6

Annotation Localization using a Weakly Supervised Model for Top-down Visual Saliency

This chapter describes the concept of Discriminant Saliency, its principles and its suitability for handwritten annotation extraction in printed documents. Section 6.1 describes Discriminant Saliency and its computational models. Section 6.2 gives an overview of the dataset created for the problem. Section 6.3 describes the features used in our work. In Section 6.4, further investigates the threshold protocols adopted for the given problem. The results are illustrated in the subsequent Section 6.5 on our dataset and two other standard datasets, namely, IAM and PRIma-NHM. We also present a comparison of our work with a discriminative classifier in the Section 6.6. Finally, Section 6.7 concludes the chapter.

6.1 INTRODUCTION

We are quickly able to interpret a scene, even if it is highly cluttered. This happens not because our perceptual system is powerful enough to quickly process the visual sensory input from the entire scene, but because there are saliency mechanisms deployed by the neural circuitry which can detect *salient* regions, which are the regions carrying higher semantic content that is most relevant to scene understanding. These salient regions constitute the relevant subsets of sensory information and are prioritized for further analysis by the visual cortex. The detection of salient regions helps in quickly recognizing the important objects and subsequently interpreting the scene.

Research related to attentional/saliency mechanisms has been carried out in psychology and neurophysiology. Experiments in psychophysics and neural signal recording have contributed to the understanding of saliency, however, such knowledge cannot be readily translated into computational models and principles for optimal saliency computation by computer vision algorithms. For tasks such as object detection, recognition and tracking, many computer vision algorithms rely on extraction of interest points, which can be considered as salient points. The purpose of these interest point detectors is to reduce the computational burden on the subsequent processing stages, which can focus on detailed processing of information around the interest points. Such interest point detectors can be tied to visual features of two types:

- Features having better stability against transformation and having mathematically well-defined properties. For example, such as edges, corners, contours, local symmetry, blobs, etc.
- Features that depend on generic principles pertaining to image complexity. For example, variance of Gabor filter responses over multiple orientations, entropy of the distribution of local intensities, values of wavelet decomposition coefficients, etc.

This translates the question of what constitutes optimally salient (or the optimality criteria for saliency), into optimal detection of specific visual attributes depicting stability or computing feature values depicting image complexities. Though these salient point detectors give good saliency judgements and are useful in many applications, these saliency judgements are not influenced by the recognition goal. Therefore, the extracted interest points may not help the object detection/recognition modules. In other words, the extracted interest points are not optimal from the point of view of recognition. In fact, it may happen that there are no interest points on the object to be detected and therefore such interest points may not offer any advantage over dense sampling or random sampling. Such saliency computation which is stimulus driven, but not aligned with recognition goal is called as *bottom-up* saliency.

Attentional mechanisms which are driven by recognition goals contribute to the *top-down* component of saliency. Top-down mechanisms are weak classifiers that extract regions of the scene that are likely to contain the object to be recognized. Computational models for top-down saliency need to be efficient so that the candidate regions likely to contain the object of interest can be quickly extracted. Such salient regions deserve attention or further processing by the brain to establish presence of the desired object.

Optimal saliency, i.e. what (properties) constitute a salient region, now gets tied to the recognition problem. [Gao *et al.*, 2009] proposed a model for recognition driven top-down saliency called as *discriminant saliency*. The model allows the design of computationally fast weak classifiers which can be trained in a weakly supervised setting. The principle of discriminant saliency specifies two fundamental tasks:

1. Feature selection: This task involves selecting features that best distinguish the object class to be recognized from other possible objects in the scene. This definition for the feature selection task translates into a computational principle of classification with minimal expected probability of error.
2. Saliency detection: This task involves assigning saliency values to the the extracted feature components and taking a call on which regions on the image are salient.

6.1.1 Computational principles for top-down saliency

The computational principle for saliency is closely related to some of the previously proposed principles for the organization of perceptual systems. These principles can be translated into a computational models for saliency, as follows:

A) Maximization of information transmission across perceptual layers (infomax): This selects optimal features as the ones which are maximally informative of the presence/absence of the target class in the field of view.
B) Inference by detection of suspicious coincidences: This selects optimal features as those whose observation is most suspicious in the absence of the target class. This principle was contributed by Barlow [Barlow, 1994; Bartlett *et al.*, 2002].
C) Classification with minimal uncertainty: This principle selects features which minimize the uncertainty about the presence/absence of the target class.
D) Discriminant Saliency: This principle selects features which give the minimum probability of error. Thus, it specifies a discriminant principle for the design of top-down saliency computation.

[Gao *et al.*, 2009] asserted that the methods inspired by the first 3 principles *can* give saliency measures that are nearly optimal w.r.t. the computational principle of discriminant saliency, i.e. in the minimum probability of error sense. [Gao *et al.*, 2009] investigated which of these methods allow a computationally more efficient method and found that the Barlow's principle of inference by detection of suspicious coincidences gives the most efficient method under reasonable simplifications. By using this method, they developed efficient algorithms for feature selection and saliency detection.

The computed saliency value for the interest points helped identify the high saliency value locations that deserve the focus of attention and the remaining low saliency value locations which can be pruned out for further processing. Discriminant saliency refers to a decision-theoretic interpretation of perception. It hypothesizes that perception involves taking a decision regarding which sensory input from the surrounding environment is salient. Taking optimal decisions would correspond to having minimum probability of error. Discriminant saliency refers to making a decision of classifying the stimuli into two classes: target and null hypothesis. Saliency refers to the confidence with which a location in the scene can be classified as containing the target. The decision making hypothesis (discriminant saliency) can be applied to top-down as well as bottom-up forms of attentional mechanisms. For bottom-up saliency detection, the `decision-making' can be made part of the center-surround image processing. For top-down saliency detection, the decision-making can be adapted to any specification of target stimuli and null hypothesis. This translates to learning a one-vs-all classification model, where the object class of interest constitutes the target stimuli and all the other object classes constitute the stimuli considered as the null hypothesis.

Vision algorithms translate the visual stimuli into features. Salient features are the features which can well discriminate the target class from the other object classes (null hypothesis). Appropriate image attributes are selected depending on the target to be recognized. The saliency measure corresponds to the confidence of classifying a portion of the scene into the target class. Salient locations are the ones which compute the highest confidence in classifying the target. It is worth mentioning that the salient features identified in the previous step will vary in their confidence while declaring a given location as salient (possibly containing the target). This variation happens because of the varying recognition context. Features that are effective in classifying the target in a given context (background) may not remain effective as the recognition context changes. Instead, another set of features may be able to better discriminate the target from its changed background.

6.1.2 Optimal feature selection for discriminant saliency

We now review the computational principles that can be used to derive optimality criteria for feature selection. Given an observation $\mathbf{x}$ that lies in the feature space $\mathscr{X}$, the following problems need to be addressed.

Task 1: How $\mathbf{x}$ can be classified as salient or non-salient?
Task 2: What is the confidence value associated with the classification of $\mathbf{x}$?
Task 3: How to choose the optimal feature space $\mathscr{X}$?

We now discuss the three computational principles, (i) Bayesian Decision Theory, (ii) Principle of minimum uncertainty, and (iii) Barlow's principle of suspicious coincidences, for top-down saliency that guide and specify the mathematical expressions for the 3 tasks.

1. Bayesian Decision Theory: The Bayes classifier models $P_{Y|\mathbf{X}}(i|\mathbf{x})$ where $i \in \{0,1\}$ denotes absence of target ($i=0$), or presence of target ($i=1$).

   - Task 1: The decision regarding whether a feature $\mathbf{x}$ belongs to the class $i=0$ or $i=1$ is based on the outcome $g^*(\mathbf{x})$, which is

   $$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \tag{6.1}$$

   - Task 2: The maximum value of $P_{Y|\mathbf{X}}$ for a class is taken as the confidence measure for classification

   $$c^*(\mathbf{x}) = \max_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \tag{6.2}$$

- Task 3: The optimal choice for $\mathscr{X}$ is the one that maximizes the expected confidence on the classification decisions

$$C^* = E_{\mathbf{X}}\left[c^*(\mathbf{x})\right] = E_{\mathbf{X}}\left[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})\right] \tag{6.3}$$

The computed value C^* is the *feature selection cost* for the Bayesian decision theory. Maximizing this expected confidence, is equivalent to minimizing the Bayes error $1 - E_{\mathbf{X}}\left[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})\right]$

2. Principle of Minimum Uncertainty:

- Task 1: The decision rule gives the outcome $g'(\mathbf{x})$ defined as follows

$$g'(\mathbf{x}) = \arg\max_i \log P_{Y|\mathbf{X}}(i|\mathbf{x}) \tag{6.4}$$

- Task 2: The confidence value $c'(\mathbf{x})$ in declaring a feature $\mathbf{x}$ as salient is obtained by relaxing the decision rule to the mean (i.e. taking an expectation)

$$c'(\mathbf{x}) = \sum_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \log P_{Y|\mathbf{X}}(i|\mathbf{x}) \tag{6.5}$$

The right hand side expression can be recognized as the negative of entropy, giving

$$c'(\mathbf{x}) = -H(Y|\mathbf{X} = \mathbf{x}) \tag{6.6}$$

- Task 3: The optimal choice for $\mathscr{X}$ is the one that minimizes the expected confidence $E_{\mathbf{X}}\left[H(Y|\mathbf{X} = \mathbf{x})\right] = -H(Y|\mathbf{X})$ which corresponds to the minimization of the uncertainty of the classification decision. The feature selection cost is, therefore, $-H(Y|\mathbf{X})$

We see that the decision rule $g'(\mathbf{x})$ is equivalent to $g^*(\mathbf{x})$ and the confidence measure $c'(\mathbf{x})$ can be seen as the relaxation to the mean of the decision rule, i.e., mean of $\log P_{Y|\mathbf{X}}(i|\mathbf{x})$, or $E_{\mathbf{X}}\left[\log P_{Y|\mathbf{X}}(i|\mathbf{x})\right]$.

