

IMPLICIT SPATIAL INFERENCE WITH SPARSE LOCAL FEATURES

Deirdre O'Regan and Anil Kokaram

Dept. of Electronic & Electrical Engineering
Trinity College Dublin, Ireland
oregandm@tcd.ie

ABSTRACT

This paper introduces a novel way to leverage the implicit geometry of sparse local features (e.g. SIFT operator) for the purposes of object detection and segmentation. A two-class Bayesian scheme is used as a framework, and the likelihood is derived from the real-valued classification of machine learning algorithm Gentle AdaBoost, whose output is transformed to a probabilistic distribution using either of two models investigated; Log-Sigmoid or Bi-Gaussian. The main contribution is a novel scheme for the injection of prior contextual spatial information. This occurs on a uniquely designed Markov Random Field defined by Delaunay Triangulation of the feature points. Our experiments show that this framework is useful for object detection and segmentation, and we achieve good, mostly invariant results in these tasks.

Index Terms— Object detection, Bayes procedures, Delaunay triangulation, Feature extraction, Geometric modeling

1. INTRODUCTION

Sparse, local feature-based object detection has become very popular, with Lowe's SIFT (Scale Invariant Feature Transform) [1] remaining a commonly-used interest operator for such tasks. Recent work in the field has leveraged the relative geometric contexts of these sparse features in classification. Popular approaches involve training probabilistic part-based models [2, 3], "vocabulary" clustering [4, 5] or feature merging [6]. Part-based models, while effective, are training-complex and clustering can be sensitive to the number of clusters and clustering method used. Star Graph-like geometric models [3, 5, 6] and Pairwise Spatial Relations [7] have also been used. These models define spatial relationships between features explicitly (i.e. as a semi-rigid, global model), which is interesting when we consider that the typical interest point operator cannot guarantee 100% feature "repeatability" (i.e. detection of completely identical feature sets across images capturing different instances of an object).

We propose an alternative technique that can group sparse object features implicitly, without the need for an explicit geometry, part representation or clustering. The idea is that we expect features to occur in a particular, but loose spatial configuration. Faces, for example, are generally contiguous regions in images, as are backgrounds. We present a traditional Bayesian modeling of this idea. But uniquely, the likelihood is produced by a machine learning algorithm, while spatial inference occurs over a Markov Random Field (MRF) modeled on a special graph created by Delaunay Triangulation of the sparse feature points. The MRF is only concerned with

This work was funded by the Irish Research Council for Science, Engineering and Technology (IRCSET) and partly by Adobe Systems Incorporated. Thanks to David Simons of Adobe Systems, and Francois Pitié of Trinity College Dublin, for their contribution to this work.

the "local" context of features, so we are not imposing a rigid, global structure on the feature geometry. Therefore feature repeatability becomes less critical, and the inherent invariance of each feature can be preserved. The result is a fairly accurate, mostly invariant object localization, coupled with a rough segmentation.

2. SPATIAL INFERENCE ON A DELAUNAY GRAPH

Consider that a test image produces K local (e.g. SIFT) features, and we wish to classify $f(\underline{x}_k) = f_k$, at image location \underline{x}_k as positive (i.e. belonging to object such as a face), or negative (i.e. belonging to background). According to Bayes rule, the probability that f_k belongs to either class is

$$p(l_k|f_k) \propto \underbrace{p(f_k|l_k)}_{\text{likelihood}} \underbrace{p(l_k|L_k)}_{\text{prior}} \quad (1)$$

where $l_k = l(\underline{x}_k) \in \{1, 0\}$ is a labeling as positive or negative respectively. $L_k = L(\underline{x}_k)$ is the spatial "neighborhood" of labels around \underline{x}_k .

The likelihood represents a connection between the observed feature point and the label that is assigned, while the prior quantifies knowledge about the label field before observation. It is here that contextual spatial information is introduced to the solution, and an MRF is used as the prior in the usual way, although it is defined on a uniquely designed Delaunay graph.

