Object Retrieval Using Image Semantic Structure Groupings

Nishat Ahmad[1], Younghun Lee[2] and Jongan Park[1]
[1] *Chosun University,* [2] *Hannam University*
japark@chosun.ac.kr

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 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. Based on the notion of proximity a transitive relation is defined, which combines basic micro level semantic structures hierarchically till such a point where semantic meanings of the structure can be extracted. The algorithm extracts line segments in an image and then forms semantic groups of these line segments based on a minimum distance threshold from each other. The line segment groups so formed can be differentiated from each other, by the number of group members and their geometrical properties. The geometrical properties of these semantic groups are used to generate rotation, translation and scale invariant histograms used as feature vectors for object class recognition tasks in a K-nearest neighbor framework.

In the second approach a semantic group based on the proximity distance is clustered and modeled as a graph vertex. The line segments which are common to more than one semantic group are defined as semantic relations between the semantic groups and are modeled as edges of the graph. This way an image object is transformed into a graph using micro level structure formations. Each vertex and edge is labeled using translation, rotation and scale invariant properties of the member segments of each vertex and edge. From a set of training images, a graph model is constructed for visual object class recognition. The graph model is constructed by iteratively combining the training graphs and frequency labeling the vertices and edges. After the combining phase, all the vertices and edges whose repetition frequency is below a threshold are removed. The final graph model consists of the semantic nodes which are highly common in the training images. The recognition is based on graph matching the query image graph and the model graph. The model graph generates a vote for the query and ties are resolved by considering the node frequencies in the query and model graph.

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. Introduction

Using equations (1 ~ 4), we can find the coordinates of the minimum distance points on the respective lines which are closest.

This gives a minimum on the edge at (s_0, t_0) where $s_0 = 0$ and $t = t_0$:

$$t_0 = v \bullet w_0 / v \bullet v \quad (1)$$

If $0 <= t_0 <= 1$, then this will be the minimum and P(0) and Q(t_0) are the two closest points of the two segments. However, if t_0 is outside the edge, then we will have to check other cases for the true minimum.

$$\text{Similarly, for } s = 1, \quad t1 = (v \bullet w_0 + v \bullet u) / v \bullet v \quad (2)$$

$$\text{for } t = 0, \quad s0 = u \bullet w_0 / u \bullet u \quad (3)$$

$$\text{for } t = 1, \quad s1 = u \bullet v \bullet u \bullet w_0 / u \bullet u \quad (4)$$

From these points we can find the distance between them and the center point of the line joining them. All the line segments which are within the minimum distance threshold and whose center points of the line joining the minimum distance points are also within a defined distance threshold are grouped together as a semantic group. This can be understood from Figure 1, SG 1 ~ 3 are three semantic groups and the mean center point of the group is shown as a black dot. The relation R1 and R2 are the line segments which are at a minimum distance with two groups and hence joining the two groups. The relations R1 and R2 are common members of both the semantic groups they are joining. This way we have transformed an image object into a relational structure of semantic groups and their inter linking. The relational structure obtained is shown in the tabular form as table 3.4 below, which can easily be represented as a linked list. In order to use this structure for identification purposes, we have to compute some of its properties and associate these with individual semantic groups and relations. The only image information considered here is the extracted line segments. Since the line angles and lengths are subject to change under various transformations, we consider the effect of transformations and try to minimize or eliminate it.

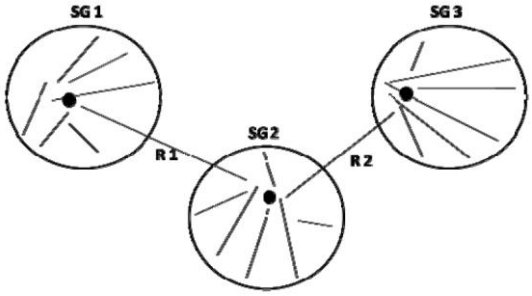

Figure 1. Semantic structures and their relations

Since Euclidean Distance Matrix (EDM) is invariant to rotation, translation and scaling equation (5), we compute EDM's from the geometric properties of the line segments of each semantic group.