3. Barlow's principle of suspicious coincidences

- Task 1: The decision rule proposed for this principle yields the outcome

$$g''(\mathbf{x}) = \arg\max_i \log \frac{P_{\mathbf{X},Y}(i,\mathbf{x})}{P_Y(i)P_{\mathbf{X}}(\mathbf{x})} \tag{6.7}$$

- Task 2: Relaxation of the decision rule to the mean gives the confidence measure $c''(\mathbf{x})$ for classification

$$c''(\mathbf{x}) = \sum_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \log \frac{P_{\mathbf{X},Y}(i,\mathbf{x})}{P_Y(i)P_{\mathbf{X}}(\mathbf{x})} \tag{6.8}$$

A simplification of the right hand side expression in terms of mutual information yields $c''(\mathbf{x}) = I(Y;\mathbf{X} = \mathbf{x})$

- Task 3: Taking expectation of the confidence measure $c''(\mathbf{x})$ gives

$$\int \sum_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \log \frac{P_{\mathbf{X},Y}(i,\mathbf{x})}{P_Y(i)P_{\mathbf{X}}(\mathbf{x})} d\mathbf{x} \tag{6.9}$$

which is the familiar expression for $I(\mathbf{X};Y)$

$$I(\mathbf{X};Y) = \sum_i \int P_{Y|\mathbf{X}}(i|\mathbf{x}) \log \frac{P_{\mathbf{X},Y}(i|\mathbf{x})}{P_Y(i)P_{\mathbf{X}}(\mathbf{x})} d\mathbf{x} \tag{6.10}$$

The optimal choice for $\mathscr{X}$ is the one that minimizes the expected confidence, i.e. $I(Y;\mathbf{X})$. Thus, the feature selection cost for this principle is $I(Y;\mathbf{X})$.

Notice that the mutual information $I(Y;\mathbf{X})$ between the class label Y and the feature vector $\mathbf{X}$ can be also written as

$$I(Y;\mathbf{X}) = H(Y) - H(Y|\mathbf{X}) \tag{6.11}$$

We see that maximizing $I(\mathbf{X};Y)$ with respect to $\mathbf{X}$ is equivalent to maximizing $-H(Y|\mathbf{X})$. Therefore, this computational principle gives a feature selection cost that is the same as the one given by the principle of minimum uncertainty even though the decision rules are different. This happens because a relaxation to the mean is applied to the Barlow's decision rule. This criterion for feature selection is referred to as the Infomax criterion.

Out of the given feature selection criteria, we need to adopt the one which is computationally least expensive.

It is seen that certain simplifications applied to the infomax feature selection criterion result in a formulation that is computationally parsimonious. Rewriting the feature vector $\mathbf{X}$ more explicitly in terms of its k feature components

$$\mathbf{X}_{1:k} = \{X_1, X_2, ..., X_k\},$$

the selection criterion of Infomax can be rewritten as:

$$I(Y;\mathbf{X}) = \sum_k I(Y;\mathbf{X}_k) + \sum_k [I(X_k;\mathbf{X}_{1,k-1}|Y) - I(X_k;\mathbf{X}_{1,k-1})] \tag{6.12}$$

Research has shown [Gao *et al.*, 2008] that some statistical properties of band pass filters, such as wavelet coefficients, extracted from natural images exhibit strongly consistent patterns of dependency across a wide range of natural image classes. However, such dependencies carry little information about the image class. This implies that the mutual information between features X_k and $\mathbf{X}_{1:k-1}$ given the knowledge of Y (i.e. $I(X_k;\mathbf{X}_{1,k-1}|Y)$) and the same same mutual information without the knowledge of Y (i.e. $I(X_k;\mathbf{X}_{1,k-1})$) are almost similar. Thus the second term of Eq. 6.12, which signifies the discriminant information is much smaller and can be ignored, thus yielding

$$I(\mathbf{X};Y) \approx \sum_k I(Y;\mathbf{X}_k) \tag{6.13}$$

Reverse analysis reveals that this approximated feature selection cost corresponds to the expectation of a new confidence measure $c'''(\mathbf{x})$ given as

$$c'''(\mathbf{x}) = \sum_k I(Y|X_k = x_k) \tag{6.14}$$

Further reverse analysis shows that the confidence measure $c'''(\mathbf{x})$ corresponds to relaxation to the mean of the following decision rules:

$$g''_k(x) = \arg\max_i \log \frac{P_{Y,X_k}(i,x)}{P_{X_k}(x)P_Y(i)}, \quad k \in \{1,...,K\} \tag{6.15}$$

Each feature channel X_k gives its individual decision rule using $g''_k(x)$. The decision rule applied to the individual channels is considered as the marginal decision rule. Maximizing the approximated feature selection criterion $I(\mathbf{X};Y) \approx \sum_k I(Y;\mathbf{X}_k)$ is easier because each term being a mutual information is positive. Therefore we can select k features having the highest values of $I(X_k;Y)$. Computing $Y(Y;\mathbf{X} = \mathbf{x})$ using Eq.6.8 is simple for bandpass filters extracted from natural images.

This kind of computationally parsimonious feature selection is applicable for only the Barlow's principle. If a similar simplification is applied to the Bayes rule for feature selection, then the approximation would be poor. For example, the feature selection cost for the minimum uncertainty principle cannot be well approximated as $H(Y|\mathbf{X}) \approx \sum_k H(Y|X_k)$ and therefore does not allow marginal Bayes decision rules such as

$$g_k^*(x) = \arg\max_i \log P_{Y|X_k}(i|x), \quad k \in \{1, ..., K\} \tag{6.16}$$

Therefore, it can be concluded that the principle of suspicious coincidences yields computationally parsimonious feature selection cost and marginal decision rules. An interesting observation is that these marginal decision rules are more consistent with the psychophysics of human saliency, since humans find it easier to distinguish between target and background along a single feature, but not along conjunction of features, such as color and orientation. The holistic confidence measure (Eq.6.14) is the sum of the marginal confidence measures for the features.

6.1.2.1 Implementation of discriminant saliency

Deploying the discriminant saliency model requires addressing first the task of choosing the optimal features (Task 3), formulating a saliency measure for a feature (Task 2), and finally taking a decision on whether a feature is salient or not (Task 1).

Task 3: How to choose the optimal feature space $\mathscr{X}$? (feature selection task)

Salient features are the ones which pass the following test:

$$\mathscr{S}_k = \{x_k \mid H(X_k|Y=1) > H(X_k|Y=0)\} \tag{6.17}$$

This follows from the observation that discriminant saliency selects features which are present in the class of interest ($Y = 1$) and mostly absent in the null hypothesis($Y = 0$). This translates into a distribution of features which is narrower and centered around 0 if the feature is absent from the null hypothesis ($Y = 0$), and leads to a broader distribution if the feature is present in the target class ($Y = 1$). A broader distribution contributes a higher value of entropy than a narrower distribution and therefore $H(X_k|Y = 1)$ is larger than $H(X_k|Y = 0)$ for the salient features. Eq.6.17 can be written as

$$\mathscr{S}_k = \left\{ x_k \left| \frac{P_{Y,X_k}(1,x_k)}{P_Y(1), P_{X_k}(x_k)} > \frac{P_{Y,X_k}(0,x_k)}{P_Y(0)P_{X_k}(x_k)} \right. \right\} \tag{6.18}$$

which can be simplified as

$$\mathscr{S}_k = \left\{ x \left| P_{X_k|Y}(x|1) > P_{X_k|Y}(x|0) \right. \right\} \tag{6.19}$$

Algorithmically, this task requires as input a set of images $\mathscr{T}_1$ that belong to the target class, a set of images $\mathscr{T}_0$ that belong to the null hypothesis, and a set of N features $X_k, k \in \{1, ..., N\}$. The target number of features to be selected is given as K. Selection is based on the value computed for $I(X_k, Y)$ for certain features X_k, as follows:

1. For each feature, compute $P_{X_k|Y}(x_k|i)$ and $P_{X_k}(x_k)$
2. Features which pass the test given by Eq.6.17 or 6.19 are retained and others are discarded.
3. For the retained features, compute $I(X_k, Y)$

4. The K features that give the largest values for $I(X_k, Y)$ are selected.

Task 2: What is the confidence value associated with the classification of $\mathbf{x}$? (Saliency computation task)

The saliency value for the kth feature x is computed as:

$$S_k(x_k) = \begin{cases} I(Y|X_k = x_k), & \text{if } x_k \in \mathscr{S}_k \\ 0, & \text{otherwise} \end{cases} \tag{6.20}$$

The overall saliency measure $S_D(\mathbf{x})$ is the sum of the saliency measure over all the feature channels.

$$S_D(\mathbf{x}) = \sum_{k=1}^{K} S_k(x_k) \tag{6.21}$$

From the marginal decision rules given in Eq.6.15, the saliency measure for the individual features x_k can be interpreted as the (log) degree of suspicion

$$S_k(x_k) = \left\langle \log \frac{P_{Y,X_k}(i, x_k)}{P_Y(i) P_{X_k}(x_k)} \right\rangle \tag{6.22}$$

where $\langle f(x) \rangle = \sum_i P_{Y|X}(i|x) f(x)$

Consider a set of interest points $\mathrm{I}_1, \ldots. \mathrm{I}_M$ extracted from a test image $\mathscr{I}$. The task of computing the saliency values for the interest points involves computing $S_D(\mathbf{x}_m)$ where $\mathbf{x}_m$ is the feature extracted for each location I_m. Algorithmically, this involves evaluating the selected features X_k at the given locations as per the following steps:

1. Given the feature value x_{km} of feature component X_k computed at location I_m, compute $P_{X_k|Y}(x_{km}|i)$ for $i \in \{0, 1\}$.
2. Compute $S_k(x_{km})$ using Eq.6.20

Task 1: How $\mathbf{x}$ can be classified as salient or non-salient? (Declaring salient regions)

Instead of making a hard classification for the mth region/interest point characterized by feature $\mathbf{x_m}$ as salient or non-salient, we can simply order the regions/interest points by decreasing discriminant saliency values $S_D(\mathbf{x}_m)$. Otherwise a suitable threshold value can be adopted to classify regions as salient or non-salient.

