and similarly for $L_{xy}(X, \sigma)$ and $L_{yy}(X, \sigma)$[7].

Hessian based detectors are more stable and repeatable. Also, approximation like Difference of Gaussians (DoG) can decrease speed at low cost. As integral images are used, the computation time is decreased. The range of interest point is kept in a 3x3x3 neighborhood so that interest point can be localized in scale and image space by using non maximum suppression.

3.1.2 Interest Point Descriptor

In order to assign a unique orientation around the detected interest point, SURF constructs a circular region. This also gives invariance to image rotations. The first step in orientation assignment is the calculation of Haar wavelet response in both x and y directions. The Haar wavelets are big in size at high scales. Hence, integral images for fast filtering [7]. As soon as the dominant orientation is approximated and accommodated in the interest point information, the next step is extracting the square region around the interest points. The regions are split up in 4x4 sub-regions. The underlying intensity pattern (first derivatives) of each sub-region is described by a vector

$$V = [\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|] \quad (2)$$

In the proposed approach, the interest points were detected and described using the descriptors for each individual image. Figure 2 shows how SURF features were detected in every individual image.

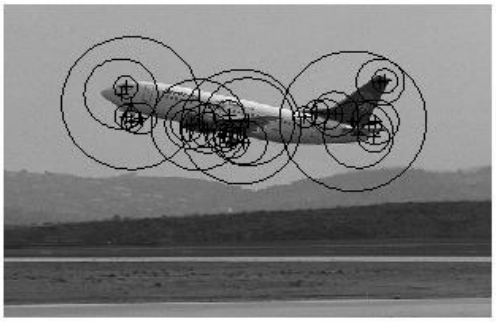


Figure 2: An example to show how SURF points are detected and plotted

3.2 Bag-of-Words

Bag-of-word model is used in image classification and document classification. While in former it gives us a bag of visual words which are a sparse vector of frequency counts of a glossary of local image features, in the latter it gives us a bag of words which then contain a sparse vector of frequency counts of words.One of the most crucial developments in image classification using keypoints and descriptors is to epitomize these descriptors using a BoW model.

These descriptors between them contain spatial and geometric relationship information which gets lost using this notion. The reason that BoW has become popular is that the intrinsic simplification gains make it highly beneficial.

The descriptors extracted from the training images are grouped into N clusters of visual words using K-means. A descriptor is categorized into its cluster centroid using a Euclidean distance metric. For our purposes, we choose a

value of N = 200. This parameter provides our model with a balance between high bias (underfitting) and high variance (overfitting).

3.3 Multi-Class SVM

The Support Vector Machines methodology comes from the application of statistical learning theory. Support vector machines (SVMs) are supervised learning methods that generate input-output mapping functions from a set of labelled training data.[2] Every occurrence in the training set contains one class label and several attributes or features or observed variables. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes [14].

SVM perform a classification for problems by determining the separating hyper plane with maximum distance closest to points of training set. Usually it is employed for two class problems. Although there are many hyper planes able to separate data into multiple classes there is only one hyper plane achieves maximum separation. SVMs classify data by making this a part of a machine-learning process, which "learns" from the historic cases represented as data points [2]. These data points may have more than two dimensions.

If there is a training set of n samples, $\{x_i, y_i\}$, the separating hyper plane can be defined which satisfies the inequality:

$$y_i(w.x_i + b) > 0 \quad \forall \text{ i= 1,2,.....N} \tag{3}$$

where $x_i \in R^d$ are the vectors of d- dimensional features and $y_i \in \{+1, -1\}$ are the labels indicating the classes.

The set is said to be linearly separable if there is such a hyper plane .This causes the SVM to select the distance to the closest class from hyper plane as $1/\|w\|$ leading (1) to be

$$y_i(w.x_i + b) \geq 1 \tag{4}$$

To find the optimal separating hyperplane we need to minimize $\|w^2\|$ under constraints (2).The margin is $2/\|w\|$ and the cases closest to the hyper planes are called support vectors[13].

Lagrange multipliers are used to minimizing it under linear constraints (2) as $\|w^2\|$ is convex. If the N non negative Lagrange multipliers associated with constraints (2) is considered as $\propto = (\propto_1, \ldots\ldots, \propto_N)$, the optimization problem is summed as to maximize

$$W(\alpha) = \sum_{i=0}^{N} \propto_i - \frac{1}{2} \sum_{i,j=1}^{N} \propto_i \propto_j y_i y_j x_i x_j \tag{5}$$

with $\propto_i \geq 0$ and under constraint $\sum_{i=1}^{N} y_i \propto_i = 0$. This can be achieved by the use of standard quadratic programming methods.