$$A = (a_{ij})$$

$$a_{ij} = \| x_i - x_j \|_2 \quad (5)$$

Where $\|.\|_2$ denotes the 2-norm on R^m

Table 1. Layout of semantic structures and their relations

Semantic group - ID	link - ID	Linked group
SG 1	R 1	SG 2
SG 2	R 1	SG 1
	R 2	SG 3
SG 3	R 2	SG 2

For each semantic group, let L = {l_i|i = 1, 2, ..., N}, be the set of line segments obtained. Then we can compute geometric properties of L: the angles formed by all segments between each other and the relative length of each segment with respect to all other line segments. For the segment joining two semantic groups, we computed its geometric properties with respect to each group member of all the linked groups equations (6 ~ 9).

The angle between two line segments can be calculated as:

$$\cos\theta = \left| \frac{u \cdot v}{|u| \cdot |v|} \right| \tag{6}$$

where, u and v are line direction vectors of two line segments. The length of segment l_i with end points (x_0, y_0) and (x_1, y_1) is given as:

$$len(l_i) = \sqrt{(x_0 - x_1)^2 + (y_0 = y_1)^2} \tag{7}$$

Relative lengths of the line segments for constructing EDM are calculated as:

$$a_{ij} = \left| l_i - l_j \right| \tag{8}$$

We normalize the relative line length data in order to bring it in [0, 1] range as follows:

Given a lower bound l and an upper bound u for a feature component x,

$$\bar{x} = \frac{x - l}{u - l} \tag{9}$$

Results in $\bar{x}$ being in the range of [0, 1]. Now we have angles in the range of (±π) and relative line lengths in the range of [0, 1].

Since every EDM is symmetric, we extract the upper triangle matrix and form a histogram from each EDM with different resolutions based on empirical testing equation (10).

$$H_{ang} = \left\{h_{b_a}^{ang}\right\}, b_c = \{1,2,3,\ldots\ldots,B_a\}$$
$$H_{len} = \left\{h_{b_l}^{ang}\right\}, b_l = \{1,2,3,\ldots\ldots,B_l\} \quad (10)$$

where B_a and B_l denote the different number of bins of the three histograms. For H_{ang}, 72 bins that correspond to a 5 degree angle resolution were used. The resolution for H_{len} was taken as 101 bins through experimentation. As shown in the below figure, for each semantic group we get two histograms and for each relation we get four histograms, two from each group. The histograms of the semantic groups are obtained from the EDM of the group members and the histograms of the relations are the relative difference of length and angles of the R1 segment with the rest of the respective group members.

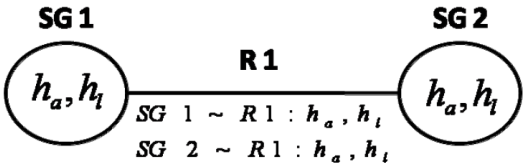

Figure 2. Properties of semantic groups and relations

Now we have a relational structure consisting of semantic groups with distinct properties and their inter-relations, also with distinct properties. This way we can simply define an image object in terms of its lower level semantic groups and their interrelations. Further we analyze the relational structure and introduce another property of frequency for repetitive semantic groups using a distance function (histogram deviation measure, equation (11))

$$d_{rd}(H,H') = \frac{\sqrt{\sum_{m=1}^{M}(H_m - H'_m)^2}}{\frac{1}{2}\left(\sqrt{\sum_{m=1}^{M}H_m^2} + \sqrt{\sum_{m=1}^{M}H_m'^2}\right)} \quad (11)$$

We combine the semantic groups which are exactly same and combine their relations with other nodes, under one semantic node. So, in the relational structure of an image object, the semantic groups or nodes are unique.