6.1.2.2 Estimating the models

An implementation for Task 3 requires constructing models for $P_{X_k|Y}(x_k|i)$, $P_{X_k}(x_k)$ and $I(X_k, Y)$, and also a way to do the condition check $H(X_k|Y = 1) > H(X_k|Y = 0)$. An implementation of Task 2 requires constructing models for $I(Y|X_k = x_k)$.

1. **Computing $P_{X_k|Y}(x_k|i)$, $P_{X_k}(x_k)$**

We make an assumption that the extracted features X_k have a probability distribution that can be well approximated by a Generalized Gaussian Distribution (GGD). A GGD can be defined using two parameters α and β as follows:

$$P_X(x; \alpha, \beta) = \frac{\beta}{2\alpha\,\Gamma(1/\beta)} \exp\left\{ -\left(\frac{|x|}{\alpha} \right)^{\beta} \right\} \tag{6.23}$$

where the Gamma function $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad t > 0$. More specifically, the parameter α governs the scale of the function and β governs the shape (rate of decay from the peak value) of the distribution. Setting $\beta = 1$ gives the the Laplacian subfamily of GGD and setting $\beta = 2$ gives the Gaussian subfamily. The parameters α and β can be estimated using methods such as the method of [Sharifi and Leon-Garcia, 1995], maximum likelihood [Do and Vetterli, 2002], and minimum mean square error [Huang and Mumford, 1999]. Following [Gao *et al.*, 2009] we adopt the method of moments that exploits the relations of α, β with two quantities: variance σ and kurtosis κ of the distribution of X.

$$\sigma^2 = \frac{\alpha^2 \Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})} \quad \text{and} \quad \kappa = \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})} \tag{6.24}$$

The quantities σ and κ are estimated from image data, as follows:

$$\sigma^2 = E_X\left[(X - E_X[X])^2\right] \quad \text{and} \quad \kappa = \frac{E_X\left[(X - E_X[X])^4\right]}{\sigma^4} \tag{6.25}$$

2. **Computing** $I(X_k, Y)$

$$I(\mathbf{X}; Y) = \sum_i P_Y(i)\ KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i) || P_{\mathbf{X}}(\mathbf{x})\right] \tag{6.26}$$

yielding,

$$I(X_k; Y) = \sum_i P_Y(i)\ KL\left[P_{X_k|Y}(x_k|i) || P_X(x_k)\right] \tag{6.27}$$

where $KL[p||q] = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$ is the Kullback-Leibler (KL) divergence between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$

The KL divergence between two GGD distributions $P_X(x; \alpha_1, \beta_1)$ and $P_X(x; \alpha_2, \beta_2)$ can be written as:

$$KL[P_X(x;\alpha_1,\beta_1)||P_X(x;\alpha_2,\beta_2)] = \log\left(\frac{\beta_1 \alpha_2 \Gamma(1/\beta_2)}{\beta_2 \alpha_1 \Gamma(1/\beta_1)}\right) + \left(\frac{\alpha_1}{\alpha_2}\right)^{\beta_2} \frac{\Gamma((\beta_2+1)/\beta_1)}{\Gamma(1/\beta_1)} - \frac{1}{\beta_1} \tag{6.28}$$

3. **Computing** $I(Y|X_k = x_k)$

The closed form expression for $I(Y|X_k = x_k)$ is given as

$$I(Y; X_k = x_k) = s\left[g(x_k)\right] \log \frac{s\left[g(x_k)\right]}{P_Y(1)} + s\left[-g(x_k)\right] \log \frac{s\left[-g(x_k)\right]}{P_Y(0)} \tag{6.29}$$

where $s(x) = (1 + e^{-x})^{-1}$ is a sigmoid function, and

$$g(x_k) = \left(\frac{|x_k|}{\alpha_0}\right)^{\beta_0} - \left(\frac{|x_k|}{\alpha_1}\right)^{\beta_1} + \log\left(\frac{\alpha_0 \beta_1 P_Y(1) \Gamma(1/\beta_0)}{\alpha_1 \beta_0 P_Y(0) \Gamma(1/\beta_1)}\right) \tag{6.30}$$

4. **Evaluating** $H(X_k|Y=1) > H(X_k|Y=0)$

   Using the closed form expression for $H(X_k|Y=i)$, which is

$$H(X_k|Y=i) = \frac{1}{\beta_i} + \log \frac{2\alpha_i \Gamma\left(\frac{1}{\beta_i}\right)}{\beta_i}, \tag{6.31}$$

   we can simplify the condition check $H(X_k|Y=1) > H(X_k|Y=0)$ as

$$\log\left(\frac{\alpha_1}{\alpha_0}\right) > \left(\frac{1}{\beta_0} - \frac{1}{\beta_1}\right) + \log \frac{\Gamma\left(\frac{1}{\beta_0}\right)\beta_1}{\Gamma\left(\frac{1}{\beta_1}\right)\beta_0} \tag{6.32}$$

6.2 DATASET

To evaluate the method of Discriminant Saliency (DS) for annotation localization we use the same dataset of images as mentioned in the previous chapter 5 in Section 5.3.

6.3 FEATURE EXTRACTION

We use the same feature set as mentioned in the previous chapter 5 in Section 5.4.

6.4 SALIENCY THRESHOLD DETAILS

In order to generalize the detection system for a variety of documents we compute the threshold according to the statistics of the input image as:

$$Threshold_{Saliency} = Threshold_{scalar} \times mean(\mathbf{S_k})$$

This $Threshold_{scalar}$ is dependent on the number of annotations present in the document. It is usually varied from $1 \leq Threshold_{scalar} \geq 5$. For pages with lesser number annotations the $Threshold_{scalar}$ must be high.

6.5 EXPERIMENTS AND RESULTS

We have defined four sets of training and testing experiments:

Set 1: The objective of this set of experiments is to localize *all* annotations. In this setting, the model is trained with the images comprising all the annotations.

Set 2: The objective of this set of experiments is to localize the individual annotations in a multi-annotated document. In this setting, the model is trained with the images comprising only individual annotations.

Set 3: The objective of this set of experiments is to localize textual annotations in a test document consisting symbolic annotations. In this setting, the model is trained with the images comprising only textual annotations with printed text as background.

Set 4: The objective of this set of experiments is to localize individual annotations in a test document not consisting other types of annotations. In this setting, the model is trained with the images comprising only individual annotations.

6.5.1 Accuracy metrics for performance evaluation

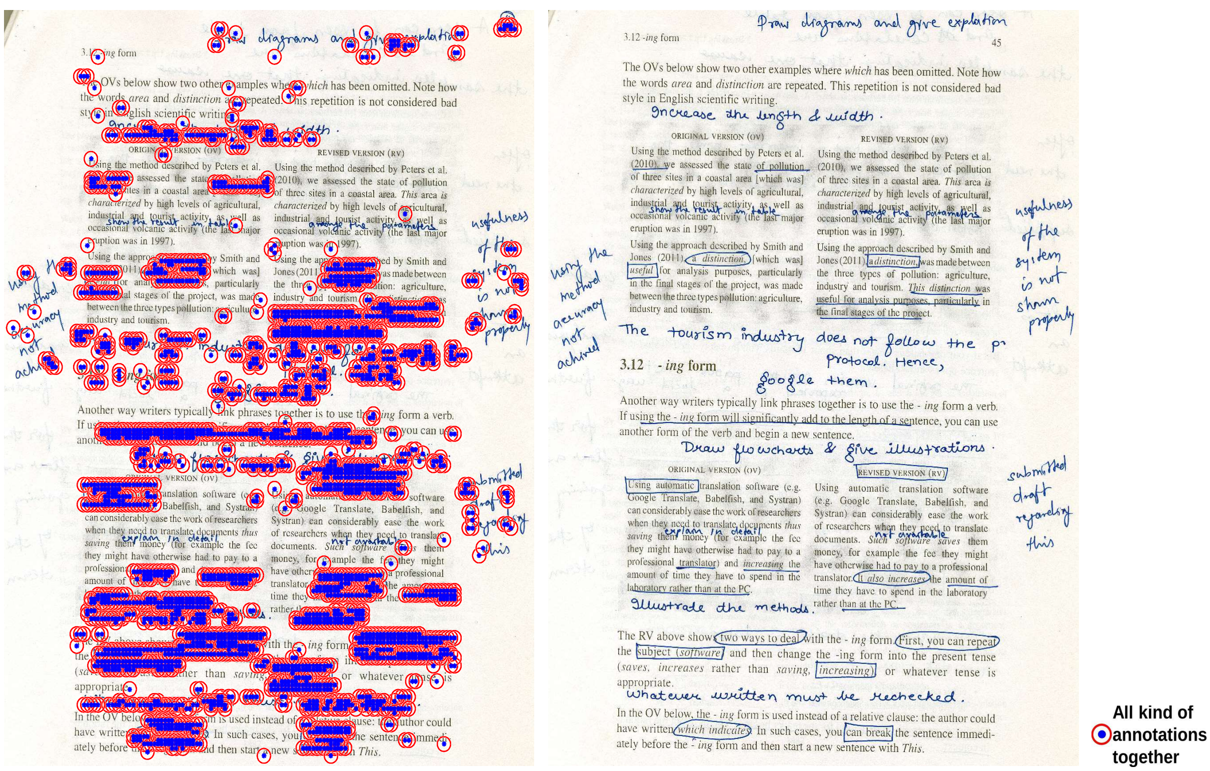
In order to evaluate the DS model for annotation localization we use the same performance measures as described in the previous chapter 5 in Section 5.6.1.

