RECOGNIZING OBJECT CATEGORIES IN NATURAL SCENES USING ROI SEGMENTATION IN SVM WITH SPATIO-GEOMETRIC CONSTRAINTS AND VISUAL SIMILARITY

S. Kumaravel [1], *Dr S. Veni* [2]

ABSTRACT

The image retrieval performance in Scalable databases, may be enhanced by certain paradigms like generating large Visual Vocabularies, Compact image representations and use of geometric information.

This paper first reviews the above approaches and then brings out the importance of using Saliency and Spatial information, to detect the Regions of Interest (ROI). The Spatial information of the visual words in the images has not been considered in the Bag-of-Words (BOW) model. This work includes the Spatial information of visual words in the images. The Visual vocabulary is constructed using K-means. The kd-tree is used for fast vector quantisation. The image representations used are: Spatial Histogram and PHOW. The low-level features i.e., Color, Shape and Contrast information, are combined using VLAD compact image representation. The combined features are then given to Linear SVM classifier, and implemented using MATLAB.

Keyword: Image Retrieval, ROI Segmentation, Categorisation of Objects, Spatio-Geometric constraints.

I. INTRODUCTION

Scene Recognition systems, for performing a large-scale database image search, are a real necessity today. Nowadays, the amount of multimedia data, have been increasing explosively. But, these large-scale multimedia data are not labelled and they also contain Noise. Moreover, manual labelling (annotating) such large scale multimedia data is quite expensive. Hence, the need for designing and developing semi-supervised or unsupervised systems, for recognizing Scene categories, arises. However, Unsupervised Scene Recognition systems are yet to become a reality, commercially. Conventional Image Understanding techniques for large scale image databases, focus only on using Global features for Categorisation into Classes like Buildings, Roads, Trees, Cows, Deer etc., and found them to be a computationally expensive task and also impractical, for real world applications like Surveillance systems. Hence, the necessity for, incorporating the Saliency paradigm (for identifying the Local Regions of Interest) and Geometric constraints (for generating Spatial information), which in turn, significantly improves the image retrieval time.

Generic Object Recognition, is a sub-domain of Computer Vision. While human beings carry out the task of Object Recognition without difficulty, it still remains arduous for Robots. Object Recognition has to solve many problems like occlusion, point of view,

[1] Research Scholar, PhD (PT) in Computer Science,
Karpagam Academy of Higher Education, Coimbatore - 21.
Email: ara28vel@gmail.com

[2] Prof.,&HOD, Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore - 21
Email: venics@kahedu.edu.in

illumination, Noise etc. Object Categorisation has to address the 'intra-class variation' issue, in addition to the issues encountered during the Object Recognition task.

As Visual Object Recognition is a vital ability required for carrying out the day-to-day activities like Walking, Reading, Communications etc., this Domain undergoes active Research in Neuroscience, Cognitive Psychology and Artificial Intelligence (AI) fields. Both, Biological and Artificial Visual Systems, while recognizing the object categories in a scene, face issues like variations in location, scale, view-angle, illumination, clutter etc.,

Generic visual categorisation, performs the task of Labelling an image, by using its semantic information. Each image is represented by one visual-word-occurrence Histogram. To obtain a Visual Vocabulary, insensitive to viewpoint and illumination, rotation or affine invariant orientation histogram descriptors of image patches, are vector quantized. Other Problems, having some relation to Visual Categorisation (VC), but are quite different from VC, are briefed below:

- Recognition: This paradigm identifies specific objects, for eg., US War-Planes or Russian War-Planes (not identified generally as War-Planes).
- Content Based Image Retrieval (CBIR): This paradigm retrieves images using the Low-level image features (color, texture etc.), given a Query image.
- Detection: This paradigm makes decisions, regarding the presence of a category member, in a given image.

The Image Categorisation problem, is more difficult than the conventional Classification problem (eg. Face recognition), because of the large within-class variations like color, shape and position, which are resolved through SALIENCY.

