FEATURES WE TRUST!

Amir M. Rahimi, Lakshmanan Nataraj, B.S. Manjunath

Department of Electrical and Computer Engineering, University of California, Santa Barbara

ABSTRACT

We investigate the problem of image classification within a supervised learning framework that exploits implicit mutual information in different visual features and their associated classifiers. In our proposed two stage hierarchical processing, visual features are first clustered with the objective of maximizing diversity. Majority vote within each cluster is used to enforce diversity. Many partitioning variations are evaluated using K-nearest neighbor to obtain the highest inter-cluster entropy. In the second step, a richer measure of discrimination is obtained using a fully connected conditional random fields (CRF) over clusters. The unary and interaction potentials are defined over mutual information within each cluster and inter-dependencies across clusters respectively. Experimenting over five distinct datasets, we demonstrate an average performance gain of 30% compared with state of the art techniques.

Index Terms— Image classification, Feature interaction, Diversity maximization, Conditional Random Fields (CRF)

1. INTRODUCTION

A typical image classification workflow includes computing one of more visual descriptors from a given image, training appropriate classifiers to discriminate among various classes, and then using the learned model to classify a given image sample. Alternatively, many different classifier models can be built from these descriptors (Figure 1) where the classifier models are then aggregated (Figure 2). There is an extensive literature on combining weak classifiers, for example, Fisher Vectors [1], VLAD [2], Random Forests [3] and AdaBoost [4] where most model averaging techniques treat weak classifiers independently. Recent relationship modeling techniques such as graphical models [5], fuzzy techniques [6], neural network [7, 8] or other relationship modeling [9, 10] have shown a better discrimination over higher order interactions. This paper presents a novel and efficient way of aggregating classifier information by exploiting the mutual information implicit in the feature descriptors from which the classifiers are trained.

Our proposed method builds on existing work on classification based on standard visual descriptors. Weak classifiers are built using these descriptors to construct class labels in a supervised framework. Multiple descriptors for each image results in a label vector, where each component of the vector is the result of classifying one visual descriptor corresponding to that image. These vectors are then clustered using inter-cluster entropy as the objective function. We then define our mutual information with the co-occurrence statistics of cluster labels. Using these clusters, we then build a second stage classifier with a fully connected conditional random fields (CRF) to obtain a richer discrimination model. This CRF model learns the implicit dependencies among the visual descriptors. Extensive experimental results of this two-level classification approach demonstrates the significant performance gain in image classification over state of the art methods.

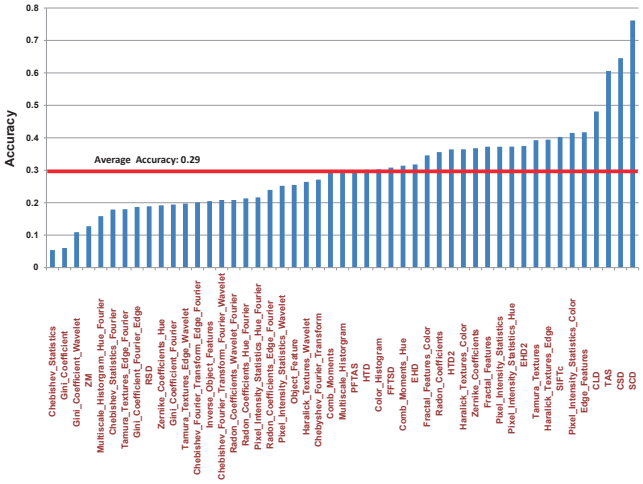


Fig. 1: Individual performance (F1-Score) of Random Forests classifier with different visual features on MIT8 Scene dataset

2. APPROACH

Consider the classification problem with M weak classifiers and N classes. Let us assume that each weak classifier returns two values; x: the classification score (hard prediction of single label), and y: corresponding regression score (soft prediction score for the prediction x). Let S be a set of images in a dataset, where each image s can only take one label l from a label set L. Given image s and an arbitrary classifier, the goal is to predict the correct label l^* such that;

$$l^* = \underset{l}{\operatorname{argmax}} \; P(l|\mathbf{x}, \mathbf{y}) \tag{1}$$

where $\mathbf{y}$ is a sequence of M classification scores and $\mathbf{x}$ is a sequence of M corresponding confidence scores,

$$\mathbf{x} = [x_1, ..., x_M] \quad and \quad \mathbf{y} = [y_1, ..., y_M] \tag{2}$$

Referring to Figure 2, for example, y_M and x_M are the label predictions and corresponding confidence score using the Color-Shape-Descriptor (CSD). In our experiments, we use Random Forests classifier for this first layer of classification since it provided the best performance on our datasets. However, we note that without loss of generality, the following discussion is applicable independent of the classifier used in the first layer.