6.5.2 All Annotation vs Printed Text

To localize all annotations together in a document, discriminant saliency method produces a recall of 0.58 for annotations and 0.82 precision for printed text. Table 6.1 and Figure 6.1 elaborates the results.

Table 6.1 : Set 1: Annotation localization when the dictionary is trained on images comprising all annotations and the testing is performed on similar images.

Annotation Category	Threshold Scalar	Accuracy	Recall	Precision	F1 Score	Execution Time (sec)
All (Fig. 6.1)	2.5	80.23%	.58	.82	.68	26.86



(a) Highlighted all Categories of Annotations Regions as Salient Objects

(b) Original Image

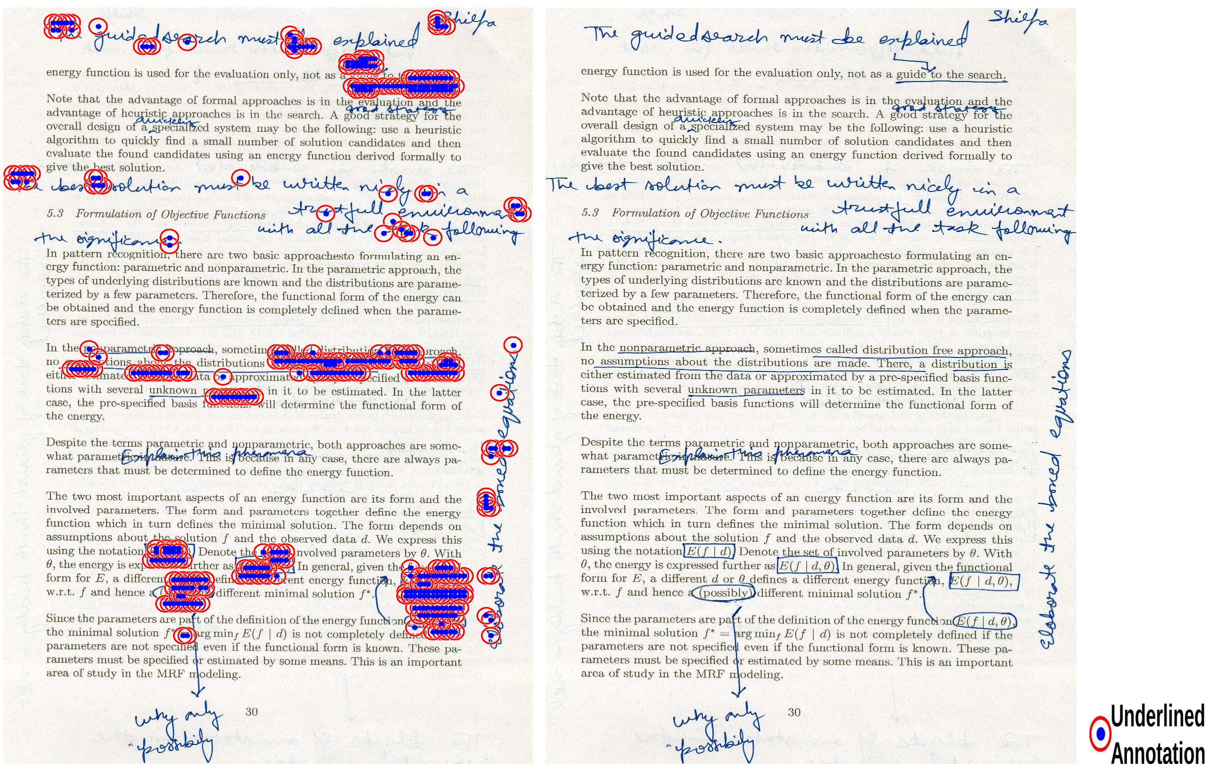
Figure 6.1 : Set 1: All kinds of Annotation Localization in Documents using DS.

6.5.3 Category-wise Annotation vs Printed Text

To localize *specific* annotations in a multi-annotated document each discriminant saliency model is trained with the images comprising only individual annotations. For underlined annotations, our model achieves a recall of 0.81 and precision of 0.47. For marginal text annotations, it produces 0.79 and 0.70 as recall and precision rates. In a similar manner, for encircled annotations, our model produces the recall and precision as 0.84 and 0.63 respectively, while for inline annotations, recall of 0.18 and precision of 0.16 is produces. Table 6.2 illustrates the results for localizing individual annotations in a multi-annotated document. In the third set of experiments,

Table 6.2 : Set 2: Annotation localization when the dictionary is trained on individual annotations and the testing is performed on the images containing all the annotations.

Annotation Category	Threshold Scalar	Accuracy	Recall	Precision	F1 Score	Execution Time (sec)
Underline (Fig. 6.2)	5	94.02%	.81	.47	.60	17.05
Marginal Text (Fig. 6.3)	5	95.04%	.7920	.7045	.7456	8.87
Encircled (Fig. 6.4)	5	95.81%	.84	.63	.72	17.11
Inline (Fig. 6.5)	6	82.03%	.18	.16	.17	8.61



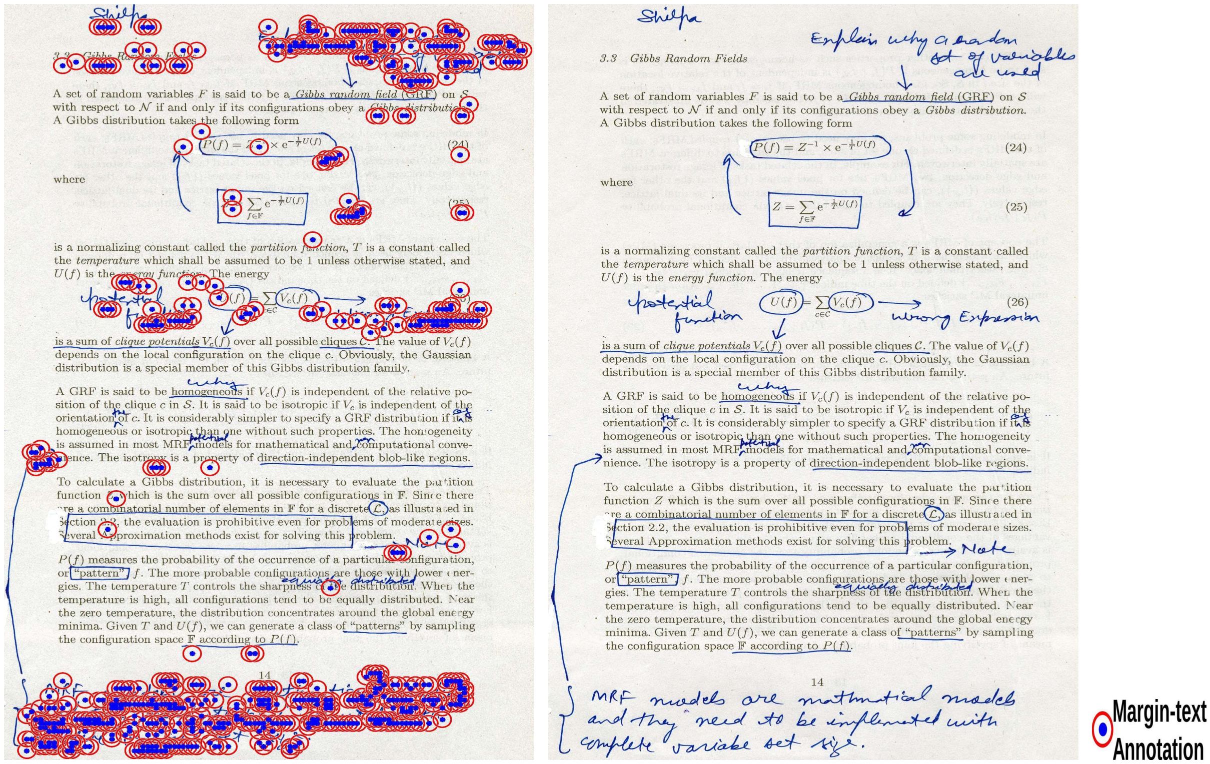
(a) Highlighted Underlined Regions as Salient Objects

(b) Original Image

Figure 6.2 : Set 2: Underlined Region Localization in Multi-annotated Images using DS

we localized *textual* and *symbolic* annotations separately in a multi-annotated document. In this setting, the model is trained with the images comprising only individual annotations. For such set of experiments discriminant saliency shows impressive results and produces a recall and precision of 0.57 and 0.78 to locate textual annotations. It also produces a recall of 0.71 and precision 0.83 of to locate symbolic annotations on a multi-annotated document. Table 6.3 depicts the results for localizing only textual and symbolic annotations in a test document.

In the fourth set, *specific* annotations in a single-class annotated document are localized. For underlined annotations, discriminant saliency achieves a recall of 0.68 and precision of 0.94



(a) Highlighted Marginal Annotated Textual Regions as Salient Objects

(b) Original Image

Figure 6.3 : Set 2: Marginal Annotation Region Localization in Multi-annotated Images using DS.

for annotations and printed text respectively. For marginal text annotations, it produces 0.91 and 0.87 as recall and precision rates. In a similar manner, for encircled annotations, a recall of .87 and precision of .70 is obtained for annotations and printed text. For inline annotations, our model obtains a recall of 0.49 and precision of 0.52 for annotations and printed text. Table 6.4 presents the results for localizing individual annotations in a single-class annotated test document.

