LEARNING TOP DOWN SCENE CONTEXT FOR VISUAL ATTENTION MODELING IN NATURAL IMAGES

S. Karthikeyan, Vignesh Jagadeesh and B.S. Manjunath

Department of Electrical and Computer Engineering, University of California Santa Barbara
{karthikeyan, vignesh, manj }@ece.ucsb.edu

ABSTRACT

Top down image semantics play a major role in predicting where people look in images. Present state-of-the-art approaches to model human visual attention incorporate high level object detections signifying top down image semantics in a separate channel along with other bottom up saliency channels. However, multiple objects in a scene are competing to attract our attention and this interaction is ignored in current models. To overcome this limitation, we propose a novel object context based visual attention model which incorporates the co-occurrence of multiple objects in a scene for visual attention modeling. The proposed regression based algorithm uses several high level object detectors for faces, people, cars, text and understands how their joint presence affects visual attention. Experimental results on the MIT eye tracking dataset demonstrates that the proposed method outperforms other state-of-the-art visual attention models.

Index Terms— Scene Context, Visual attention modeling, Eye Tracking.

1. INTRODUCTION

Humans are able to swiftly process a rich stream of visual data and extract informative regions suitable for high level cognitive tasks. Therefore, understanding the manner in which humans process visual stimuli in a free viewing scenario has been an interesting problem in the scientific and engineering community. Several applications in computer vision (object recognition [1], visual tracking [2], text detection [3]), graphics (non-photo realistic rendering [4]), multimedia (video summarization [5], video compression [6]) and robotics (robot localization [7]) can benefit from better understanding of human visual attention. A detailed overview of various saliency algorithms and its applications are presented in [8].

Early visual attention models [9, 10] are pure bottom up approaches and use multiple low level image features such as intensity, color, orientation, texture and motion to determine regions of interest in natural images. In these approaches feature specific saliency maps are computed for every low level feature and the final master map is a linear or non linear combination of individual feature specific saliency maps. However, in meaningful scenes, top down factors such as task at hand and image semantics play a major role in capturing attention.

Recent research [11] suggests that when subjects view natural scenes, faces and text primarily attract attention. A mathematical model using this information to improve human attention prediction was proposed in [4] which utilizes multiple object detectors (car, person and face) and low level saliency maps. A linear SVM is trained on these features to predict human attention regions in an image. This approach essentially learns a single weight for each feature vector. In practice, a single weight for each object irrespective of the scene content can be a severely limiting assumption. Consider an example shown in Fig.1, the car is highly salient in the left image, but not in the right image due to the presence of the large face (other salient objects in general) in the scene. The weight vector learnt by [4] for car will not reflect the true saliency of the car as it ignores the presence of other object(s) in the scene. Hence, modeling context and competition between multiple interacting objects in a scene is essential to build systems which are better at predicting where humans look in images. In the absence of meaningful semantic objects in the scene, low level features primarily drive visual attention. However, presence of interesting semantic objects initiates change in visual attention from low-level to high level context and current visual attention models do not explicitly model this transition.

Fig. 1. In the left figure, the car is the only important object in the image and it captures most of the visual attention (overlay in red). In the right figure, the presence of the person's face diverts attention from the car. Therefore, the context of the scene plays a major role in deciding the objects of interest. Should be viewed in color.

Also, recent research using controlled experiments [12, 13] highlight the importance of object co-occurrence and context for visual search tasks. These works indicate that other objects in a scene can provide a distracting(sometimes positive) effect for visual search of a specific object using reaction time studies. In a similar perspective, our effort aims to model the effect of object co-occurrence for a free viewing task and helps in creating a better organization of interesting regions in a scene.

In addition, previous learning based approach [4] artificially generates a classification problem by thresholding the attention map to estimate visual saliency. However, as the attention maps are continuous, it naturally presents itself as a regression problem. To summarize, the primary contribution of our work is to create a novel framework which predicts visual attention by modeling object co-occurrence in a scene using a regression approach. A comparison of our algorithm with existing state of the art visual attention algorithms in the MIT eye tracking dataset yields encouraging results.

2. THE PROPOSED APPROACH

Our saliency algorithm learns a regression model from features extracted at multiple levels in an image. Apart from low, mid and high

level features used in [4], the proposed technique also uses scene context features which model the interaction between these features. The following sections explain the proposed feature extraction and learning steps.

2.1. Feature extraction

We incorporate features at four levels - low, mid, high and scene context to train our classifier.