2.1. Diversity Maximization with Clustering

The first level of classification results in the $(\mathbf{x}, \mathbf{y})$ vectors as explained above. Recall that each component y_i of $\mathbf{y}$ vector corresponds to a decision based on a particular feature (as in y_M corresponding the label based on CSD shown in (Figure 2) We would

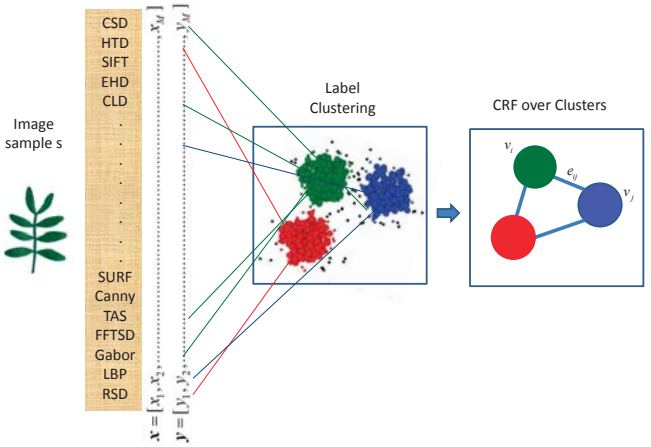


Fig. 2: Overview of descriptor aggregation: A number of descriptors are computed for each image and each of these descriptors are independently classified, resulting in a label vector y whose dimensionality M corresponds to the number of descriptors. The elements of y are then clustered based on label similarities (Algorithm 1). Mutual information among these computed clusters are then modeled with conditional random fields (CRF). This two stage clustering/classification method is then used to label unknown image samples.

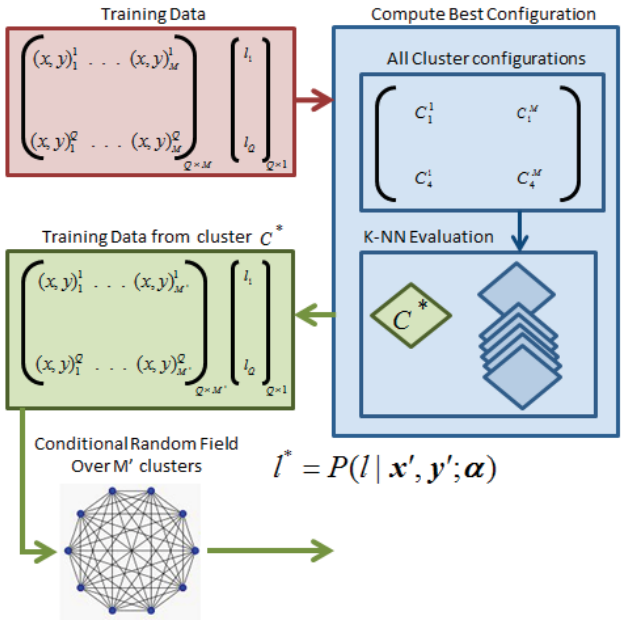


Fig. 3: Overall framework of proposed method. The training data $\mathbf{X}^{train}_{Q\times M}$ and $\mathbf{Y}^{train}_{Q\times M}$ (red box) is the output of arbitrary classifiers with the corresponding ground truth labels. The rows correspond to Q training images and the columns correspond to M label prediction using a particular visual feature. Columns of input matrix are then clustered using four clustering methods. All computed clusters are evaluated with K-nearest-neighbor to obtain the best configuration C*. A CRF model is then used to model dependencies between the descriptors and their associated labels to further improve the classification performance.

now like to cluster these different visual descriptors based on how similar their decisions are. For this we explore different clustering configurations with different parameter settings, with the objective of finding the best descriptor subsets that are similar in their labels over the entire training set. We would like to emphasize here that this clustering results in forming clusters of visual features and NOT the data items associated with the feature (as is typical with most image classification methods). At this point, we would have a very large number of cluster sets, each set could potentially correspond to a particular choice of the clustering method with the particular number of clusters. An exhaustive search over these cluster configurations results in identifying the best partitioning of visual descriptors. We use the maximum entropy criterion that maximizes inter-cluster separation while minimizing the intra-cluster entropy. This is explained by algorithm 1 and the overall framework is shown in Fig. 3.