2.1. Obtaining the Likelihood

The likelihood would normally be obtained by proposing some parametric model in feature space, such as a Gaussian distribution for each class. This would lead naturally to yet another Bayesian classifier for the problem. However, given the success of recent object detection schemes based on machine learning algorithms [6, 7], we propose here to use the output of such a classifier to determine the likelihood.

This might seem strange, but consider the output of a real-valued AdaBoost classifier, such as the GML implementation of Gentle AdaBoost (also called GentleBoost) [8] which we use for our experiments. It is a point-wise classifier which yields a real-valued measure of "confidence" that we assume is related to the likelihood of $l = 0$ or $l = 1$ at each site. Specifically, GentleBoost labels feature f_k according to the rule

$$l_k = \begin{cases} 1 & \text{for } c_k > h \\ 0 & \text{for } c_k \leq h \end{cases} \quad (2)$$

where $c_k \in R$ is a soft classification, and h is a hard labeling threshold. The greater the margin $z_k = |c_k - h|$, the more confident we can be about l_k .

After feature classification with GentleBoost, the distribution $c_{k..K}$ must be adapted to probability space. We propose two models that will transform these values into two-class likelihood distributions. The first naively assumes no class overlap in the distribution of confidence values. This Log-Sigmoid (L-S hereafter) model is defined as

$$p_o(f_k|l_k) = \begin{cases} \frac{1}{1+e^{-(c_k)}} & \text{for } l_k = 1 \\ \frac{1}{1+e^{-(-c_k)}} & \text{for } l_k = 0 \end{cases} \quad (3)$$

The Bi-Gaussian (B-G hereafter) model, which accounts for class intersection in the confidence distribution, is defined as

$$p_g(f_k|l_k) = \begin{cases} \frac{1}{(\sigma_p)\sqrt{2\pi}} e^{-(c_k-c_{\mu_p})} & \text{for } l_k = 1 \\ \frac{1}{(\sigma_n)\sqrt{2\pi}} e^{-(c_k-c_{\mu_n})} & \text{for } l_k = 0 \end{cases} \quad (4)$$

where c_{μ_p} and c_{μ_n} , σ_p and σ_n are the means and standard deviations obtained by GentleBoost classification of the positive and negative training features respectively.

2.2. Modeling the Prior

The assignment of MRFs to sparse feature points is interesting. Our novel approach connects the points in a graph via Delaunay Triangulation, as can be seen in Fig. 1 (a). We can compute a feature point's Delaunay neighbors of degree n from this graph, allowing us to experiment with varying neighborhood sizes. Figs. 1 (b) and (c) demonstrate this idea. We can model the prior probability as

$$p_q(l_k|L_k) \propto \exp -\left\{ \sum_{s \in S} \lambda_{sk} |l_s \neq l_k| \right\} \quad (5)$$

where l_s are the labels of all sparse features linked to f_k by Delaunay connections. We choose to weight each term by $\lambda_{sk} = (1 + 1/d_{sk})$, where d_{sk} is the Euclidean distance from f_k to each neighbor, f_s . We exclude edge-crossing penalization from this prior, since objects can have some internal edges (e.g. faces have nose, mouth and eyes).

We seek to maximize the posterior probability associated with each feature point. This is equivalent to minimizing the energy, E_k , since $E(p) = -\log(p(\dots))$. Eqn. 1, simplified and expressed in terms of log energy becomes

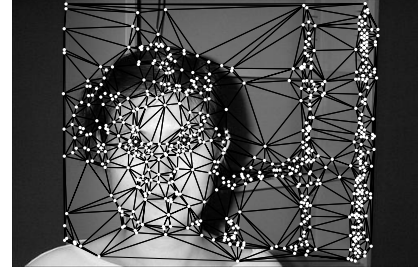
$$E_k^o(l = 1, 0) = \Lambda_p, \Lambda_n \{E_k(p_o) + \alpha E_k(p_q)\} \quad (6)$$

$$E_k^g(l = 1, 0) = \Lambda_p, \Lambda_n \{E_k(p_g) + \alpha E_k(p_q)\} \quad (7)$$