Once the vector $\propto^0 = (\propto_1^0, \ldots\ldots, \propto_N^0)$solution of the maximization problem (3) has been found, the OSH (w_0, b_0) has the following expansion:

$$w_0 = \sum_{i=1}^{N} \propto_i^0 y_i x_i \tag{6}$$

The *support vectors* are the points for which $\propto_i^0 > 0$ satisfy (2) with equality.

Considering the expansion (4) of w_0, the hyperplane decision function can thus be written as

$$f(x) = sgn\left(\sum_{i=1}^{N} \propto_i^0 y_i x_i . x + b_0\right) \tag{7}$$

Although SVMs are usually used for binary classification, they can be adapted to multi-class problems. A multi class SVM classifier can be obtained by training several classifiers and combining their results. Two of the most used strategies to develop multi-class SVM are "*one-against-one*" or "*one-against-all*".

In this paper, linear SVMs classifiers are used using the above provided inputs. Once the bag-of-words features for all training images are obtained, they are given into SVMs. They find a hyper plane that separates the training data by maximal margin. "O*ne-against-all*" approach is used in the framework as it achieves comparable performance with faster speed than *"one -against-one"*. In the one against all implementation of SVM, n hyper planes are implemented, where n is the number of classes. Each hyper plane can be used to separate one class from the other classes.

3.4 Annotation of Images

For a given image to be queried, each extracted descriptor from this image is mapped into its nearest cluster centroid. A histogram of counts is assembled by incrementing a cluster centroid's number of occupants each time a descriptor is placed into it. The result is that each image is represented by a histogram vector of length N.The images are classified and the class in which they belong is their annotation label. Figure 3 shows how an image is annotated.

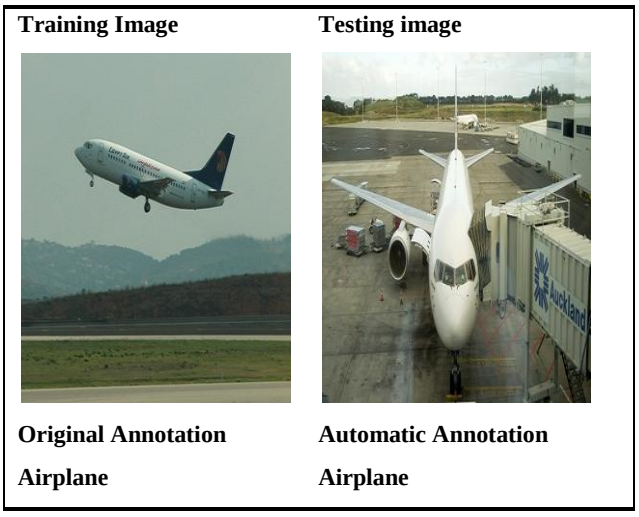


Figure 3: Automatic annotation compared with original annotation

3.5 Proposed Algorithm

Begin

1. Input Image To Train
   1.1. Compute SURF descriptors of Training Images
      1.1.1. Localize interest points
      1.1.2. Form Interest Points Descriptors
   1.2. Compute Bag-Of-Words
      1.2.1. Decide Vocabulary Size
      1.2.2. Cluster SURF Descriptor using K-means clustering
      1.2.3. find how many features from each cluster is present in image.
      1.2.4. Form Histograms of m bins for each image to be trained.
2. Train MultiClass SVM
   While m Training images
   Input Histograms of a training image to SVM of that class for that particular concept
   End while.
3. Annotate the test image
   3.1. Compute SURF descriptor of image.
   3.2. Compute BoW for the image.
   3.3. Input Image to MultiClass SVM.
   3.4. Class of image is determined.

End

4. EXPERIMENTAL RESULTS

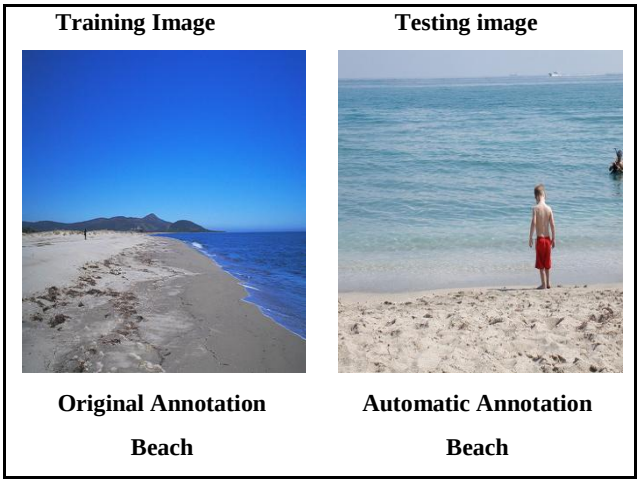
In this section the dataset that has been used is shown and also experimental results that has been found by proposed approach has been discussed. At last, some examples of different categories used have been shown.