Training data $\mathbf{Y}^{train}_{Q\times M}$ contains a total of Q training samples where M is the total number visual descriptors per sample. The data is organized as rows in the matrix $\mathbf{Y}$. The elements of this matrix correspond to the predicted label for the particular data sample using one of M visual descriptors. Similarly, we construct the validation matrix $\mathbf{Y}^{validation}_{V\times M}$. The label vectors from the training data are clustered in many different ways. Each cluster configuration C_m has a number of data partitions that are a function of both clustering method used and the associated parameter settings (e.g., number of desired partitions). Assuming T such configurations, we have $C_m \in \{C_1, C_2, ..., C_T\}$. Note that each configuration C_m partitions the set of visual descriptors, where the number of such partitions range from a minimum of 1 (all descriptors together) to a maximum of M (each descriptor is its own group). Consider C_m with M' partitions $C_m = \{C_m(1), \cdots, C_m(l), \cdots, C_m(M')\}$. We take a majority vote for the labels associated with each of these partitions $\{C_m(l)\}$, for each data in the training set, to create a new vector $\mathbf{y}'$ of length M'. This is our signature vector that characterizes the corresponding data item for the cluster configuration C_m. Each cluster configuration is then evaluated using a validation set with known ground truth for the labels as follows.

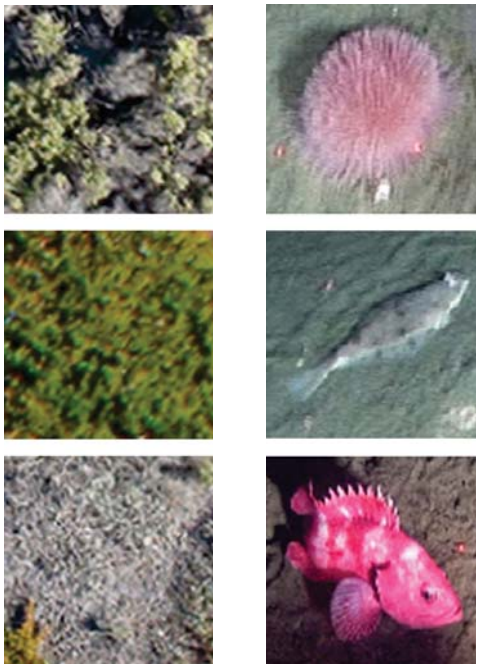
For each item in the validation set and using the cluster configurations C_m computed as above, we map the data items to the corresponding reduced-dimensional label vectors $y'(validation)$. Using the Hamming distance as the metric, we then compute the K-nearest neighbors of each $y'(validation)$ from the $y'(training)$. With the choice of hamming distance we are able to directly measure label similarities in computing the distance between two predictions. Finally, a majority vote is taken from this K-nearest neighbor set (using the ground truth labels associated with the training data) to determine the label for the data item from the validation set. If two labels have the same plurality within each cluster we then select one at random. Finally, the F-score (based on precision/recall) is computed over the entire validation set, by comparing these predicted labels with the associated ground truth labels (for validation set), thus giving a performance metric for the corresponding cluster configuration.

These computations are repeated for each of the T cluster configurations and the configuration with the best performance is selected for the second layer. Note that these computations need to be done only once. In the second layer, a CRF is trained over the lower dimension feature vectors $\mathbf{y}'$.

Why Clustering?: In principle the CRF model can be built using all M values of prediction labels. The complexity of CRF model depends highly on the number of connected nodes. A fully connected graph with the original set of M features is not only computationally