The relational structure can easily be represented in the form of a 4-tuple labeled graph g as:

$$g = (V, E, \alpha, \beta) \quad (12)$$

where : - L denotes the finite set of labels for nodes

- fv denotes the finite set of node frequencies
- M denotes the finite set of edge property labels for edges
- V is the finite set of vertices (semantic nodes)
- $E \subseteq V \times V$ is the set of edges (semantic relations)
- $\alpha : V \rightarrow L$, fv is a function assigning labels and frequency to the vertices
- $\beta : E \rightarrow M$ is a function assigning labels to the edges (semantic relation properties)

2. 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.

Given two structures $g_1 = (V_1, E_1, \alpha_1, \beta_1)$ and $g_2 = (V_2, E_2, \alpha_2, \beta_2)$. For each v_1 in V_1 and v_2 in V_2, construct a node of associated graph 'g' labeled (v1, v2) if v1 and v2 have the same properties [$\alpha_1(v_1, L)$ if $\alpha_2(v_2, L)$ $\forall v_1, v_2$]. Thus the nodes of G denote assignments, or pairs of nodes, one each from V_1 and V_2, which have similar properties. Now connect two nodes (v_1, v_2) and (v'_1, v'_2) of 'g' if they represent compatible relations, that is, if the pairs satisfy the same relations $\beta_1(e_1)$ if $\beta_2(e_2)$ $\forall e_1, e_2$. The properties can be compared using a suitable distance function (histogram deviation measure). A match between the two relational structures is just a set of assignments that are mutually compatible.

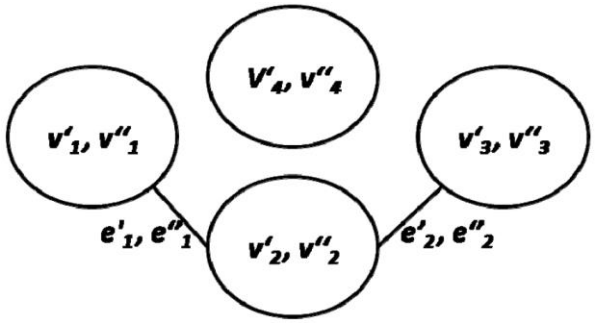

Figure 3. Association Graph

In the above figure, three nodes of the association graph are linked as they have compatible relations and the 4th node (v'_4, v''_4) has no compatible edges with any other node pair. The

association graph g = (V, E, α, β) obtained this way, gives us the similarity pattern between two object images of the same visual class.

Here we take the union of the two relational structures and add two more properties using the association graph, for intergroup frequency count of nodes (semantic groups) and edges (semantic relations). These labels show us the importance of a semantic group in the overall structure formation of that object class and also in deciding which semantic group would most likely be part of which object class in case there is a tie in the classification task. The union of the two structures is calculated as follows:

We have graphs 'g_1', 'g_2' and the association graph 'g', such that $g \subseteq g_1$ and $g \subseteq g_2$. The difference of $g_1 - g$ and $g_2 - g$ is a graph g'= (V', E', α', β') and g"= (V", E", α", β"), where.

$$V' = V_1 - V \qquad V'' = V_2 - V$$

$$E' = E_1 \cap (V' \times V') \qquad E'' = E_2 \cap (V'' \times V'')$$

$$\alpha'(v') = \alpha_1(v') \text{ for any } v' \in V' \qquad \alpha''(v'') = \alpha_2(v'') \text{ for any } v'' \in V''$$

$$\beta'(e') = \beta_1(v') \text{ for any } v' \in V' \qquad \beta''(e') = \beta_2(v'') \text{ for any } v'' \in V''$$

The difference graphs g' and g" above are obtained by removing the sub-graph g from g_1 and from g_2, including the edges that connect g_1 and g_2 with the rest of the graph. These edges are removed by finding the embedding of g with g_1 and g_2. The embedding of g in g_1 and g_2, emb(g, g_1) and emb (g, g_2) is the set of edges that connects g with $g_1 - g$ and $g_2 - g$.

$$g_{emb}(g, g_1)\ E1 \cap [(V \times (V_1 - V) \cup ((V - V) \times V)] \tag{13}$$

Where $\beta(e_{emb}) = \beta 1(e_{emb})$ for any $e_{emb} \in$ emb (g, g1).