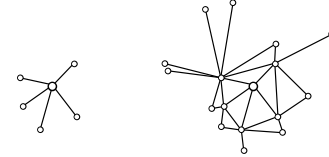
where $E_k(p_o)$ and $E_k(p_g)$ are the likelihood energies (either modeled by L-S or B-G), and $E_k(p_q)$ is the prior energy. Λ_p and Λ_n reflect the ratios of total energy associated with the positive and negative classes, and α signifies the relative influence of prior to observed knowledge. We compute the posterior using Iterated Conditional Modes (ICM) [9]. The ICM process works by minimizing the defined Gibbs Energy at each feature site, iteratively converging to a local minimum over all sites.

2.3. Rough Segmentation

Our Bayesian classification activates a tight network of positive feature points, f_p , on the object. We experiment with object segmentation by centering normalized Gaussian masks on each f_p , having variance v_p proportional to the scale, ω_p , of each feature. Applying a global threshold, t_p , to the test image yields a rough object segmentation, M_p .



(a) Delaunay Triangulation of SIFT points



(b) $n = 1$

(c) $n = 2$

Fig. 1. Feature points connected on a Delaunay graph, and $n = 1$ and $n = 2$ degree MRF neighbors of a feature point, where the point being evaluated is central to the neighborhoods.

3. EXPERIMENTAL SETUP

We choose the Caltech-4 face database [2] to test our algorithm. Although this database is not regarded as being particularly challenging [10], we can certainly use it to highlight some of the pros and cons of our Bayesian framework. The database is comprised of multi-sized images with varying lighting, including subjects with beards, glasses, some occlusion and cartoons.

We split the set of 435 images into training and test sets. The 218 training images are manually cropped for our supervised training approach. We create our own ground truth masks for remaining 217 test images which are "stricter" than the official Caltech ones - encapsulating the outlines of the faces with ellipses. We evaluate both our rough object segmentation and detection rate with these masks. We take 550 unmodified images from the Caltech background database as the negative training set. The default parameters specified in [1] are used to obtain around 40,000 positive and 142,000 negative 128-element SIFT descriptors from the training sets. GentleBoost is then trained with these features using 400 iterations of boosting.

During testing, we compare the rough segmentation mask, M_p and ground truth mask, M_{gt} for each image. Fig. 2 (b) show a merging of the two masks pertaining to Fig. 2 (a). White pixels indicate true positive areas, and gray pixels reveal false positive or false negative areas, outside or within M_{gt} respectively. A circle centered on, and with area equivalent to M_p is useful for visualizing detections.

3.1. Experiments

We compare the performance of the L-S and B-G likelihood models for object detection and segmentation in two experiments:

1. Varying the degree of the Delaunay neighborhood, $n = [1 : 5]$, with $\alpha = [0 : 20 : 100]$ (see Eqns. 6 and 7) to investigate the power of our unique spatial prior.
2. Testing the invariance of our model over image rotations of $[0 : 60 : 300]^\circ$ and scalings of $[0.6 : 0.2 : 1.4]$ times original image size.

The other parameters are fixed at $t_p = 0.6$, $v_p = (30 * \omega_p)$, $h = -1.5$, $\Lambda_p = 1$ and $\Lambda_n = 2$ (see Sec. 2). We have previously investigated the significance of these variables, but do not discuss these experiments here. Further results and graphs, along with the training, test and ground truth images are available from our webpage: www.deirdreoregan.com/OD_ICIP08.html.