Salient Regions, are those regions which are more noticeable than the surrounding regions, in an image. Saliency Detection methods function on a single scale or Multi-scale. The Saliency values, per pixel, using the features' orientation, luminance and color are computed at different scales which form the Final saliency map. The salient features, thus obtained, are given to the SVM classifier, to identify whether the required category is present in the reference images.

Support Vector Machine (SVM) is one of the effective Pattern Classifiers, available. They have a Good generalization capability and are based on Structural Risk Minimization (SRM). The objective of SVM is, to maximize the margin between the data and the separating hyperplane. They construct a hyperplane, which separates the Positive and Negative Classes, with the maximum margin. The closest data points are called Support Vectors. SVM may be linear or non-linear. The basic SVM is Binary. SVMs have been used in many applications like object recognition, face recognition, gender classification, handwritten character recognition etc. Multi-Class SVMs are employed for Multi-Object Recognition in Scenes.

The Principal Benefactions, of this paper are:

- Applying one of the best Visual Feature Descriptors VLAD.
- Applying SALIENCY, to find the ROI local region.
- Including the Spatial information, in the Feature Descriptor.

- Visual Similarity, measured using one of the following descriptors: PHOG, PHOW, MSER, FV, VLAD etc.

The arrangement of the paper is as follows: Section 2 discusses the Related Works, Section 3 outlines the Methodology, Section 4 details the Results and Discussions and Section 5 concludes this Work.

2 RELATED WORKS

The following 3 approaches are generally used to improve the retrieval performance.

2.1. USING VERY LARGE VOCABULARIES

Sivic et al. [2] used the viewpoint invariant region descriptors for object representation. The regions, which survived for three frames or more, were considered stable. The vocabularies were built using k-means clustering. But, the performance decreased with the increasing vocabulary size.

Philbin et al. [3] used the approximate nearest neighbour technique for the object retrieval. The technique used randomized k-d trees for generating a forest, near the cluster centres. These trees, on combination resulted in a feature space, with overlaps. The features near the boundary, were not considered as Nearest Neighbors of the Result images. LO-RANSAC algorithm was used for a RE-RANKING of the Result images. This algorithm used lesser hypothesis, considering only one correspondence, and employed a few transformation classes, with 'shape' feature applied on Test images.

Philbin, Chum et al. [4] represented the Local Descriptors using soft quantisation, where each image patch descriptor was compared with the neighbouring clusters (visual words) in the local region for a match. Earlier methods had applied hard assignment, where each image patch descriptor was compared with ONLY the nearest SINGLE cluster.

Tong, Li et al. [5] implemented keypoint quantisation and image matching processes. This work unified the above two processes into a single framework. The matching between the query image and the gallery images were ascertained through a Kernel density function.

2.2. COMPACT IMAGE REPRESENTATIONS

Turcot, Lowe et al. [6] reduced the number of detected features per image, in large database recognition problems. Among all the local image features which were quantised, only a small subset (called "useful features") was selected, based on their geometrical information. Such "useful features" might be invariant to the point of view and also had consistency between them, geometrically.

Chum et al. [7] used a compact geometric representation-Hashing algorithm. The Hash keys were designed, with the help of additional Geometric information. The use of geometric (spatial) information for selecting the "secondary" features closer to the "central" feature, helped to increase the Hashing performance.

Perronnin et al. [8] used compressed Fisher vector, to reduce the memory requirement, as the Fisher kernel was high dimensional and dense.

Douze et al. [9] performed a Hamming Embedding on each of the GIST descriptors, to improve the web scale image search performance. They also observed that Local Representations gave significantly better results in Object Recognition applications, while the global GIST Descriptors gave High Accuracy percentage, in near-duplicate detection applications.