Algorithm 1 Clustering Optimization with K-NN

input: Prediction label data (Training) $\mathbf{Y}_{Q\times M}^{train}$
Prediction label data (Validation) $\mathbf{Y}_{V\times M}^{validation}$
Ground truth labels $L_{Q\times 1}^{train}$ and $L_{V\times 1}^{validation}$
Q: Number of samples in training set
V: Number of samples in validation set
T: Maximum number of cluster configurations
K: Parameter of K-nearest neighbor
output: Optimum cluster configuration C^*,
for all $C_m \in \{C_1, ..., C_T\}$ **do**
1. Cluster $\mathbf{Y}^{train}$ (Column-Wise)
2. Cluster $\mathbf{Y}^{validation}$ with identical partitions as (1)
3. Compute *majority vote* for all clusters
4. Using (3) create a new data $\mathbf{Y'}_{Q\times M'}^{train}$ & $\mathbf{Y'}_{V\times M'}^{validation}$
for all $\mathbf{y'}^{validation} \in \mathbf{Y'}^{validation}$ **do**
a. $d_{Hamming}(\mathbf{y'}^{train}, \mathbf{y'}^{validation}) = \sum_{m=1}^{M'} \mathbb{I}(y_m'^{train} \neq y_m'^{validation}))$
Where $\mathbb{I}(.)$ is an indicator function.
b. Sort $\mathbf{y'}^{train}$ with ascending order
c. Select top K corresponding $L^{validation}$
d. Assign l^{knn} with a *Majority Vote* at (c)
e. Push the estimated label at d to $L_{V\times 1}^{knn}$
end for
5. Evaluate F_1 $Score(L^{knn}, L^{validation})$ for given C_m
6. Update C^* for the best performance at (5)
7. Empty $L_{V\times 1}^{knn}$
end for



(a) UAV (b) Bering Sea

Fig. 4: Sample images from two of the datasets

expensive (learning parameters) but also requires an exponentially increasing number of data items to avoid overfitting [11, 12]. Further, the above described clustering steps helps to group "similar" visual features together, thus creating a robust subset of features on which the random field model can be built.

2.2. CRF Over Clusters

Following the clustering stage, a graphical model over the clusters C^* is learnt to discriminatively label each image pattern. Recall that each cluster within C^* contains a number of labels $\{y\}$ and their corresponding confidence values $\{x\}$. Similar to the cluster label vector $\mathbf{y}'$, a corresponding cluster confidence vector $\mathbf{x}'$ is computed by averaging over confidence values within each cluster. The new decision and confidence vectors $\mathbf{x}'$ and $\mathbf{y}'$ are then used as inputs to the second layer of CRF model computations.

The training data is used to construct the cluster confidence and label matrixes, $\mathbf{X}'_{Q\times M'}$ and $\mathbf{Y}'_{Q\times M'}$. We now construct a fully connected graphical model $G = (V, E)$ such that every node v_i corresponds to a partition i. Note that M' is the number of partitions as computed by Algorithm 1. In the standard CRF formulation, an unknown sample is assigned the label l^* as follows:

$$l^* = \underset{l}{\operatorname{argmax}}\ P(l|\mathbf{x}', \mathbf{y}'; \boldsymbol{\alpha}) = \frac{exp\{\mathbf{F}(\mathbf{x}', \mathbf{y}'; l)\}}{Z(\mathbf{x}', \mathbf{y}'; l, \boldsymbol{\alpha})} \tag{3}$$

where $Z(.)$ is the partitioning function,

$$\mathbf{F}(\mathbf{x}', \mathbf{y}'; l') = \sum_{i=1}^{M'} \boldsymbol{\alpha}_n \mathbf{f}_n(x'_i; l) + \sum_i \sum_{j<i} \boldsymbol{\alpha}_e \mathbf{f}_e(y'_i, y'_j; l) \tag{4}$$

Alternatively, we can rewrite the conditional likelihood as,

$$P(l|\mathbf{x}', \mathbf{y}'; \boldsymbol{\alpha}) = \frac{\exp\{\sum_i \boldsymbol{\alpha}_n \mathbf{f}_n(i; l) + \sum_i \sum_{j<i} \boldsymbol{\alpha}_e \mathbf{f}_e(i, j; l)\}}{Z(\mathbf{x}', \mathbf{y}'; l, \boldsymbol{\alpha})}$$

Unary and Interaction Potentials: Unary potentials measure the influence of each node and interaction potentials capture the influence of each possible pairwise configurations among clusters. We average the confidence measure at each node as an estimate of the significance of that node,

$$\mathbf{f_n}(i; l) = \frac{1}{N_i} \sum_k x_{i:k} \quad \forall x \in \text{cluster } k \tag{5}$$

