Object Retrieval Using Image Semantic Structure Groupings

Nishat Ahmad[1] and Jongan Park[1],

[1] Dept. of Information & Communications Engineering, Chosun University,
309 Pilmun-daero Dong-gu Gwangju, South Korea
{Nishat Ahmad, Jongan Park, nischat}@gmail.com

Abstract. This paper explores basic level of semantic structure formation in the human vision inferential processes in line with Gestalt laws and proposes micro level semantic structure formations and their relational combinations. Using this approach two sets of semantic features have been derived for visual object class recognition. The algorithms have been applied to classify 101 object classes at one time. The results have been compared with existing state-of-the-art approaches and are found promising. Results from above approaches show that low level image structure and other features can be used to construct different type of semantic features, which can help a model or a classifier make more intelligent decisions and work more effectively for the task compared to low level features alone. Our experimental results are comparable, or outperform other state-of-the-art approaches. We have also summarized the state-of-the-art at the time this work was finished. We conclude with a discussion about the possible future extensions.

Keywords: Object recognition, Semantic structures, Graph theoretic

1 Graph model for classification

Since there is quite a considerable amount of variability at a structure level between the objects of the same semantic category, a graph model should take into account the commonality and variability in the semantic structures of the objects from same visual class. Going back to the basic argument of the idea that semantic objects are a combination of micro level semantic groups and their relations, we chart these factors for building a graph model. We build a graph model by iteratively merging the graphs of the test dataset and counting the frequency of the recurring semantic groups and relations. Groups and relations below a threshold are considered not essential in basic semantic labeling and are dropped. The resulting model graph is quite small and is neither a subset, nor a super set of any image graph in the test data set and captures the variability in all the test samples. It contains the set of those semantic groups and relations which are common in at least few test sets. Retention of semantic groups and relations which are common over a spectrum of test samples from the same object class can be considered as a basic semantic skeletal structure which is essential to identify an object.

For building a graph model we use a general relational structure matching approach which is less restricted than graph isomorphism, because nodes or edges may be missing from one or the other graph. Also, it is more general than sub-graph isomorphism because one structure may not be exactly isomorphic to a substructure of the other. A more general match consists of a set of nodes from one structure and a set of nodes from the other and a 1:1 mapping between them which preserves the compatibilities of properties and relations. In other words, corresponding nodes (under the node mapping) have sufficiently similar properties, and corresponding sets under the mapping have compatible relations. We use association graph techniques of general relational structure matching to build a graph model encompassing similarities and variability in visual object classes.

2 Classification steps

The classification task has been reduced to the graph matching between the model graph and the query graph. We constructed graph models of all the object classes in the test data set using 15 and 30 training images, selected randomly. The remaining images form the test data set. For the purpose of matching the query and model graph we used the association graph technique[1] and constructed a relational graph from the query and model graphs. We used relative histogram deviation measure equation (11) as a distance function for building nodes and arcs of the relational graph.
From the relational graph we find the maximum cliques in the graph[1]. The decision is based on the voting by each model based on the maximum cliques as follows.

$$vote = \sum_{i=1}^{n} a \times x \tag{1}$$

Where, a = {2, 3, 4...} are the a-clicks, i = {1, 2, 3 ...n} is the number of total click counts and x = {1, 2, 3 ...} is the frequency of an instance of a-click. In case of a tie, node and edge frequencies in the model graph are used as an additional vote for the nodes in the cliques. The vote for a node in case of a tie is calculated as: node vote = $f_g \times f_n$ Where as, f_g is inter-object node frequency from the model and f_n is the node frequency from the query image. Final vote is formulated by counting the maximum number of node votes.

3 Experiments and results

For testing the algorithm we have used the Caltech 101 data set as a number of previously published papers have reported results on this data set, thereby making comparisons more meaningful. In literature multiclass object categorization has been dealt in a less frequency. Many authors have reported the classification rates of their algorithms on a subset of the data and on class-wise classification methodologies, i.e. a classifier was trained in order to discriminate a single class among the subset from a background class consisting of arbitrary images. For comprehensive comparisons, we have shown results from published work on multiclass object categorization using whole of the Caltech 101 dataset. The algorithm was tested with the benchmark methodology of [2], where a number (in this case 15 and 30) of images are taken from each class uniformly at random as the training image, and the rest of the data set is used as test set. The “mean recognition rate per class” is used so that more populous (and easier) classes are not favored. This process is repeated 10 times and the average correctness rate is reported.

Table 1. Classification results : Comparison with published results using whole of Caltech 101

Model	**15 training images/cat**	**30 training images/cat**
Fei-Fei et al.[3]	18	--
Holub et al.[4]	37	43
Grauman & Darrell[2]	50	58
Nishat and park	**43**	**57**

When looking at the classification results of individual visual object categories, we find that our algorithm performed better for the classes which have distinctive semantic structure like airplane, motorbikes, grand piano, minaret, etc. The categories which were difficult to categorize are semantically more diverse, having greater shape variability due to greater intra-category variation and no-rigidity. A scrutiny of misclassification errors show that the misclassified objects have structural similarities, which needs additional features to be considered. The most common confusions are schooner vs. ketch (both are sail boats with three or four sails, commonly indistinguishable by uninitiated) and lotus vs. water lily (both are almost similar flowers).

4 Conclusion

The field of Content Based Image Retrieval (CBIR) has evolved very quickly due to the rapid advancement in technology, making possible unmanageable collections of image and multimedia data. The emphasis in future will be to make the CBIR systems more and more intelligent, mimicking human vision and intelligence. In this thesis work, effort has been made to understand the underlying principles of human vision perception and explore them to make the computer vision systems more intelligent in the task of image retrieval. David Marr wrote, “The true heart of visual perception is the inference from the structure of an image about the structure of the real world outside” [5]. This is the main objective of this thesis, to be able to infer a real world object from the structure of an image.

The thesis explores basic level of semantic structure formation in the human vision inferential processes in line with Gestalt laws and proposes micro level semantic structure formations and their relational combinations. Using this approach two sets of semantic features have been derived for visual object class recognition. The first algorithm uses the hypothesis in line with Gestalt laws of proximity that; in an image, basic semantic structures are formed by line segments (arcs also approximated and broken into smaller line segments based on pixel deviation threshold) which are in close proximity of each other. In the second approach a semantic group based on the proximity distance is clustered and modeled as a graph vertex.

The algorithms have been applied to classify 101 object classes at one time. The results have been compared with existing state-of-the-art approaches and are found promising. Results from above approaches show that low level image structure and other features can be used to construct different type of semantic features, which can help a model or a classifier make more intelligent decisions and work more effectively for the task compared to low level features alone. Our experimental results are comparable, or outperform other state-of-the-art approaches. We have also summarized the state-of-the-art at the time this work was finished. We conclude with a discussion about the possible future extensions.

References

1. Ballard, D. H., Brown, C. M., "Computer Vision", Englewood Cliffs, New Jersey: Prentice-Hall(1982)
2. Grauman and T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features", Technical Report MIT-CSAIL-TR-2006-020(2006)
3. Li Fei-Fei, R. Fergus and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories", IEEE CVPR2004, Workshop on Generative-Model Based Vision(2004)
4. Holub, M. Welling, P. Perona, Exploiting unlabelled data for hybrid object classification, Proc. NIPS Workshop on Inter-Class Transfer(2005)
5. David. Marr, "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information", W. H. Freeman and Co., ISBN 0-7167-1284-9(1982)