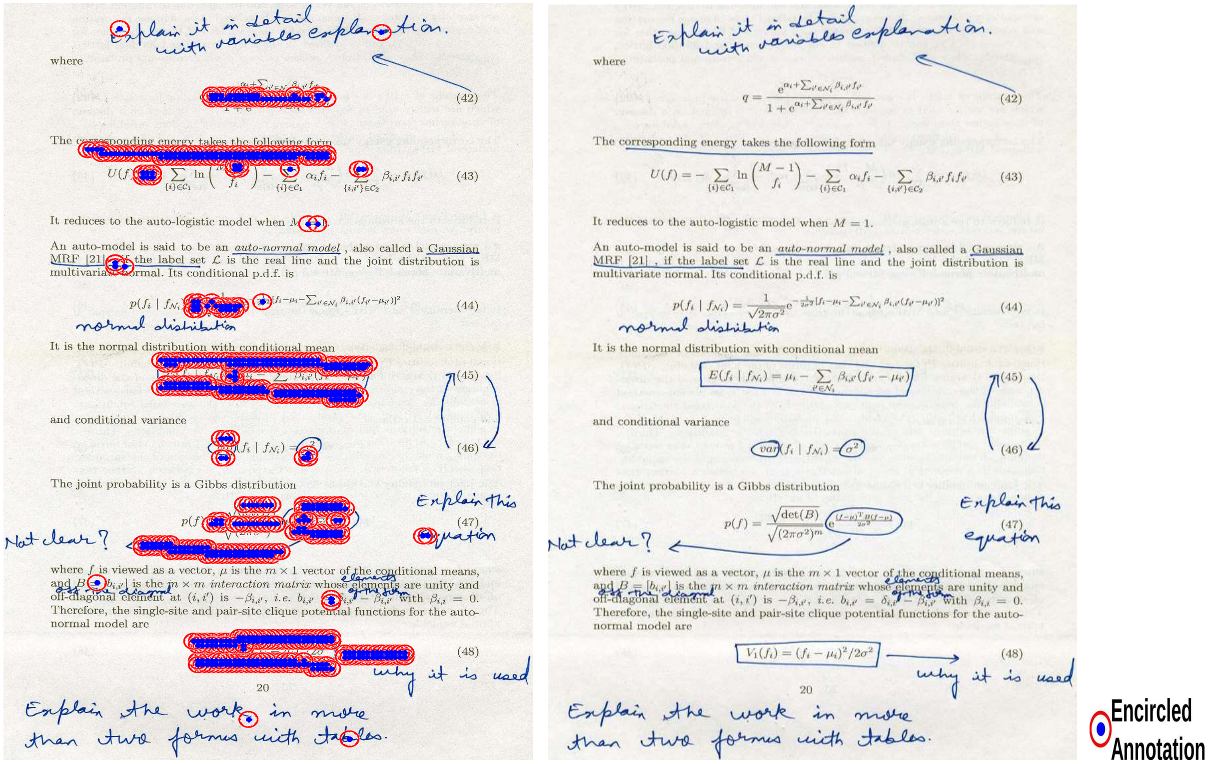
6.5.4 Results on Standard Datasets

It must be noted that there is a non-availability of a multi-annotated dataset. Therefore, likewise as stated in previous Chapter 5 in Section 5.6.4 we apply the DS method on IAM and PRImA-NHM datasets.

Our method shows impressive results on IAM dataset with a recall of .98 for handwritten text and a precision of .99 for the printed text. Similarly, we achieve a recall of .77 for handwritten text and precision of .66 for printed text in PRImA-NHM dataset. Table 6.5 presents the result for both the dataset using weakly supervised visual saliency. Figures 6.12 and 6.13 pictorially presents the results of DS learning for IAM dataset and PRImA-NHM dataset.

6.6 COMPARISON WITH SVM

Table 6.6 and 6.7 along with Figure 5.17 presents the effect of applying SVM on image patches. We applied RBF kernel and used two-class SVM. From the results it is clear that SVM is unable to capture the difference among the closed overlapping feature space among the printed



(a) Highlighted Encircled Regions as Salient Objects

(b) Original Image

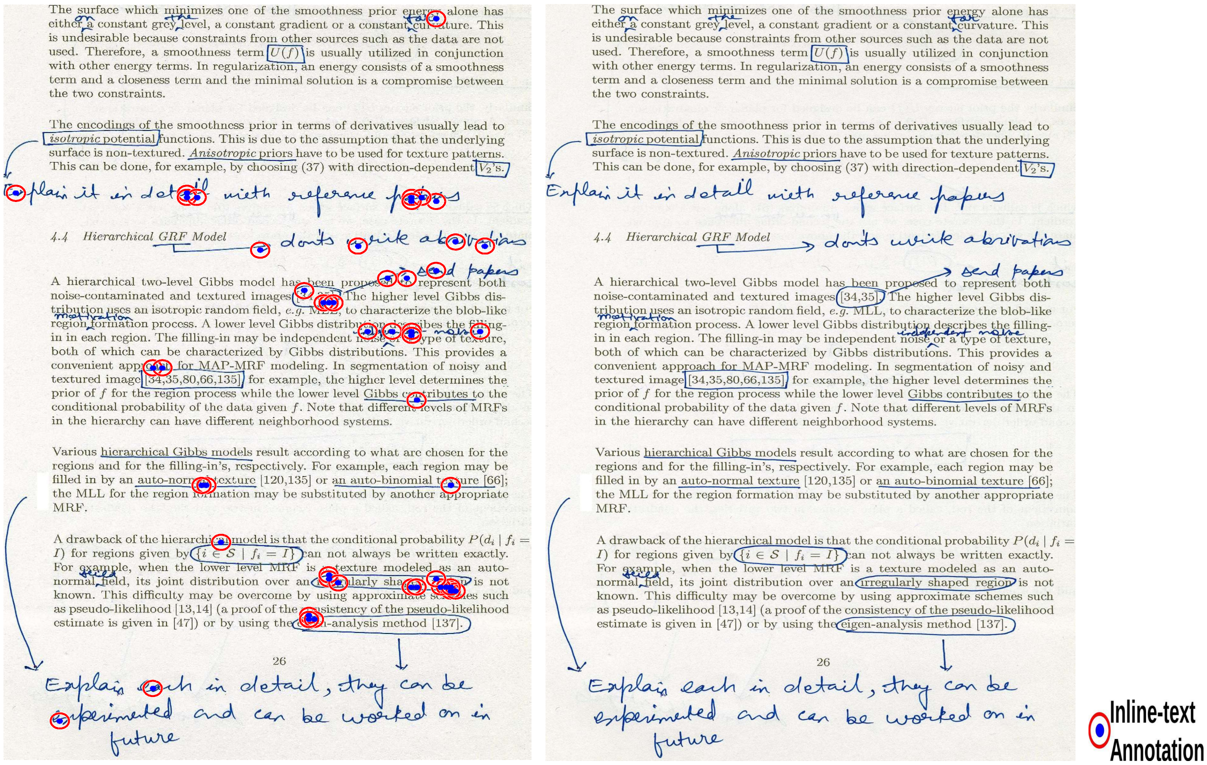
Figure 6.4 : Set 2: Encircled Region Localization in Multi-annotated Images using DS

text and annotations. The most probable reason could be the presence of variety of annotations in a document rather than only text.

6.7 CONCLUSION

Discriminant saliency has been demonstrated to identify specific annotations in a multi-oriented cluttered document. Our experimental results corroborate that discriminant saliency produces better results in comparison to a discriminative classifier such as SVM. It shows comparable results with the CRF based supervised saliency model. It is observed that the overall recall produced for all the experiments is high for supervised saliency model mentioned in previous Chapter 5. Our weakly supervised learned model for annotation extraction performs well for densely annotated documents. While dealing with unconstrained handwriting environment for annotations, in the subsequent chapter 7 we propose a method to detect baseline for unconstrained handwritten word. This allows to separate the core zone from the ascenders and descenders and therefore leads to effective extraction of features for preprocessing and writer identification.

...