3.2. Test Metrics

In terms of object detection, we define the criterion for a True Detection, TD , similarly to [5, 6]. $\%TD$, is the percentage of times this criterion is met in an experiment with the set of 217 test images.

$$\frac{area(M_p \cap M_{gt})}{area(M_p \cup M_{gt})} > 0.5 \Rightarrow TD \quad (8)$$

$$\%TD = \frac{\#TD}{217} \quad (9)$$

Since our detector does not evaluate windows, we cannot measure false positives as a percentage of all windows evaluated. Instead, we propose the metric of False Detection Area, FDA , leading to the alternative evaluation of $\%FDA$ over the entire test set.

$$FDA = area\{M_p \cap (M_{gt} \cup TDM)^c\} \quad (10)$$

$$\%FDA = \frac{\sum_j^{217} FDA_j}{\sum_j^{217} area\{(M_{gt_j} \cup TDM_j)^c\}} \quad (11)$$

where TDM is the True Detection Mask of any TD recorded for the image. The area under more than one TD per test image j is added to FDA_j .

Object segmentation results are evaluated by the metrics suggested in [4]. Formulated in terms of our own experiments, Recall, R , and Precision, P , are defined as

$$R = \frac{\sum_j^{217} area(M_{p_j} \cap M_{gt_j})}{\sum_j^{217} area(M_{gt_j})} \quad (12)$$

$$P = \frac{\sum_j^{217} area(M_{p_j} \cap M_{gt_j})}{\sum_j^{217} area(M_{p_j} \cap M_{gt_j}) + FDA_j} \quad (13)$$

4. RESULTS

Figs. 3 (a) and (b) graph object detection results for the two likelihood models. We plot $\%TD$ versus $\%FDA$ over all test images for neighborhood size, n varying with α from Eqns. 6 and 7. Fig. 3 (c) plots R versus $1 - P$, showing object segmentation results for B-G only.

Fig. 4 reveals the performance of both models in invariance tests with fixed $n = 3$ and $\alpha = 20$. The leftmost column shows the test image scaling, and the other entries average scores over all tested rotations. Breakdowns of all results are available on our webpage mentioned in Sec. 3, along with R versus $1 - P$ results for invariance tests.

4.1. Influence of the Prior

When the prior (spatial) energy has no influence (i.e. $\alpha = 0$), we can achieve a reasonable detection rate. When $\alpha > 0$ however, the value of n can influence a degradation ($n < 2$) or improvement ($n > 2$) in results, whereas the actual value of α is less critical. Neighborhood size $n = 3$ works well for both models, but B-G performs best overall achieving $\%TD = 87.6\%$ and $\%FDA = 0.03\%$ with $n = 3$

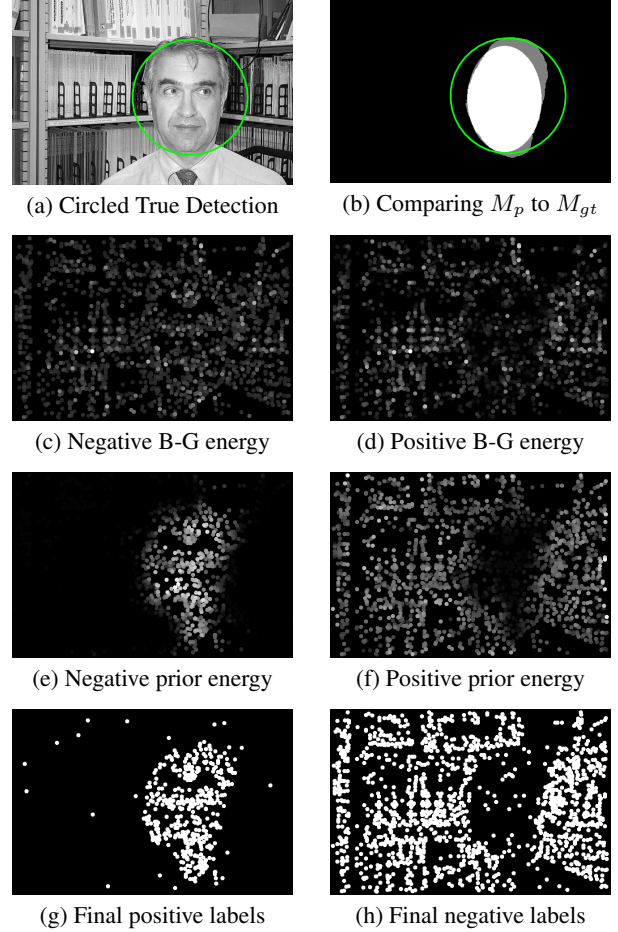


Fig. 2. A detection and segmentation, the point-wise B-G likelihood and spatial energies, and posterior feature labels dilated for clarity.