4.1 Dataset

For evaluation purpose, Social20 dataset [6] is used. The Social20 set has 20 visual concepts and a total of 19,972 images. Under each concept category 1000 images are present. The concepts are very sundry: airplane, beach, boat, bridge, bus, butterfly, car, cityscape, classroom, dog, flower, harbor, horse, kitchen, lion, mountain, rhino, sheep, street, and tiger.

4.2 Illustrative Examples of Image Annotation

As the social20 dataset has 20 visual concepts, we have used only 10. Hence for each concept, 200 images were trained. Hence in total 2000 images were trained. This section shows some illustrative examples of the annotations generated by the approach.All 10 categories used are shown.



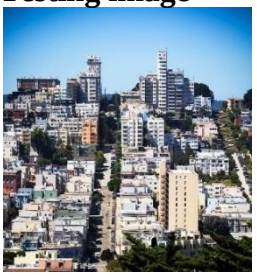
Class: Beach

Training Image **Testing image**


Original Annotation Bridge **Automatic Annotation Bridge**

Class: Bridge

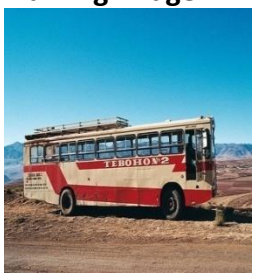
Training Image **Testing image**

Original Annotation Cityscape **Automatic Annotation Cityscape**

Class: Cityscape

Training Image **Testing image**




Original Annotation Bus **Automatic Annotation Bus**

Class: Bus

Training Image **Testing image**


Original Annotation Street **Automatic Annotation Street**

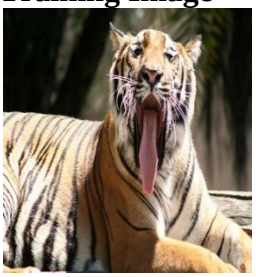
Class: Street

Training Image **Testing image**


Original Annotation Butterfly **Automatic Annotation Butterfly**

Class: Butterfly

Training Image **Testing image**




Original Annotation Tiger **Automatic Annotation Tiger**

Class: Tiger

Training Image **Testing image**


Original Annotation Horse **Automatic Annotation Horse**

Class: Horse

Training Image **Testing image**


Original Annotation Flower **Automatic Annotation Flower**

Class: Flower

Figure 4: Some examples to show how annotation is done for each category

4.3 Analysis of Results

Images are reclaimed from the vocabulary to appraise the annotation performance. The relevance of the retrieved images can easily be judged by looking at the real (manual) annotations of the images.

Specifically, Table 1 shows the average annotation results of images from the 10 categories, which have distinct semantics and have been widely used in the peer retrieval or annotation systems.

This experiment demonstrates the following:

- The overall annotation accuracy is 91.25%.
- The accuracy of 3 categories, namely Bridges, Horses and Tigers are close to 100%.
- All other categories achieve an average annotation accuracy of above 80 % with the exception of Beach and Mountain categories, whose average annotation accuracy is about 75%.

The recall is the number of correct annotations divided by the number of occurrences of keyword with ground truths in the test dataset. In other words, the recall of any classifier is computed as dividing the correctly classified positives by total positive count of images that are been tested [2].

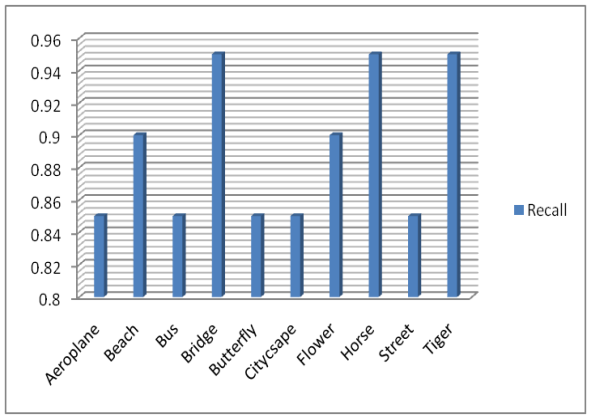


Figure 5

The precision is the number of correct annotations divided by no of predicted annotations. In other words, it is number of correctly retrieved images divided by the number of retrieved images [12].

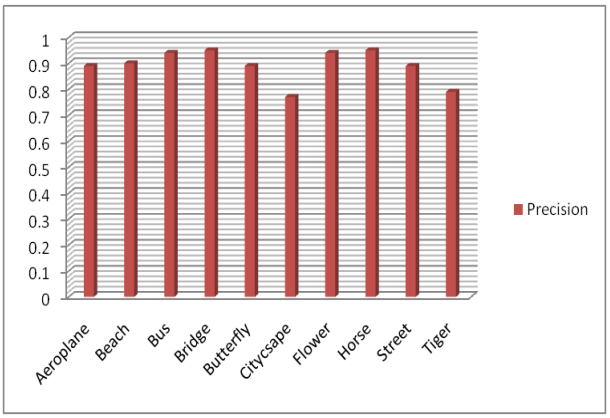


Figure 6

To combine recall and precision in a single efficiency measure, the harmonic mean of precision and recall is calculated. It is called F-measure. This is one of the aggregated performance measures.