2.1.1. *Low level features*

Our model utilizes the following low level features due to their importance in bottom up saliency.
Itti and Koch saliency: Early saliency model [9] motivated by linear filtering and center surround operation provides intensity, color and orientation channels which are suitable for bottom up visual attention modeling.
Steerable pyramid filters: Provides filter responses which correlate well with visual attention and therefore local energy of steerable pyramid filters [14] in four orientations and three scales are used.
Torralba Saliency: Provides a holistic representation of a scene [15] using spectral and coarsely localized information.
Color histogram features: The values of the red, green and blue channels and the probabilities of each of these channels are used according to [4].
Signature Saliency: Provides a saliency map [16] using the theoretical framework of sparse signal mixing which spatially approximates image foreground.
Graph Based Visual Saliency: Jointly models feature extraction and activation map creation in a unified manner by defining edge weights using saliency and dissimilarity [17].

2.1.2. *Mid level features*

Horizon detection is performed using using gist descriptor [15]. It is especially important in outdoor scenes where salient objects are present near the ground plane .

2.1.3. *High level features*

High level objects such as faces and text have high visually saliency. We utilize automatic object detectors for face [18], person [19], car [19] and text [20] in our model. The source code for [20] is not publicly available and we used our implementation in this paper.

2.1.4. *Scene Context features*

In addition to high level features, we propose a novel set of features which model the pairwise interaction between multiple high level features. High level features typically model attention gain in the locality of semantic objects. However, presence of interesting objects in a scene also incurs attention loss in other objects (in general other high, mid and low level features) in a scene. This attention loss scene context features can be described using a *cause-effect* mechanism. Let there be N possible objects in a scene and $f^{SC}_{i,j}(x,y)$ denote the scene context feature between object i and j at position(x,y). The scene context vector models the attention loss in the scene incurred on the pixels of object i (*effect*) due to presence of object j (*cause*). Now let the total number of objects corresponding to label i in the scene be denoted by N_i and the number of object $i's$ in position (x,y) be $n_i(x,y)$. Let the image be denoted as $\mathcal{I}$, the scene context vectors are defined as

$$f^{SC}_{ij}(x,y) = \begin{cases} N_j - n_j(x,y) & \text{if } n_i(x,y) > 0 \\ 0 & \text{if } n_i(x,y) = 0 \end{cases} \quad (1)$$

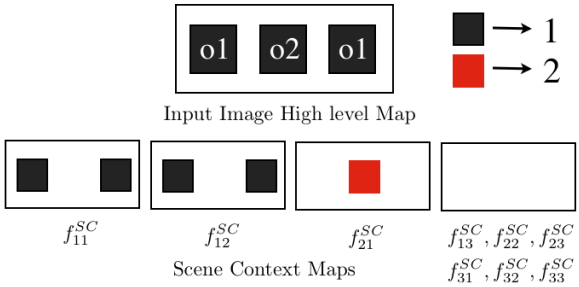
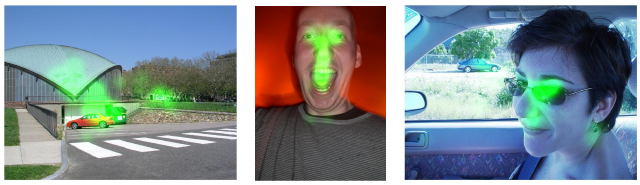


Fig. 2. An example high level object layout with N=3 and the corresponding 9 scene context feature maps

Fig.2 presents an example of scene context features where $N = 3$. Object 3 is not present in the image and therefore $f^{SC}_{ij} = 0$ if $i = 3$ or $j = 3$. Next, f^{SC}_{11} informs the regions where object 1 is present about another object 1 detection. Further, f^{SC}_{21} indicates the pixels of object 2 the presence of two object 1 detections.

Typically, the number of distinct objects which can occur in an image is large and modeling the cause-effect relationship between every pair of objects is necessary. However, it is not possible to obtain such large number of pairwise interactions in practical eye tracking datasets. Therefore, to reduce the dimensionality of the scene context features, we propose an approximation technique which clusters the cause and effect features separately which have similar properties.



(a) Car (b) Face (c) Car and Face

Fig. 3. The left figure has only one salient object (small car) and the center image has one salient object (large). The right image has both the salient objects. We observe the large face significantly draws attention (green overlay) away from the small car in the left image. Best viewed in color.