(a) Highlighted Inline Annotation Regions as Salient Objects

(b) Original Image

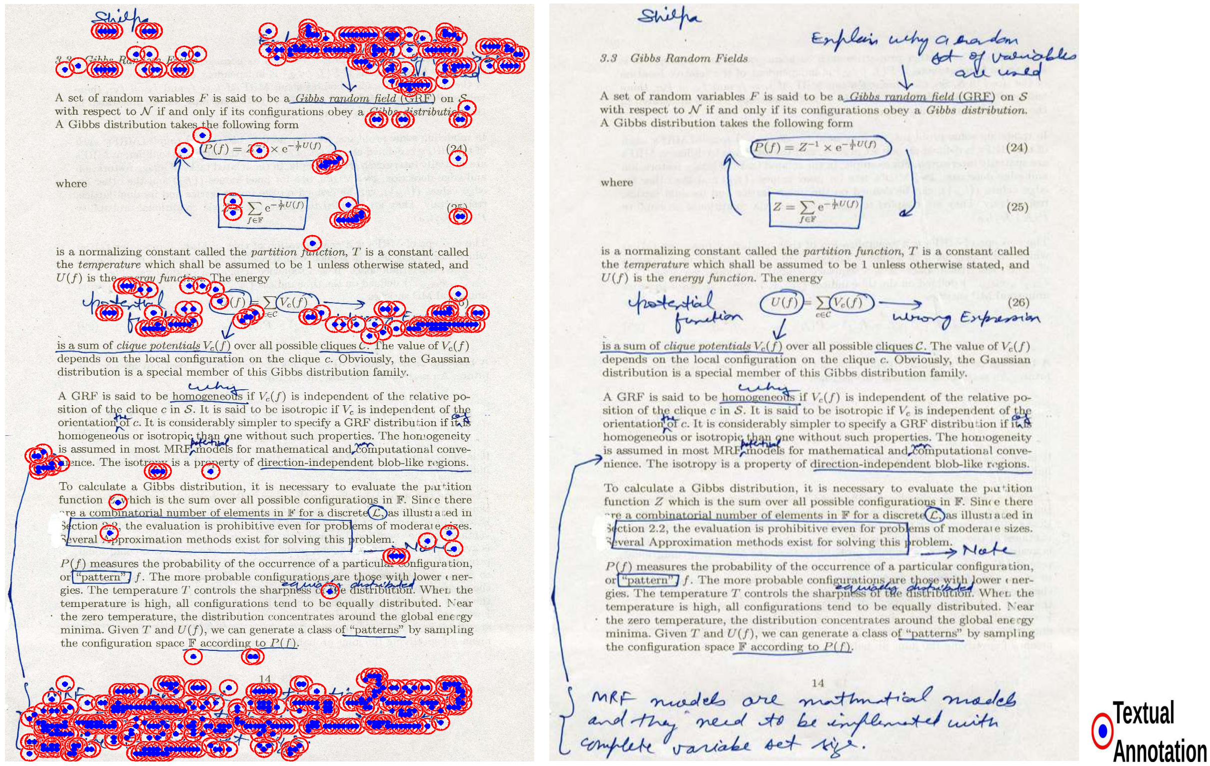
Figure 6.5 : Set 2: Inline Annotation Region Localization in multi-annotated Images using DS

Table 6.3 : Set 3: Annotation localization when the dictionary is trained on images comprising only textual annotations and the testing is performed on similar images, and vice versa.

Annotation Category	Threshold Scalar	Accuracy	Recall	Precision	F1 Score	Execution Time (sec)
Textual (Fig. 6.6)	4	88.30%	.57	.78	.66	20.25
Symbolic (Fig. 6.7)	5	92.89%	.71	.83	.77	17.54

Table 6.4 : Set 4 : Annotation localization when the dictionary is trained on individual annotations and the testing is performed on the images containing individual annotations.

Annotation Category	Threshold Scalar	Accuracy	Recall	Precision	F1 Score}	Execution Time (sec)
Underline (Fig. 6.8)	2.5	92.39%	.68	.93	.79	18.02
Marginal Text (Fig. 6.9)	5	94.03%	.91	.87	.89	15.61
Encircled (Fig. 6.10)	6	92.12%	.87	.70	.77	20.56
Inline (Fig. 6.11)	6	80.13%	.49	.52	.51	8.61



(a) Highlighted Textual Regions as Salient Objects (b) Original Image

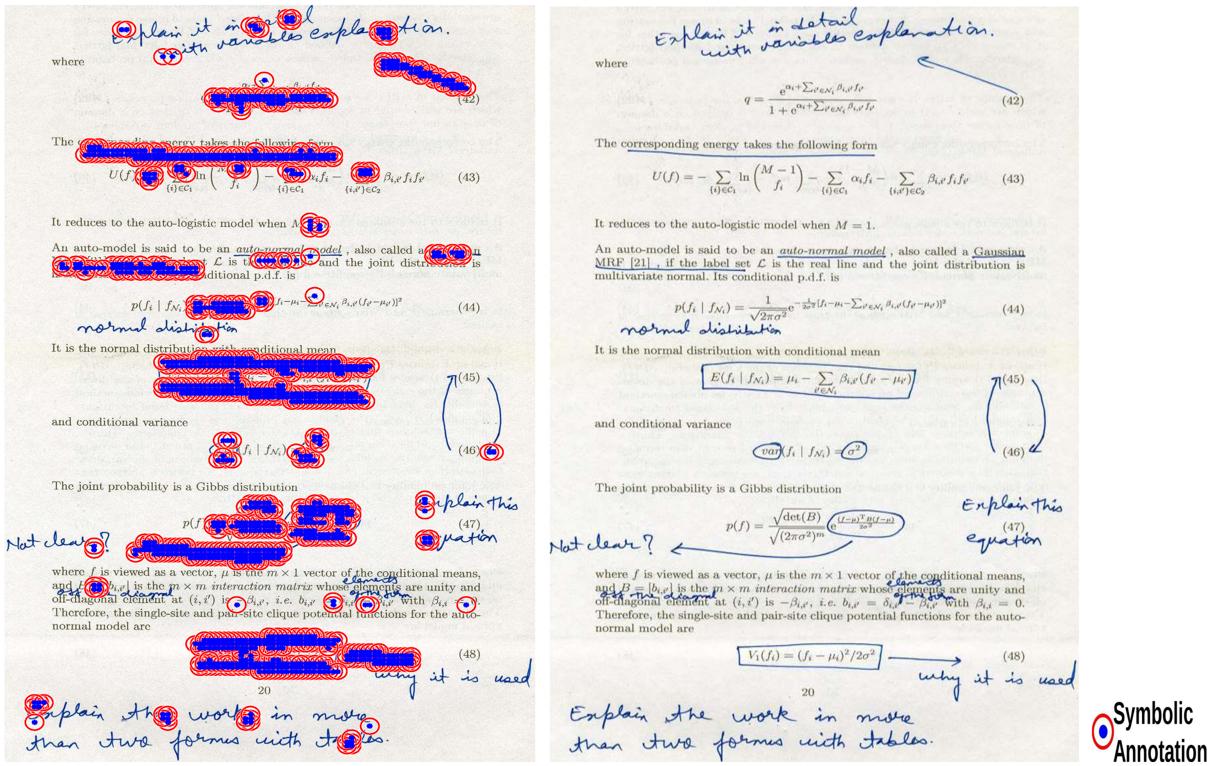
Figure 6.6 : Set 3: Textual Region Localization in Documents comprising both Textual and Symbolic Annotations using DS

Table 6.5 : Annotation localization on IAM and PRImA dataset by DS textual annotation detection model

Dataset	Testset	Accuracy	Precision	Recall	F1-score	Execution Time (sec)
IAM Dataset	100 images	98.73%	.99	.98	.98	.20
PRImA NHM Dataset	100 images	75.95%	.66	.77	.71	.23

Table 6.6 : Comparison of SVM with DS on multi-annotated documents.

Annotation Category	All Annotations		Underline		Marginal Text		Encircled		Inline	
	DS	SVM	DS	SVM	DS	SVM	DS	SVM	DS	SVM
Accuracy (%)	80.23	44.09	94.02	89.09	95.04	55.89	95.81	87.17	82.03	49.19
Precision	.82	.50	.47	.23	.70	.21	.63	.46	.16	.42
Recall	.58	.41	.81	.36	.79	.13	.84	.06	.18	.06
F1 Score	.68	.45	.59	.28	.75	.16	.72	.11	.17	.10
Execution Time (sec)	26.86	14.49	17.05	8.17	8.87	.41	17.11	17.15	8.61	18.49



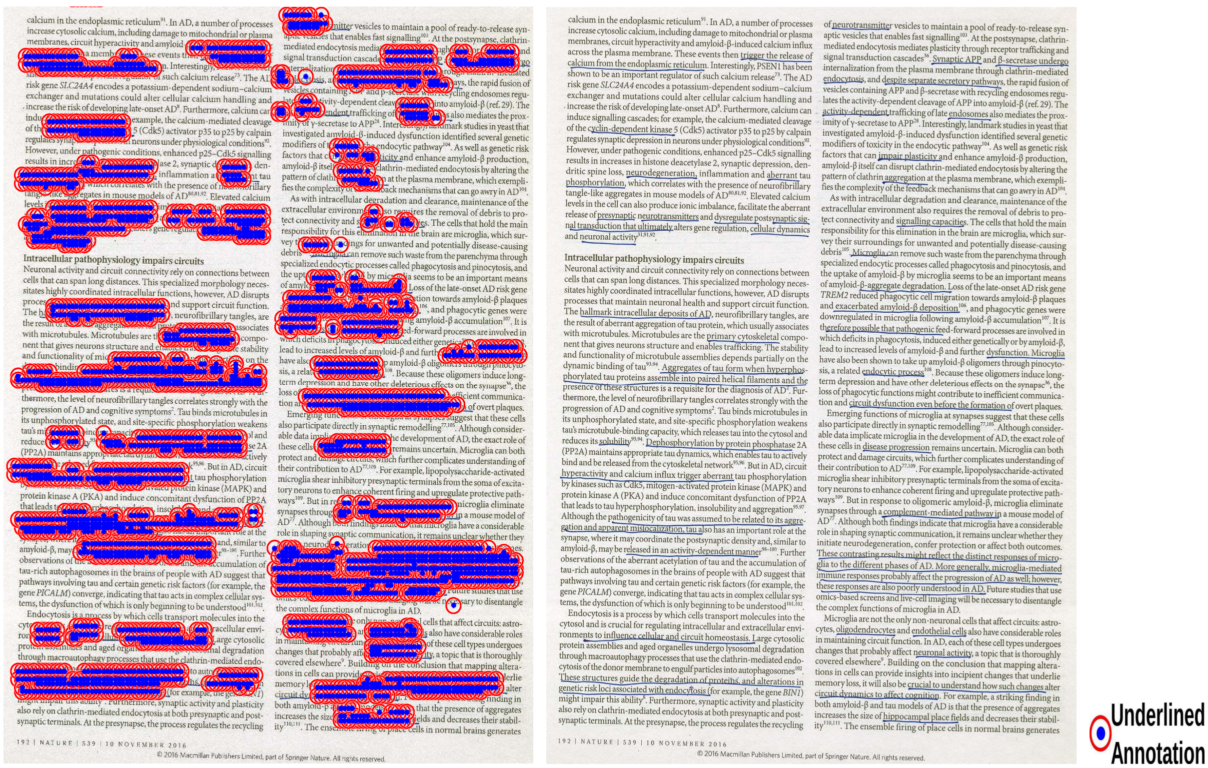
(a) Highlighted Symbolic Annotated Regions as Salient Objects

(b) Original Image

Figure 6.7 : Set 3: Symbolic Region Localization in Documents comprising both Textual and Symbolic Annotations using DS

Table 6.7 : Comparison of SVM with DS on single-class annotated documents.