Fmeasure=

$$\frac{2}{\frac{1}{precision}+\frac{1}{recall}}$$

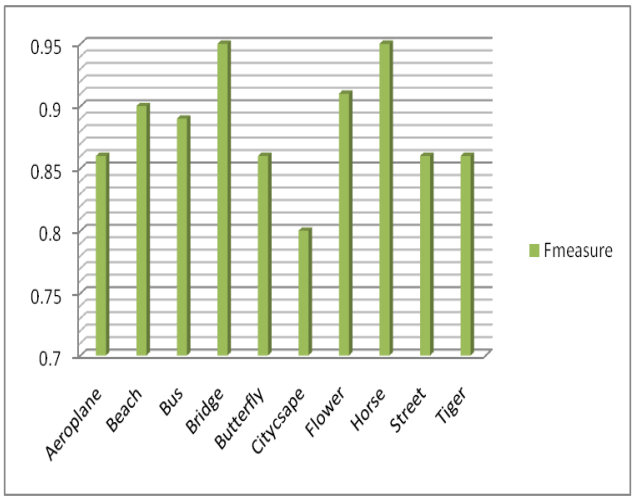


Figure 7

The figures 5, 6 and 7 show the recall, precision and f-measure of the experiments done on the dataset.

Table 2: Comparison of Systems

	ACCURACY (%)	Training time
Fusion Svm	88.8	~1.5
Proposed	91.25	~1.25

The proposed system is also compared with Multiclass SVM using fusion of MIL and Global descriptors [7] using images from the 10 categories. Table 2 recapitulates the performance of these systems in terms of the overall average annotation accuracy and the approximate average training time in minutes for one binary SVM. It clearly shows that our proposed system performs the better.

5. CONCLUSIONS

In this paper, an efficient and effective automatic image annotation system is presented. This can be easily integrated into image retrieval system. The proposed approach shows that SURF features can be used very efficiently to give us a good way to automatically annotate the images. As we can see that accuracy is very high of this approach, we need to find ways to make it better. Future work for this would be to use a more vigorous clustering algorithm that can replace K-means. Also, incorporation of keywords in the approach is also an area to work on in the future.

6. REFERENCES

[1] Gonzalez R., Woods R, Eddins S, Digital Image Processing Using MATLAB, 2nd ed., Pearson Education.

[2] Olson**,** David L**.,** Delen, Dursun, Advanced Data Mining Techniques, 2008, XII, Springer Publications

[3] Han, Kamber, Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann.

[4] Dengsheng Zhang, Md. Monirul Islam, Guojun Lu, "A review on automatic image annotation techniques", Pattern Recognition, Volume 45, Issue 1, Pages 346-362, January 2012.

[5] Lei, Yinjie, et al. "An HMM-SVM-based automatic image annotation approach." Computer Vision–ACCV 2010, Pages.115-126, 2011.

[6] Li, X., Snoek C., Worring M., "Learning Social Tag Relevance By neighbor voting", IEEE TRANSCATION MM, 11(7):1310-1322, 2009.

[7] Qi Xiaojun and Han Yutao., "Incorporating multiple SVMs for automatic image annotation", *Pattern Recognition.* 40, 2, Pages.728-741, February 2007.

[8] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." Computer Vision–ECCV 2006, Pages.404-417, 2006.

[9] Goh, K.-S.; Chang, E.Y.; Li, B.; , "Using one-class and two-class SVMs for multiclass image annotation," Knowledge and Data Engineering, IEEE Transactions on , vol.17, no.10, pp. 1333- 1346, Oct. 2005.

[10] Shi R., Feng H., Chua T.S., Lee C.H., 'An adaptive image content representa- tion and segmentation approach to automatic image annotation", Proceedings of the International Conference on Image and Video Retrieval, pp. 545–554, 2004.

[11] Cusano C., Ciocca G.,Schettini R., "Image annotation using SVM", Proceedings of the Internet Imaging IV, vol. 5304, SPIE, 2004.

[12] Jeon, Jiwoon, Victor Lavrenko, and Raghavan Manmatha. "Automatic image annotation and retrieval using cross-media relevance models." Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003.

[13] Chapelle, O.; Haffner, P.; Vapnik, V.N. , "Support vector machines for histogram-based image classification ," Neural Networks, IEEE Transactions on , vol.10, no.5, pp.1055-1064, Sep 1999.

[14] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.