N_i is the number of components (i.e., descriptors that are grouped together) in cluster i. The edge potentials are computed as proportional to the co-occurrences of labels associated with nodes i and j,

$$\mathbf{f}_e(i, j; l) = -\log P(l|y'_i, y'_j; \mu_{ij})$$

$$= -\log \frac{1}{1 + \exp -\{\sum_{\mu_{ij}} \varphi^{ij}(y'_i, y'_j; l)\}}$$

where CRF hyper-parameter μ_{ij} indicates the connectivity of node i and j which in our setup is the fully connected structure and $\varphi^{ij}(.)$ is the weighted sum of different co-occurrences. The above is a standard CRF formulation adapted to our current problem. For more details on the CRF formulation and parameter estimation we refer to [13]. The training process includes using a labeled set of images and the associated descriptors to learn the CRF model for each possible label. During testing, the CRF model is used to estimate the label with the highest conditional probability as given by Equation (3).

List of Datasets	Number of Labels	Number of Features	Independent Base Accuracy	Independent Max Accuracy	CRF Avg Accuracy with C^*	CRF Max Accuracy with C^*
BeringSea	20	29	37.20	53.42	**59.81**	**77.53**
Flower	20	29	22.69	69.56	**73.14**	**79.01**
UAV	5	29	63.41	83.78	**88.81**	**93.99**
ImageClef(07)	20	29	78.14	82.03	**82.14**	**83.34**

Table 1: Experimental results over 5 datasets: The second column contains the number of labels in every dataset, the third has the number of features used for every dataset, the fourth indicates the base performance when using Random Forests with independent visual features averaged over all labels, the fifth shows the performance of the label with highest accuracy, the sixth column is the performance of final classification (after CRF) averaged over all labels and the seventh shows the final performance (after CRF) of the label with highest accuracy

3. EXPERIMENTS

Datasets: We used five datasets to demonstrate the efficacy of the proposed method: **The ImageCLEF 2011 (IC)** plant identification dataset contains 5436 images of 71 tree species from French Mediterranean area [14]. **MIT Scene (MIT)** contains 2600 images of 8 scenes [15]. **The Flower (FL)** dataset [16] contains 8189 images with the 20 most common classes out of 102 available. **The UAV Dataset (UV)** is aerial view dataset acquired with our UAV and contains 800 high resolution samples for 5 specie of vegetation (Fig. 4a). **Bering Sea Canyons (BS)** dataset contains 23 hours of annotated HD video. The dataset has 54 different species [17] of which we choose 19 most common classes (Fig. 4b).

Clustering Techniques: For clustering, we used the following four widely used methods: KMeans [18], Spectral Clustering [19], Hierarchical clustering [18], and Affinity Propagation [20]. For each method, the entire range of possible number of clusters is explored by tuning the appropriate method-specific parameters, and Algorithm 1 is used to identify the optimal clustering strategy. Figure 3 shows an example of such clustering for one of the datasets.

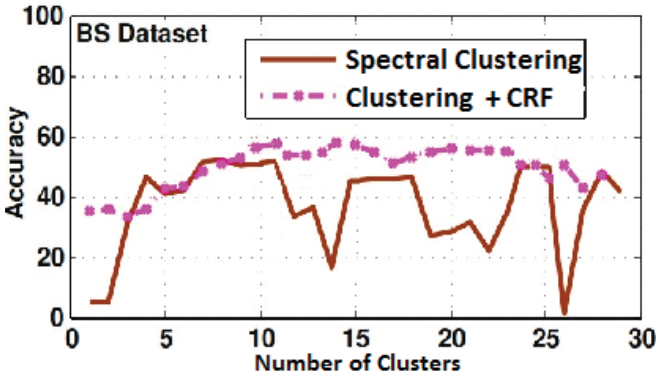


Fig. 5: Performance improvement in classification of Bering Sea Canyons (BS) Dataset when CRF is built over the clusters using 29 visual features. The solid line indicates the performance of K-nearest neighbor with Spectral Clustering as the number of clusters is varied from 1 to 29 clusters. Dashed line indicates the final performance using each cluster configuration. Similar patterns are observed with other datasets.