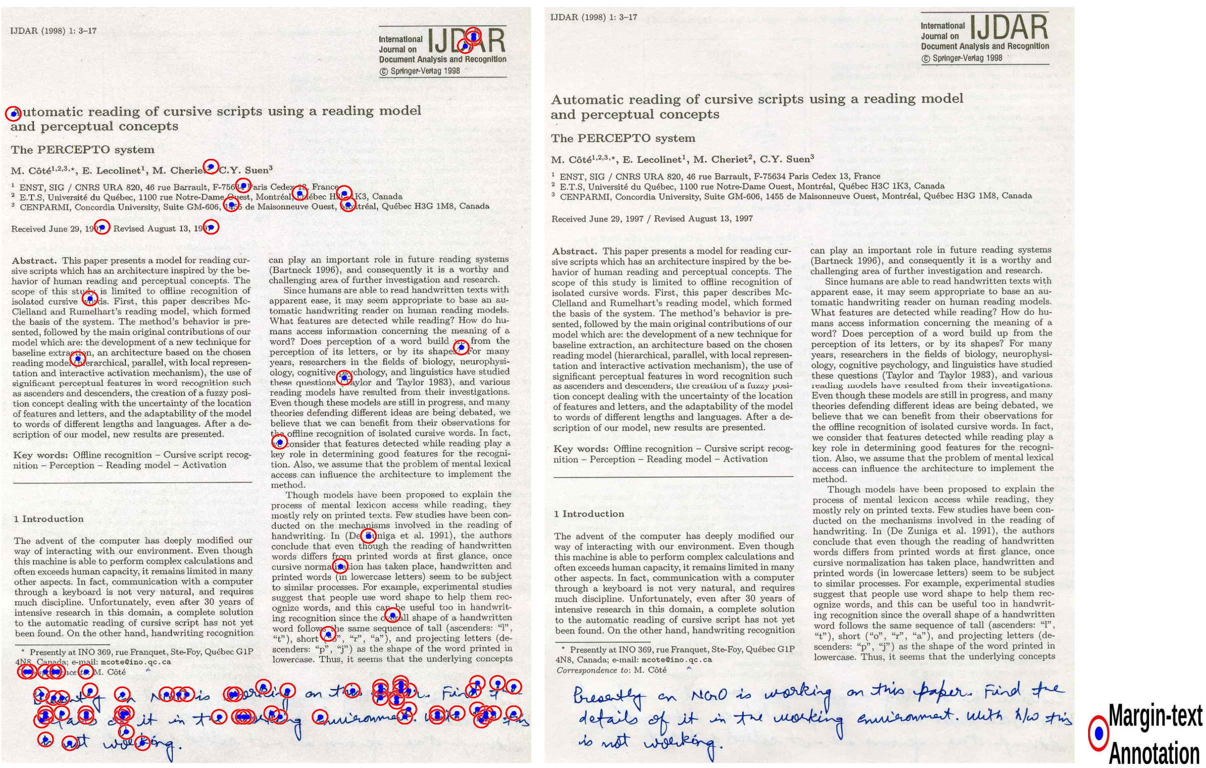
Annotation Category	Underline		Marginal Text		Encircled		Inline	
	DS	SVM	DS	SVM	DS	SVM	DS	SVM
Accuracy (%)	92.39	84.97	94.03	80.59	92.12	84.19	80.13	74.09
Precision	.94	.91	.87	.96	.70	.92	.52	.82
Recall	.68	.30	.91	.04	.87	.01	.49	.01
F1 Score	.79	.45	.89	.08	.77	.01	.51	.02
Execution Time (sec)	18.02	18.35	15.61	24	20.56	21.49	8.61	19.23



(a) Set 3: Highlighted Underlined Regions as Salient Objects

(b) Original Image

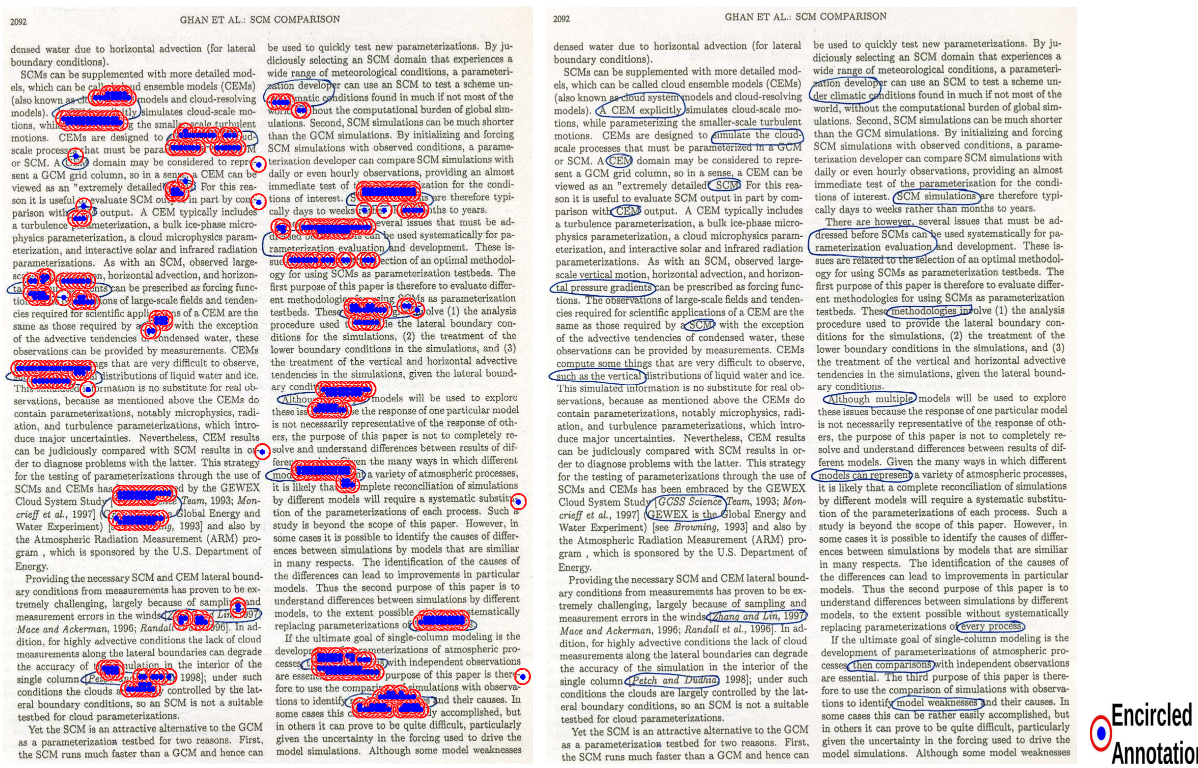
Figure 6.8 : Set 4: Underlined region localization in single-class annotated images using DS



(a) Highlighted Marginal Text Regions as Salient Objects

(b) Original Image

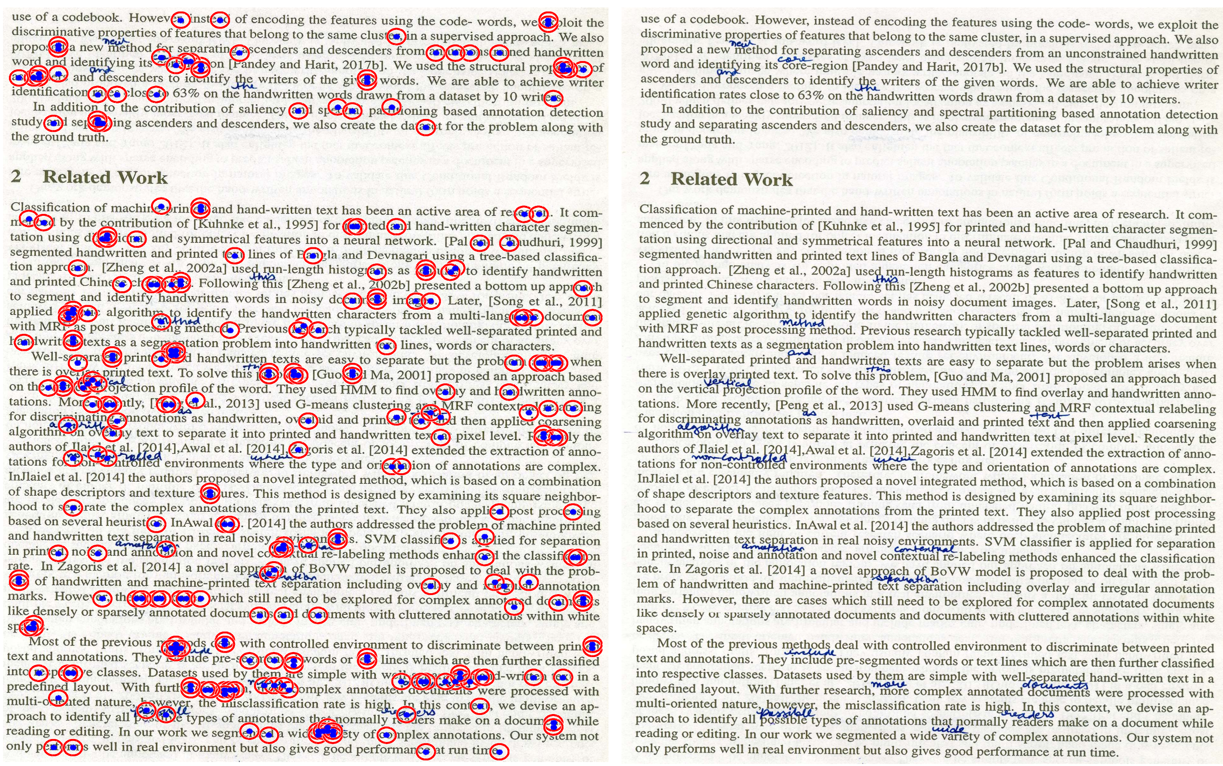
Figure 6.9 : Set 4: Marginal annotations localization in single-class annotated images using DS