Horster et al. [10] used Latent Dirichlet Allocation (LDA) to obtain a compact image representation, suitable for large scale image retrieval. Each image in the database was represented as, a mixture of topics i.e., objects, which were unsupervisingly learned from their probability distributions. Scattered Interest Points detected in DoG Pyramid [18] were given the scale, position and orientation to form the SIFT local descriptor. Visual words were generated, by applying k-means Clustering, upon the subset of SIFT features, chosen randomly from the entire set. The occurrences of Visual Words generated above, were searched in every image of the database. Bin Wang et al. [11] performed Segmentation with, an Implicit Active Contour model called Level-set method, which was region-based. The model was augmented with the three-order tensor representation, by weighting the distances. The above representation was unified, by considering the Gabor features with grey values along with Gaussian smoothed image. The model was advantageous, due to its Noise Robustness.

Weiming Hu et al. [12] learned the Tensor subspace of the objects ONLINE, and consequently less resources (time of computation and memory) were consumed. Incremental SVD and Tensor decomposition were used, for the adaptive updation of the sample mean and eigen basis.

2.3. USING GEOMETRIC CONSTRAINTS

Jegou et al. [13, 14] observed a performance improvement, while searching large scale image databases, by using Hamming Embedding (HE) and Weak Geometric Constraints (WGC) jointly, within an inverted file. The matching of the visual words, was refined by using binary signatures, generated from Hamming Embedding (HE). Inconsistent matching descriptors were been filtered by WGC.

Wu, Ke et al. [15] generated local groups (Large Regions), by bundling of the image features, which were more Discriminative than a single SIFT feature and detected using MSER [16]. Geometric constraints within a Local group, were easily implemented. Two Local groups (of SIFT features) matched, resulting in only a portion of them being matched, which made this representation Robust to Occlusion, illumination etc. Such Large regions contained several Local patches, which were detected by SIFT [17].

Zhang, Jia, Chen [19] used geometry-preserving visual phrases (GVP), to provide Spatial information, during retrieval itself. Visual words grouped in a specific layout constituted GVP. GVP also generated the spatial layout of the words, both local and Long-range, other than the co-occurrences.

Shen, Lin et al. [20] used spatially constrained similarity measure (SCSM), during a large-scale image search, where the spatial consistency of the "visual words which matched" was also to be taken into consideration. SCSM was determined by using a Spatial Voting Map. The re-ranking of the Result images was done, by using the k-nearest neighbours, of the Query.

Philbin, Sivic and Zisserman [21] generated a model gLDA, having as its nucleus, the Geometric relations. This model recognized Multiple Unique Objects, by representing images as, Mixtures of Topics. To make this method a scalable one, initially a matching graph was constructed and then used with Clustering techniques, to divide the database into smaller image groups. The expensive gLDA, was then applied upon these smaller image groups.

3 METHODOLOGY

The object (s) in the Query image, is/are automatically acquired by the system, or may be selected by the user through a bounding box.

The BOVW model, is mostly used in image recognition application. However, the Spatial information of the visual words, is not considered in this model. The Saliency is used, to identify the Local Regions of Interest (ROI) in Query and Database images. For improving the retrieval performance, compact visual representation Vector of Locally Aggregated Descriptors (VLAD), which have high discriminative ability, are to be used.

This work applied the following techniques :

- K-means, for constructing a Visual Word Dictionary.
- Kd-tree, a fast Vector Quantisation technique.
- Spatial Histograms, as image descriptors.
- PHOW features (dense multi-scale SIFT descriptors).
- VLAD, being extremely compact (32 Kb per image), is very discriminative and efficient, in retrieval and classification tasks. VLAD, performs a pooling of local image features. This encoding maps the visual words in the database image and Query image. This descriptor accesses the Feature Co-ordinates of the image, which provides the Spatial information.
- The Linear SVM Classifier takes the Query image Histogram and the database images Histograms, as inputs for Recognition and implemented using MATLAB functions.