Evaluation and Results: Our experiments, summarized in Table 1, over five different datasets demonstrate that the proposed method consistently outperforms other commonly used aggregation methods, including Random Forests. The maximum gain in base accuracy after CRF (Column 4 and Column 6 of Table 1) is **50.45%** for the Flower dataset, while the least gain iss 4% for the ImageClef(07) dataset. Similarly, the maximum performance gain for the best label after CRF (Column 5 and Column 7 of Table 1) is **24.11%** for the Bering Sea dataset, while the least gain is 1.2% for the ImageClef(07) dataset. The average performance gain over all five datasets is **30.4%**. In Table 2 we see that there is a significant Performance Loss in base accuracy when using Random Forests on visual features that are averaged over the labels (Column 3). Similarly, there is also a Performance Gain (Column 4) after using CRF on the clusters.

Dataset	Accuracy MV on Features	Independent Base Accuracy	CRF Avg. Accuracy with C^*	Number of Clusters in C^*
BeringSea	51.14	37.20	**59.81**	**11**
Flower	68.65	22.69	**73.14**	**18**
Aerial	79.03	63.41	**88.81**	**12**
ImageClef(07)	81.19	78.14	**82.14**	**10**
MIT	63.55	29.14	**78.56**	**33**

Table 2: Here we compare Random Forests (RF) classifier with CRF on clustered features. The second column shows the base performance when majority vote (MV) is performed over M label predictions using RF (as explained in Section 2.1). The third column indicates the base performance when using RF with independent visual features averaged over all labels. From the third column, it is clear that there is a significant performance loss during averaging the label predictions. The fourth column shows the Performance Gain obtained after using CRF over the clusters. Finally, the fifth column shows the number of clusters for the best cluster configuration.

Improved Classification using CRF: In Figure 5 we show the performance of CRF over different number of clusters. As shown in the graph, the CRF built over the clustered data consistently performs better over a wide range of clustering configurations. Note that increasing the number of clusters does not result in any significant performance gain, while a larger number of clusters results in high computational cost. For instance, for the 11-cluster configuration, the CRF training took about 2 hours on a standard desktop whereas with 29 clusters it took 5 days on the same computer. For an unknown sample classification, most of the computing complexity is in the descriptor calculations which can be easily parallelized.

4. CONCLUSION

We described a two step hierarchical approach for image classification. In the first step visual features are clustered with the objective of maximizing inter-cluster entropy. We then take the majority vote within each cluster. In the second stage, a CRF model is used to capture the inter-cluster dependencies for enhancing the discriminative power. Our extensive experiments over 5 datasets show that by leveraging the mutual information implicit in the visual features, significant performance gain can be obtained for image classification.

Acknowledgements:

This work is supported in part by ONR Grants # N00014-12-1-0503 and ONR # N00014-14-1-0027.

5. REFERENCES

[1] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3304–3311.

[2] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, pp. 143–156. Springer, 2010.

[3] L Breiman, "Random forests," *Machine Learning*, pp. 45, 532.

[4] Yoav Freund and Robert E Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.

[5] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie, "Objects in context," in *International conference on Computer vision*. IEEE, 2007, pp. 1–8.

[6] Theam Foo Ng, Tuan D Pham, and Xiuping Jia, "Feature interaction in subspace clustering using the choquet integral," *Pattern Recognition*, vol. 45, no. 7, pp. 2645–2660, 2012.

[7] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov., "Improving neural net-works by preventing co-adaptation of feature detectors.," in *CoRR*, 2012, vol. arXiv:1207.0580, 2012.

[8] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[9] Henry Schneiderman and Takeo Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*. IEEE, 1998, pp. 45–51.

[10] William J McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, 1954.

[11] Michael P Perrone and Leon N Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," Tech. Rep., DTIC Document, 1992.

[12] Richard Maclin and David Opitz, "Popular ensemble methods: An empirical study," *arXiv preprint arXiv:1106.0257*, 2011.

[13] C.Elkan., "Log-linear models and conditional random fields," Tech. Rep., 2008.

[14] ImageCLEF, "Imageclef," http://www.imageclef.org/2011/plants, 2011.

[15] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[16] Visual Geometry Group, "Flower data set," http://www.robots.ox.ac.uk/ vgg/data/flowers/, 2006.

[17] AM Rahimi, RJ Miller, DV Fedorov, S Sunderrajan, BM Doheny, HM Page, and BS Manjunath, "Marine biodiversity classification using dropout regularization," in *Computer Vision for Analysis of Underwater Imagery (CVAUI), 2014 ICPR Workshop on*. IEEE, 2014, pp. 80–87.

[18] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani, *The elements of statistical learning*, vol. 2, Springer, 2009.

[19] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al., "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[20] Brendan J Frey and Delbert Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.