(a) Highlighted Encircled Regions as Salient Objects

(b) Original Image

Figure 6.10 : Set 4: Encircled annotation localization in single-class annotated images using DS



use of a codebook. However, instead of encoding the features using the code- words, we exploit the discriminative properties of features that belong to the same cluster, in a supervised approach. We also proposed a new method for separating ascenders and descenders from an unconstrained handwritten word and identifying its core-region [Pandey and Harit, 2017b]. We used the structural properties of ascenders and descenders to identify the writers of the given words. We are able to achieve writer identification rates close to 63% on the handwritten words drawn from a dataset by 10 writers.

In addition to the contribution of saliency and spectral partitioning based annotation detection study and separating ascenders and descenders, we also create the dataset for the problem along with the ground truth.

2 Related Work

Classification of machine-printed and hand-written text has been an active area of research. It commenced by the contribution of [Kuhnke et al., 1995] for printed and hand-written character segmentation using directional and symmetrical features into a neural network. [Pal and Chaudhuri, 1999] segmented handwritten and printed text lines of Bangla and Devnagari using a tree-based classification approach. [Zheng et al., 2002a] used run-length histograms as features to identify handwritten and printed Chinese characters. Following this [Zheng et al., 2002b] presented a bottom up approach to segment and identify handwritten words in noisy document images. Later, [Song et al., 2011] applied genetic algorithm to identify the handwritten characters from a multi-language document with MRF as post processing method. Previous research typically tackled well-separated printed and handwritten texts as a segmentation problem into handwritten text lines, words or characters.

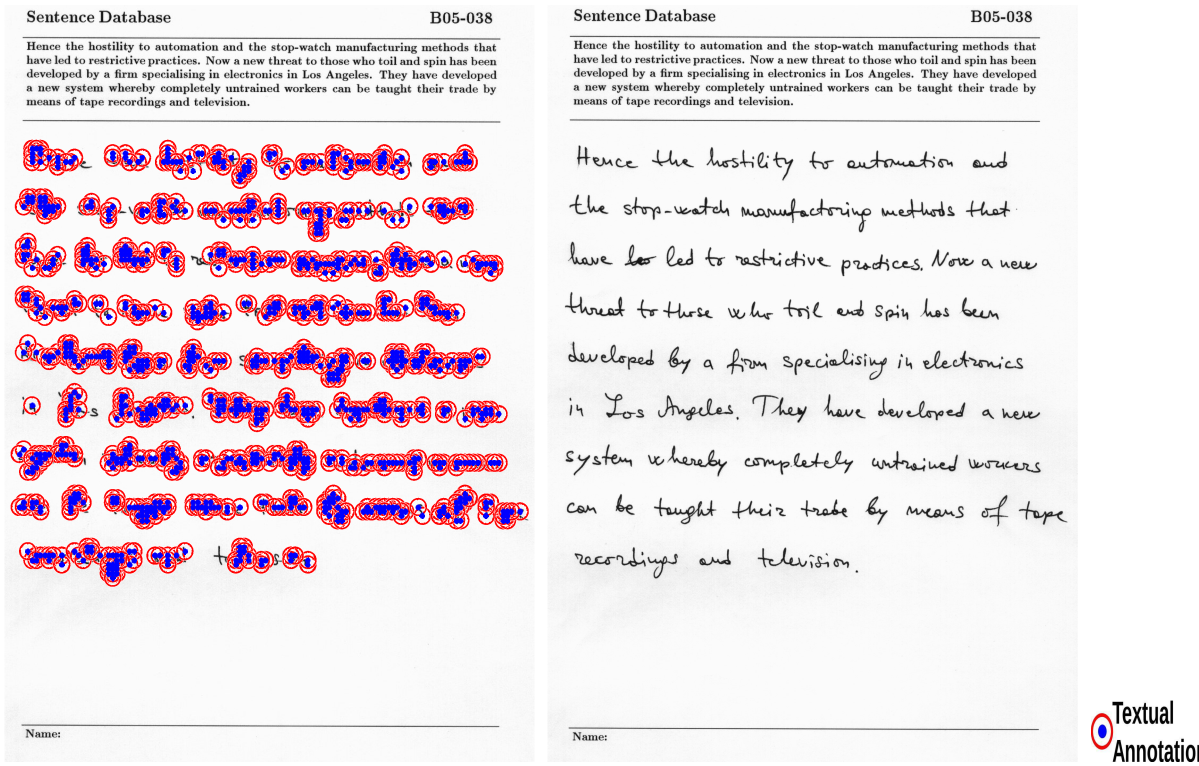
Well-separated printed and handwritten texts are easy to separate but the problem arises when there is overlay printed text. To solve this problem, [Guo and Ma, 2001] proposed an approach based on the vertical projection profile of the word. They used HMM to find overlay and handwritten annotations. More recently, [Peng et al., 2013] used G-means clustering and MRF contextual relabeling for discriminating annotations as handwritten, overlaid and printed text and then applied coarsening algorithm on overlay text to separate it into printed and handwritten text at pixel level. Recently the authors of Jlaiel et al. [2014],Awal et al. [2014],Zagoris et al. [2014] extended the extraction of annotations for non-controlled environments where the type and orientation of annotations are complex. InJlaiel et al. [2014] the authors proposed a novel integrated method, which is based on a combination of shape descriptors and texture features. This method is designed by examining its square neighborhood to separate the complex annotations from the printed text. They also applied post processing based on several heuristics. InAwal et al. [2014] the authors addressed the problem of machine printed and handwritten text separation in real noisy environments. SVM classifier is applied for separation in printed, noise and annotation and novel contextual re-labeling methods enhanced the classification rate. In Zagoris et al. [2014] a novel approach of BoVW model is proposed to deal with the problem of handwritten and machine-printed text separation including overlay and irregular annotation marks. However, there are cases which still need to be explored for complex annotated documents like densely or sparsely annotated documents and documents with cluttered annotations within white spaces.

Most of the previous methods deal with controlled environment to discriminate between printed text and annotations. They include pre-segmented words or text lines which are then further classified into respective classes. Datasets used by them are simple with well-separated hand-written text in a predefined layout. With further research, more complex annotated documents were processed with multi-oriented nature, however, the misclassification rate is high. In this context, we devise an approach to identify all possible types of annotations that normally readers make on a document while reading or editing. In our work we segmented a wide variety of complex annotations. Our system not only performs well in real environment but also gives good performance at run time.

(a) Highlighted Inline Annotated Regions as Salient Objects

(b) Original Image

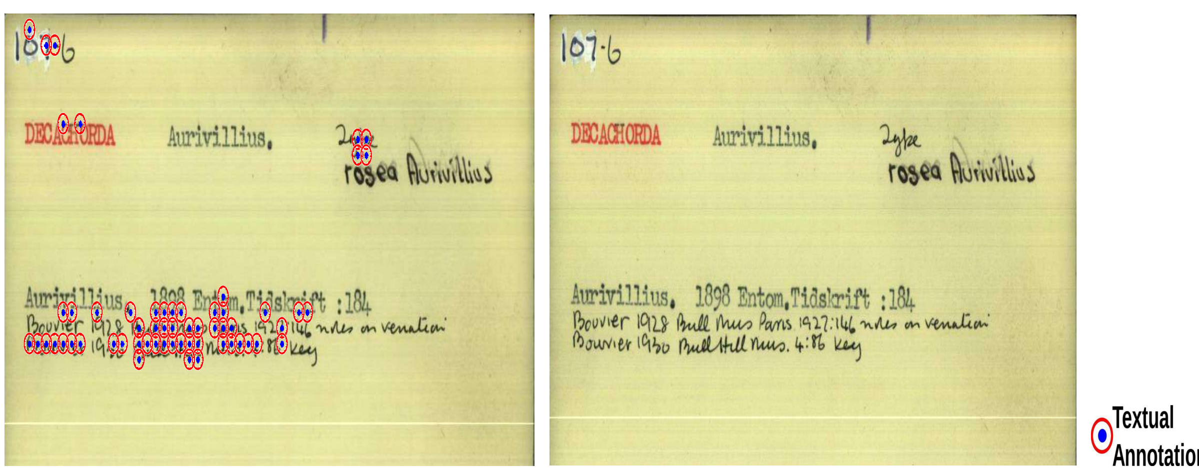
Figure 6.11 : Set 4: Inline annotation localization in single-class annotated images using DS



(a) Highlighted Textual Annotated Regions as Salient Objects

(b) Original Image

Figure 6.12 : Textual region localization in IAM images using DS.



(a) Highlighted Textual Annotated Regions as Salient Objects

(b) Original Image

Figure 6.13 : Textual region localization in PRImA images using DS.