and $\alpha = 20$. The importance of the spatial energies is very clearly seen when juxtaposed with the B-G modeled likelihood energies in Figs. 2 (c-f). It is obvious that the final labeling, seen in Figs. 2 (g) and (h), has been positively influenced by inclusion of the prior.

Note that $\%TD$ in Figs. 2 (a) and (b) plummets while $\%FDA$ soars for $n > 4$. This is probably due to the fact that $area(M_p)$ becomes too large to be counted as TD (see Eqn. 8). These large, missed detections are then accumulated in $\%FDA$. This hypothesis is supported by Fig. 3 (c), where we see that as n and α increase, segmentation R improves as P takes an increasing hit. In the absence of spatial energy, B-G seems more robust than L-S. Furthermore, B-G with $\alpha = 20$ and $n = 3$ or $n = 4$ has the best trade-off between R and P .

4.2. Invariance of the Framework

It is clear from Fig. 4 that our framework is robust to a image scaling within a range, but not completely scale invariant. We attribute this to fixing the size of the Delaunay MRF neighborhood, n . Perhaps we could vary the extent of n relative to the scale of the feature being evaluated, or combine results over a range of n instead. This is an interesting problem for future work. Although not reflected in Fig. 4, similar results are observed over rotations at a fixed image scale, so

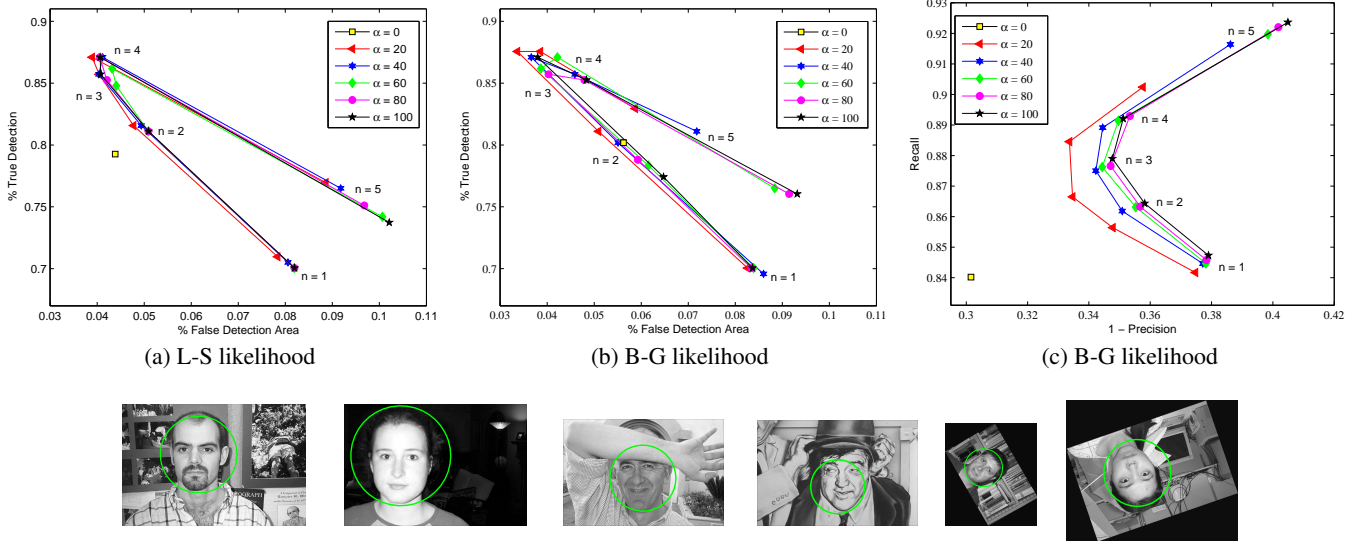


Fig. 3. Comparison of the different likelihood models in detection tests, B-G evaluated for object segmentation, and various circled True Detections using B-G in invariance tests.