4 RESULTS & DISCUSSIONS

The recognition experiment uses the Caltech 101 dataset. The Performance Metrics used are: Precision and Recall.

PRECISION is the ratio of the number of correctly identified objects with the number of identified objects. RECALL is the ratio between the number of correctly identified objects and the number of correct objects in a scene.

The results are given as follows:

Figure 1 and Table 1 compare the Accuracy Recognition Percentage, between the 2 Paradigms. Using spatio-geometric information increases the Recognition Percentage. Figure 2 and Table 2 compare the Average Missing Rate, between the 2 Paradigms. Figure 3 and Table 3 compare the Time taken for ROI SEGMENTATION, between the 2 Paradigms. It is empirically found that, the Average Miss Rate and Time taken for ROI Segmentation is significantly reduced, while using the spatio-geometric information. The SVM CLASSIFIER PARAMETERS are given in Figure 4. The Ranking of the Test Images are given in Figure 5. Using the Precision and Recall Performance Metrics, the ACCURACY Percentages are given in Figure 6 (TRAINING dataset) and Figure 7 (TEST dataset).

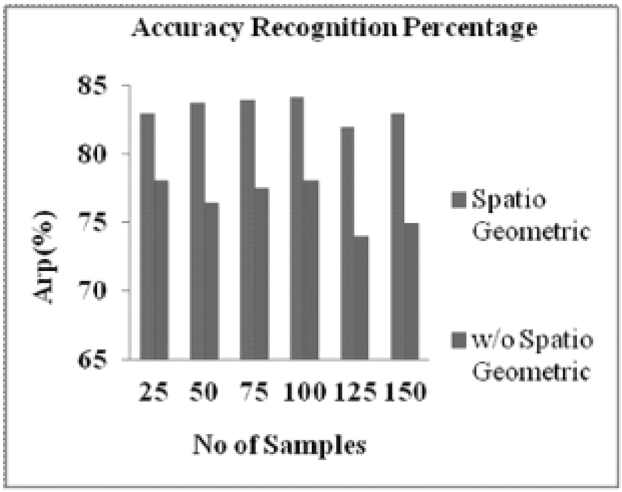
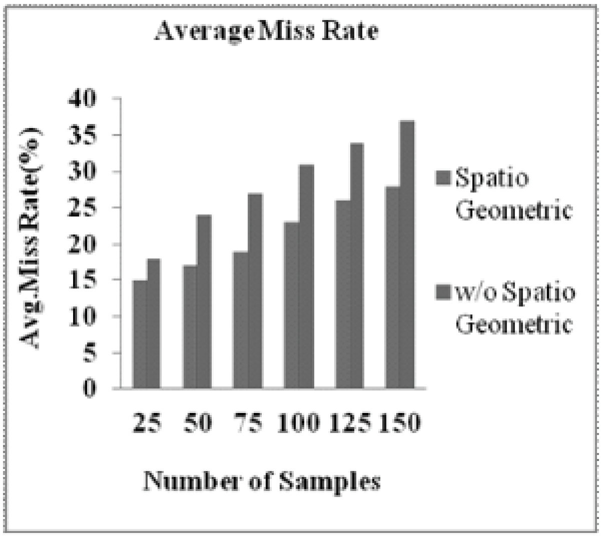


Fig 1 Accuracy Recognition Percentage compared

Table 1 Comparison of ARP % for the 2 paradigms

Number of Samples	Spatio Geometric	w/o Spatio Geometric
25	83	78
50	83.8	76.5
75	84	77.6
100	84.2	78
125	82	74
150	83	75



Number of Samples	Spatio Geometric	w/o Spatio Geometric
25	83	78
50	83.8	76.5
75	84	77.6
100	84.2	78
125	82	74
150	83	75