Cause effect clustering: Our aim is to model the factors affecting cause and effect of per pixel attention loss (attention density loss) which will be predicted by the learning algorithm. In order to gain insight to factors affecting cause and effect of attention loss, consider an example in Fig.3 which shows individual images of a small car (left), large face (center) and an image where they co-occur (right) with the corresponding attention map overlays in green. The attention sum of the two large faces in (b) and (c) are 0.7 and 0.6 and densities (normalized) are 6 and 5 respectively. The attention sum of the two cars in (a) and (c) are 0.1 and 0.05 and densities (normalized) are 20 and 8 respectively. The attention density in the large face remains relatively unchanged due to the presence of the small car, but the presence of the large face significantly contributes to attention loss in the small car as its density is almost halved. This illustrates the requirement to cluster the attention loss cause and effect features

with different metrics (presented in Algorithm 1). Objects which cause a large attention shift are typically objects which have a large overall attention sum contained within them(typically objects which occupy a large portion of viewing area, eg. large face in Fig.3). Also, the objects which take the greatest impact or effect have high concentration or attention density (small prominent objects, eg. small car in Fig.3).

Data: $\{\mathcal{I}_i\}$ - Input Training Images. $\{\mathcal{A}_i\}$ - Attention maps.
$\{\mathcal{H}_i^{jk}\}$ - Binary high level maps denoting the k^th occurrence of high level object index j in image index i
$N-$number of training images, $M-$number of high-level objects
$i \in [1, N]$ and $j \in [1, M]$
Result: $\mathcal{C}$ - Cause Clusters, $\mathcal{E}$ - Effect Clusters.
Initialization: $n^j = 0 \forall j \in [1, M]$ (counter)
$dA = \emptyset$ (Attention density vectors) $sA = \emptyset$ (Attention sum vectors);
for *i=1*$\rightarrow N$ **do**
  for *j=1*$\rightarrow N$ **do**
    for *k=1*$\rightarrow |\mathcal{H}_i^{j*}|$ **do**
      $n^j = n^j + 1;$
      $sA(n^j) = \sum_{p \in \mathcal{P}} \mathcal{A}_i(p) \bigodot \mathcal{H}_i^{jk}(p);$
      $dA(n^j) = \frac{\sum_{p \in \mathcal{P}} \mathcal{A}_i(p) \bigodot \mathcal{H}_i^{jk}(p)}{\sum_{p \in \mathcal{P}} \mathcal{H}_i^{jk}(p)};$
    end
  end
end
$\mathcal{C}$ = kmeans(sA); $\mathcal{E}$ = kmeans(dA);

Algorithm 1: Training algorithm

2.1.5. Center Prior

Finally, as images typically contain the object of interest in the center, we utilize a center prior map [4].

2.2. Learning

The features are pre-computed in all the images and a regression model is used to predict where subjects look in new images. For this purpose, the dataset is divided into training and test sets in a 10-fold cross validation setting. From each image in the training dataset, we randomly pick equal number of pixels from the top 20% and bottom 80% attention regions (to have adequate representation for high attention regions) to create a pixel level training subset. A regression model is learnt from this subset and pixel wise attention density in new images are predicted using this regression model.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

The experiments are conducted in the large publicly available MIT eye tracking dataset [4]. This dataset consists of images collected from Flickr and LabelMe datasets.The dataset consists of eye tracking data on 1003 images collected from 15 different subjects. From the eye tracking fixation maps, the attention map is obtained by smoothing the fixation maps using a gaussian filter. In the current experimental setup, the images were resized such that the smallest dimension has 100 pixels while maintaining the original aspect ratio.

3.2. Setup

For every image, low level saliency features and high level object detections are computed as described in Sec. 2.1. Object size plays a significant role in determining attention density per pixel. Therefore, objects are distinguished as small or large by thresholding their size (2500 pixels in our attention maps). Hence, the number of distinct objects double from 4 to 8.

This setup requires modeling 64 (8$\times$ 8) pairwise interactions and as explained in Sec. 2.1.4, we resort to k-means clustering of cause and effect attention loss features using attention sum and attention density respectively according to Algorithm 1. Fig.4 indicates that the elbow for clustering error in both attention loss cause and effect clusters is obtained for two clusters each. The object groupings in the cause cluster are { large face, large car, large person and large text } and {small face, small car, small person and small text}. This intuitively makes sense as large objects typically have a large attention sum and vice versa. The two attention loss effect clusters are {large text, small text, large face, small face} and {large person, small person, large car, small car }. This signifies that attention densities of faces and text are considerably higher than other objects irrespective of size.