our framework is rotation invariant. L-S performs best in one test with a result of $\%TD = 89.4\%$ and $\%FDA = 0.01\%$ at a rotation of 300° and image scale 1.2.

| Scale | L-S likelihood | | B-G likelihood | |
|-------|----------------|---------|----------------|---------|
| | $\%TD$ | $\%FDA$ | $\%TD$ | $\%FDA$ |
| 0.6 | 52.3 | 0.09 | 53.2 | 0.09 |
| 0.8 | 75.0 | 0.05 | 75.3 | 0.05 |
| 1 | 83.6 | 0.03 | 83.4 | 0.03 |
| 1.2 | 88.6 | 0.02 | 87.3 | 0.06 |
| 1.4 | 85.7 | 0.02 | 86.0 | 0.02 |

Fig. 4. Detection invariance test results for both likelihood models

5. CONCLUSION

We have devised a way to inject geometric context into a Bayesian solution for the classification of sparse local object features with the aim of object detection and segmentation. By connecting the feature points on a graph formed by Delaunay Triangulation, we can obtain a unique MRF neighborhood for each point, allowing us to infer spatial energies. Interestingly, we have derived likelihood energies from a point-wise machine learning algorithm, and investigated two methods of modeling the object and background likelihood distributions.

Our Log-Sigmoid (L-S) and Bi-Gaussian (B-G) likelihood models perform moderately at simple feature classification. Augmented with a spatial prior, however, the framework becomes much more powerful, yielding good object detection and segmentation results that are largely rotation and semi-scale invariant. As a less naïve likelihood model, B-G seems to perform slightly better than L-S, but it is clear that inclusion of contextual spatial inference is significantly more important than the choice of likelihood model.

Improving robustness to variations in scale would allow for evaluation of the system’s performance on more challenging tasks. We hypothesize that full scale invariance is achievable through further

research on the assignment of MRFs via Delaunay Triangulation or some other meshing scheme, highlighting potential further work in this area. We contribute this interesting problem to the field of local feature-based object detection and localization.

6. REFERENCES

- [1] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [2] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” in *CVPR*, 2003.
- [3] P. Moreno et al, “A comparative study of local descriptors for object category recognition: Sift vs hmax,” in *IbPRIA*, 2007.
- [4] S. Agarwal and D. Roth, “Learning a sparse representation for object detection,” in *ECCV*, 2002.
- [5] K. Mikolajczyk, B. Leibe, and B. Schiele, “Multiple object class detection with a generative model,” in *CVPR*, 2006.
- [6] A. Opelt, A. Zisserman, and A. Pinz, “Fusing shape and appearance information for object category detection,” in *BMVC*, 2006, vol. 1, pp. 117–216.
- [7] W. et al, “Object class recognition using multiple layer boosting with heterogeneous features,” in *CVPR*, 2005.
- [8] A. Vezhnevets and V. Vezhnevets, “Modest adaboost - teaching adaboost to generalize better,” in *Graphicon*, 2005, <http://research.graphicon.ru/machine-learning/gml-adaboost-matlab-toolbox.html>.
- [9] J. Besag, “On the statistical analysis of dirty pictures,” *Journal of the Royal Statistical Society, Series B*, no. 48, pp. 259–279, 1986.
- [10] J. Ponce et al, “Dataset issues in object recognition,” *Toward Category-Level Object Recognition*, Springer-Verlag LNCS, pp. 29–48, 2006.