Fig 1 Accuracy Recognition Percentage compared

Table 1 Comparison of ARP % for the 2 paradigms

Table 3 Comparison of ROI segmentation time of the 2 paradigms

Number of Samples	Spatio Geometric	w/o Spatio Geometric
25	12	16
50	15	19
75	19	23
100	22	27
125	26	32
150	29	35

Fig 4 SVM Training Parameters

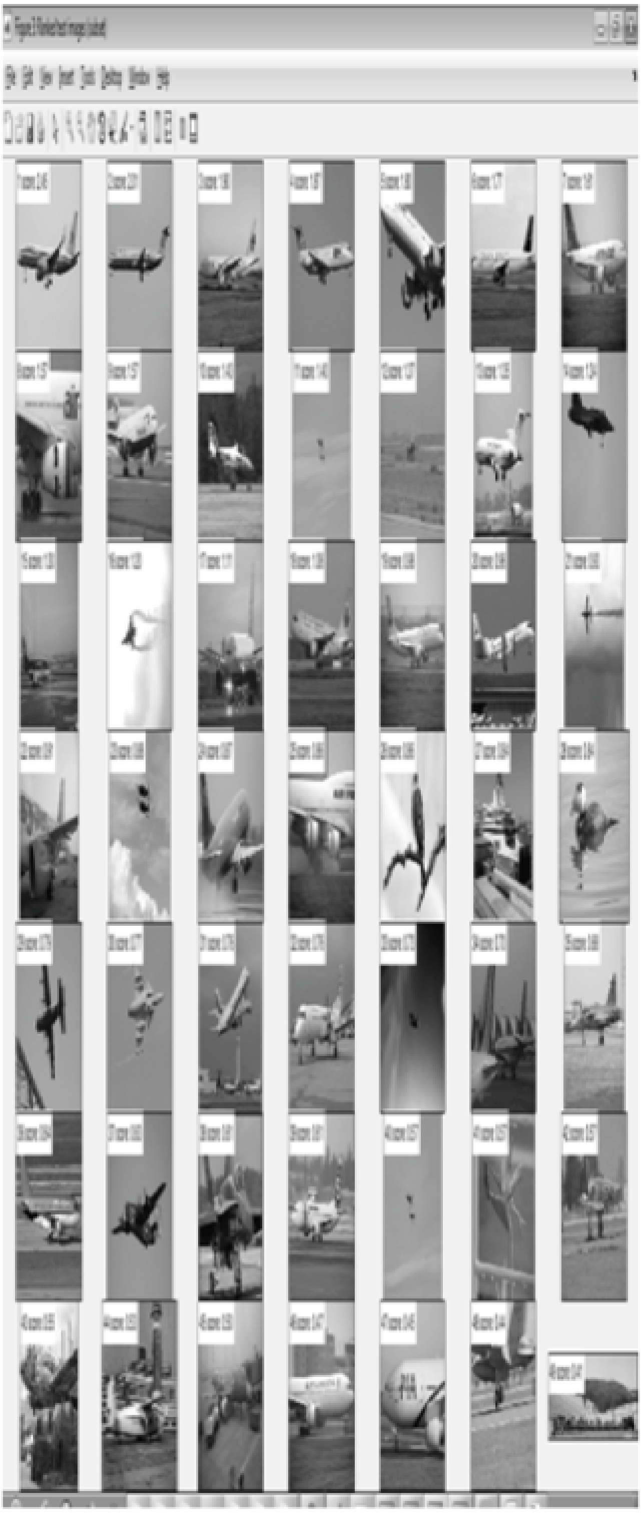


Fig 4 SVM Training Parameters

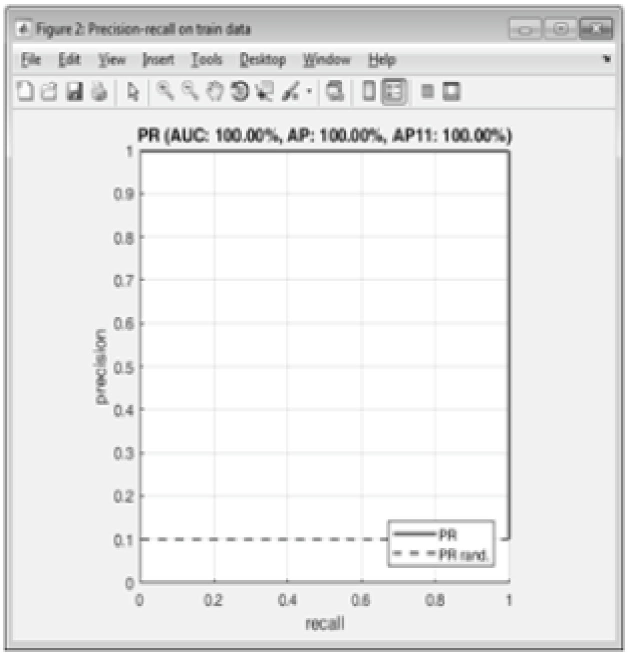


Fig 6 Precision recall on train data

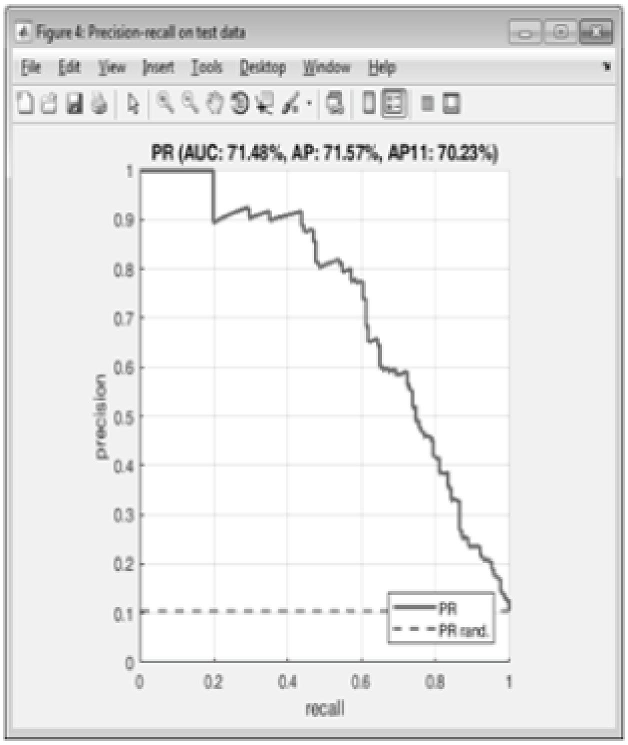


Fig 7 Precision recall on test data

5 CONCLUSION

The recognition of certain categories was experimented using the SVM Classifier. The Recognition performance was significantly improved using the Salient Local Regions. The retrieval was further enhanced, by using the Spatial and Geometric information, along with the Local Features. The selection of a better Visual Descriptor significantly improved the retrieval time.

This work has been successfully employed for object retrieval in natural scenes.