From the association graphic we can count the inter-object frequency of the semantic group and relations and add two more labels f_g and f_1, such that:

- f_g denotes the finite set of group frequencies
- f_l denotes the finite set of edge (link) frequencies
- $\alpha : V \rightarrow L$, f_g is a function assigning labels and frequency to the vertices
- $\beta : E \rightarrow M$, f_1 is a function assigning labels to the edges

  (semantic relation properties)

Now we have graphs g'= (V', E', α', β'), g"= (V", E", α", β") and g = (V , E, α ,β) with V' IV" IV = Φ . We find the union graph G = ((g' ∪ g") ∪ g) . Let E ⊆ (V' × V") ∪ (V" × V') is a set of edges with labeling function : → M . The union of g' and g" is the graph g'" (V'", E'", α'", β'") where

$$V''' = V' \cup V''$$

$$E''' = E' \cup E'' \cup \overline{E}$$

$$\alpha'''(v''') = \begin{Bmatrix} \alpha'(v''') < if > v''' \in V' \\ \alpha''(v''') < if > v''' \in V'' \end{Bmatrix} \qquad \beta'''(e''') = \begin{Bmatrix} \beta'(e''') < if > e''' \in E' \\ \beta''(e''') < if > e''' \in E'' \\ \overline{\beta}(e''') < if > e''' \in \overline{E} \end{Bmatrix}$$

We repeat the process iteratively for a set of training images. The semantic groups and relations which are essential to give semantic meanings to the visual object will have intra-object frequency greater than a threshold 't'. So, all the semantic groups with intra-object frequency $f_g \leq t$ are considered redundant and are removed along with any edges they have with other nodes. The resulting model graph contains the repetitive patterns in a training set of images containing a visual object class. The removal of redundant groups reduced the model size to a considerably small level. Using the same procedure, we build up relational structure models for all the visual object classes we want to test.

3. 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 from 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.

$$vote = \sum_{i=1}^{n} a \times x \tag{14}$$

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 fn is the node frequency from the query image. Final vote is formulated by counting the maximum number of node votes.

4. 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.

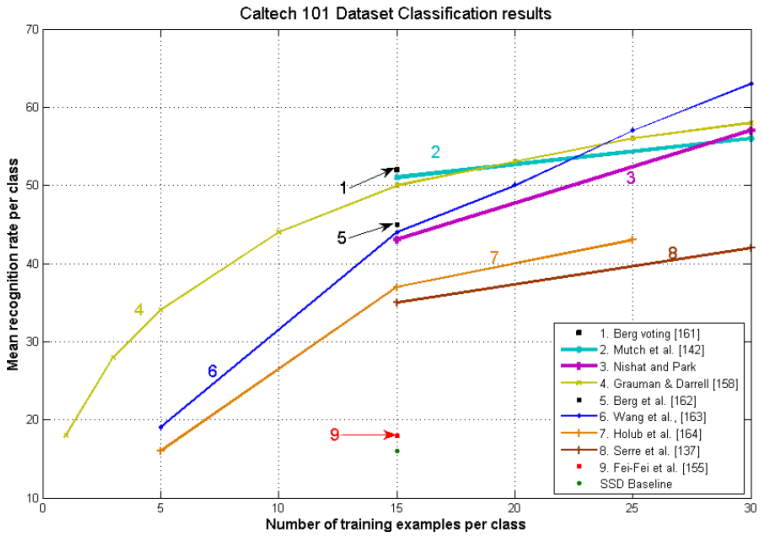

Figure 4. Classification rates Caltech 101 database

In Figure 4, number of training images per class is shown on x-axis and mean recognition rate per class on y-axis. The figure shows that our approach is comparable to other methods and has performed well above all except few. For the purpose of clarity, we have shown the published classification rates (correctness rates) using 15 and 30 training images per class, in Table 2. The blank cells indicate the unavailability of results in that category. Results for our algorithm are the average of 10 independent runs using all available test images. Scores shown are the average of the per-category classification rates.