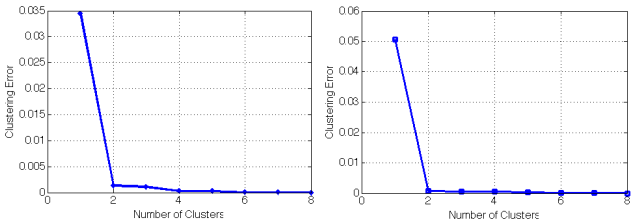


Fig. 4. Error plots of clustering the average density (left) and average sum (right) of visual attention over all 8 object classes. The elbow occurs at two clusters for both the plots but for different cluster sets.

Our framework also models the manner in which high level objects initiate attention transition from low level cues to higher order context. As the center prior, signature and GBVS maps have the highest predictive power among low level features to understand where subjects look (in Fig.5) they are added as 3 additional entities to the attention loss effect clusters. This amounts to 2 attention loss cause features and 5 effect features, providing 10 pairwise cause-effect features. To sum up, *we have 30 low-level features, 1 mid-level feature, center prior, 8 high level features and 10 scene context features totaling 50 features.*

3.3. Performance

Performance of saliency algorithms is presented using ROC curves, which are computed as follows. All visual attention algorithms generate a saliency map with a predicted pixel-level saliency. This saliency map is thresholded at k=1,3,5,10,15,20,25 and 30 percent to obtain binary saliency maps. The percentage of human fixations contained within each binary map is the performance measure.

Firstly, Table 1 indicates that modeling the attention prediction problem using a regression approach provides us some gain in performance as it avoids quantizing the training attention maps. Here, we compare the classification algorithm (liblinear SVM) used in [4] to three regression algorithms (linear [21], random forest [22] and PLS-regression [23]). We notice that PLS-regression using 45 basis vectors consistently achieves the best performance.

A comparison of performance of the proposed approach using PLS regression(with and no context features) with Judd et al., GBVS, Itti and Koch saliency, Torralba saliency, Signature saliency, Center prior (DistCenter), other low (color and subband), mid (hori-

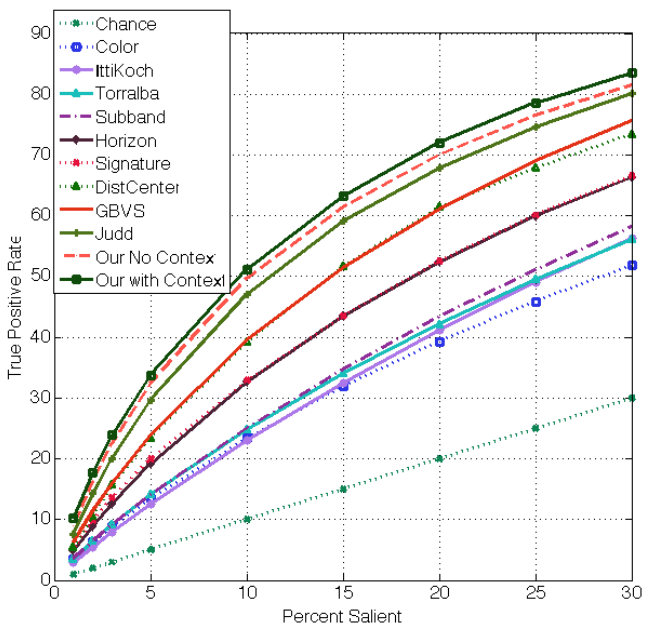


Fig. 5. Comparison of the performance of our algorithm to other low level saliency and learning based algorithms using ROC curves. Best viewed in color.

	1	5	10	20	30
Lib. Linear SVM	9.51	32.73	50.02	70.92	82.62
Linear Reg.	9.61	33.04	50.31	71.16	82.91
Random Forest Reg.	9.56	32.91	50.51	71.25	82.99
PLS Reg.	**9.76**	**33.37**	**50.77**	**71.60**	**83.19**

Table 1. Comparison of various regression algorithms (linear, Random Forest and PLS) to lib-linear SVM classification used in [4] when the output saliency map is thresholded at multiple levels. Regression based algorithms consistently outperform lib-linear SVM. Bold indicates best performance.

zon) level features and chance is shown in Fig.5. Our algorithm outperforms other saliency algorithms by a significant margin. Judd et al. [4] also utilizes high level objects (car, person and face) but does not model scene context and learns only a single weight for every object irrespective of the size of the object and scene statistics. Hence, on an average the proposed approach obtains 3.5% gain over [4] and the split is about 1.1% due to utilization of better low level features, 0.5% due to gains of regression over classification model and 1.9% gain (difference between our approach results with and no context, in Fig.5) can be attributed to modeling scene context using object co-occurrence.