REFERENCES

1. G.Csurka, C.R.Dance, L.Fan, J.Willamowski and C.Bray, "Visual Categorisation with bags of keypoints", in Proc. Workshop Statistical Learning in Computer Vision, ECCV, p 1-22, 2004.
2. J.Sivic and A.Zisserman, "VideoGoogle: A text retrieval approach to object matching in videos", in Proc .Int .Conf. Computer Vision Vol 2, p 1470-1477, Oct 2003.
3. J.Philbin, O.Chum, M.Isard, J.Sivic and A.Zisserman., "Object retrieval with large vocabularies and fast spatial matching", in Proc. IEEE Conf. Computer Vision and Pattern Recognition, p 1-8, 2007.
4. J.Philbin, O.Chum, M.Isard, J.Sivic and A.Zisserman, "Lost in quantization : Improving particular object retrieval in large scale image databases", in Proc. IEEE Conf Computer Vision and Pattern Recognition, p 1-8, Jun 2008.
5. W.Tong, F.Li, T.Yang, R.Jin and A.Jain, "A kernel density based approach for large scale image retrieval", in Proc. 1 st ACM Int. Conf Multimedia Retrieval Ser ICMR' 11, NY, USA, 2011.
6. P.Turcot and D.Lowe, "Better matching with fewer features : The selection of useful features in large database recognition problems", Proc ICCV Workshop, p 2109-2116, Sep 2009.
7. O.Chum, M.Perdochand J.Matas," Geometric min-hashing: Finding a (thick) needle in a haystack", in Proc IEEE Conf CVPR 2009, p 17-24, 2009.
8. F.Perronnin, Y.Liu, J.Sanchez, and H.Poirier, "Large-scale image retrieval with compressed Fisher vectors", in Proc IEEE Conf CVPR 2010 , p 3384-3391, 2010.
9. M.Douze, H.Jegou, H.Sandhawalia, L.Amsaleg, and C.Schmid," Evaluation of GIST descriptors for web-scale image search", in Proc. ACM Int Conf CIVR '09, p 19.1- 19.8, 2009.
10. E.Horster, R.Lienhart, and M.Slaney," Image retrieval on large-scale image databases", in Proc 6 th ACM Int Conf CIVR'07, p 17-24, 2007.
11. Bin Wang, Xinbo Gao et al. "A Unified Tensor Level Set for Image Segmentation", in IEEE Trans. SMC- Part B Cybernetics: Vol 40, No 3, Jun 2010.
12. Weiming Hu, Xi Li et al. "Incremental Subspace Learning and its applications to Foreground Segmentation and Tracking ", Int. J. Comp. Vision, Vol 91, p 303-327, 2011.
13. H.Jegou, M.Douze, and C.Schmid," Hamming

embedding and weak geometric consistency for large scale image search" in Proc Euro.Conf Computer Vision, 2008

14. H.Jegou, M.Douze, and C.Schmid, " Improving Bag-of-features for large scale image search", in Int J. Comp. Vision, Vol 87, p 316-336, 2010.

15. Z.Wu, Q.Ke, M.Isard,and J.Sun," Bundling features for large scale partial-duplicate web-image search", in Proc IEEE Conf CVPR'09, p 25-32, 2009.

16. J.Matas, O.Chum, U.Martin and T.Pajdla, "Robust wide baseline stereo from maximally stable extremal regions", in Proc British Machine Vision Conf., London, UK, Vol 1, p 384-393, 2002.

17. D.G.Lowe, " Distinctive image features from scale-invariant key-points", Int J. Comp. Vision, Vol 60, p 91-110, 2004.

18. S. Kumaravel and C. Chellappan, "Parallel Neuro-Fuzzy Knowledge-Based System for facial expression synthesis", in Proc's of 5th Int., Conf., on Advanced Computing, India, p 308-311,1997.

19. Y.Zhang, z.Jia, and T.Chen, " Image retrieval with geometry-preserving visual phrases", in Proc IEEE Int. Conf. CVPR'11, p 809-816, 2011.

20. X.Shen, Z.Lin, J.Brandt, S.Avidan and Y.Wu, " Object retrieval and localization with spatially constrained similarity measure and k-NN re-ranking", in Proc IEEE Conf CVPR' 12, p 3013-3020, 2012.

21. J.Philbin, J.Sivic, and A.Zisserman," Geometric latent Dirichlet allocation on a matching graph for large-scale image datasets", Int J. Comp. Vision, Vol 95, No 2, p 138-153, Nov 2011.

22. Gonzalez-Diaz, C.E.Baz-Hormigos, M.Berdonces, and F.D.de Maria, "A Generative Model for concurrent image retrieval and ROI SEGMENTATION", in Proc 7th Int Workshop CBMI '12, France, Jun 2012.