Table 2. 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	--
Serre et al.[4]	35	42
Holub et al.[5]	37	43
Berg et al.[6]	45	--
Mutch et al.[7]	51	56
Nishat and park	**43**	**57**
Grauman & Darrell[2]	50	58
Berg voting[8]	52	--
Wang et al.[9]	44	63

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).

5. Conclusion and Future Work

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 out-side" [10]. 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. Based on the notion of proximity a transitive relation is defined, which combines basic micro level semantic structures hierarchically till such a point where semantic meanings of the structure can be extracted. The algorithm extracts line segments in an image and then forms semantic groups of these line segments based on a minimum distance threshold from each other. The line segment groups so formed can be differentiated from each other, by the number of group members and their geometrical properties. The geometrical properties of these semantic groups are used to generate rotation, translation and scale invariant histograms used as feature vectors for object class recognition tasks in a K-nearest neighbor framework.

In the second approach a semantic group based on the proximity distance is clustered and modeled as a graph vertex. The line segments which are common to more than one semantic group are defined as semantic relations between the semantic groups and are modeled as edges of the graph. This way an image object is trans-formed into a graph using micro level structure formations. Each vertex and edge is labeled using translation, rotation and scale invariant properties of the member segments of each vertex and edge. From a set of training images, a graph model is constructed for visual object class recognition. The graph model is constructed by iteratively combining the training graphs and frequency labeling the vertices and edges. After the combining phase, all the vertices and edges whose repetition frequency is below a threshold are removed. The final graph model consists of the semantic nodes which are highly common in the training images. The recognition is based on graph matching the query image graph and the model graph. The model graph generates a vote for the query and ties are resolved by considering the node frequencies in the query and model graph.

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.

For the semantic hierarchical relational features, the most important highlight of the comparisons is the choice of a classifier for the object categorization task. Boiman, *et al.*, (2008) [11] and Zhang, *et al.*, (2006) [12] have proposed to use modified or hybrid versions of knn classifier for better performance. In the future work we would like to test and improve the algorithm performance with modified and improved classifiers and incorporate additional features to reduce the classification confusion further down. In case of Graph model, we would like to test and improve the algorithm performance with modified and improved models and incorporate additional features to reduce the classification confusion further down. Since color and texture forms very important components in recognition; their inclusion into the proposed features in a semantic perspective can further improve the performance in recognition.

References

[1] D. H. Ballard and C. M. Brown, 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] F. F. Li, 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] T. Serre, L. Wolf and T. Poggio, "Object recognition with features inspired by visual cortex", Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, **(2005)**, pp. 994-1000.
[5] Holub, M. Welling and P. Perona, "Exploiting unlabelled data for hybrid object classification", Proc. NIPS Workshop on Inter-Class Transfer, **(2005)**.
[6] C. Berg, T. L. Berg and J. Malik, "Shape matching and object recognition using low distortion correspondence", Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR), **(2005)**.
[7] J. Mutch and D. G. Lowe, "Object class recognition and localization using sparse features with limited receptive fields", International Journal of Computer Visio, vol. 80, no. 1, **(2008)**, pp. 45-57.
[8] C. Berg, "Shape Matching and Object Recognition", PhD thesis, Computer Science Division, University of California, Berkeley, **(2005)**.
[9] Wang, Y. Zhang, L. Fei-Fei, "Using dependent regions for object categorization in a generative framework", Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR), **(2006)**.
[10]D. 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)**.
[11]O. Boiman, E. Shechtman, M. Irani, "In Defense of Nearest-Neighbor Based Image Classification", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **(2008)**, pp. 1-8.
[12]Zhang, A. Berg, M. Maire and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition", Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR), **(2006)**.