Our framework models scene context as features which capture attention loss. Therefore, a linear regression algorithm automatically needs to learn negative weights for these feature vectors. Fig.7 shows that weights for all co-occurrence features are automatically learnt negative without enforcing them in the training stage which indicates our scene context design works in an expected manner. Fig.6 compares some saliency map outputs of our algorithm to [4].

4. CONCLUSION

Scene context plays a crucial role in determining where people look in images. This paper is a pioneering effort to understand the role of scene context in task free viewing scenario. Apart from low, mid and high level features we propose scene context features which communicates to each object (and few other features) the presence

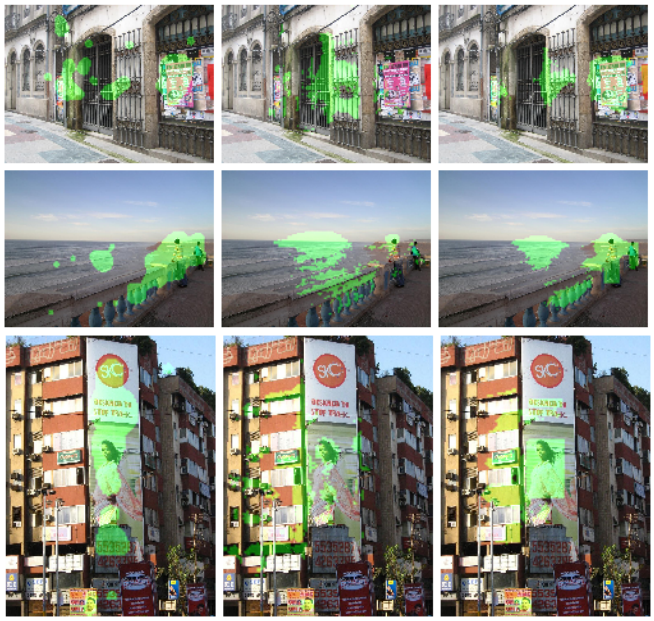


Fig. 6. The left column shows the ground truth fixation maps, the center column indicates the saliency map of [4] and the right column is the saliency map of the proposed approach (top 15% pixels overlay in green). In the first row, we outperform [4] as we include text detections in our learning framework. In the second row, the center prior primarily biases [4] and our attention map is less prone to such bias as the center pixels are aware of the presence of other people in the image and therefore our attention map is more representative of the ground truth. In the third row also our approach learns better attention maps closer to the ground truth than [4] due to scene context modeling. Best viewed in color.

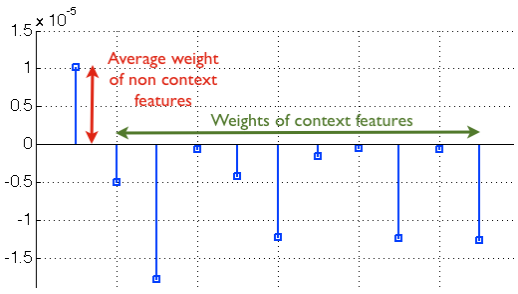


Fig. 7. Weights of the 10 context features compared to the average of non-context features. Non-context features typically have an additive effect and as scene context models attention loss, its weights are learnt negative.

of other objects in a scene. This results in loss of attention in the object of interest and is automatically learnt as negative weights in a linear regression setting. Our work also compares classification to linear and non-linear regression techniques for learning attention maps and outperforms state of the art in visual attention modeling.

Acknowledgements This work was supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Our other funding sources are N00014-12-1-0503, NSF III-0808772 and OIA-0941717.

5. REFERENCES

[1] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," *Advances in neural information processing systems*, vol. 17, no. 481-488, pp. 1, 2004.

[2] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1007–1013.

[3] S. Karthikeyan, V. Jagadeesh, and BS. Manjunath, "Learning bottom-up text attention maps for text detection using stroke width transform," *IEEE International Conference on Image Processing*, 2013.

[4] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th international conference on*. IEEE, 2009, pp. 2106–2113.

[5] Y.F. Ma, L. Lu, H.J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the tenth ACM international conference on Multimedia*. ACM, 2002, pp. 533–542.

[6] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 185–198, 2010.

[7] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *Robotics, IEEE Transactions on*, vol. 25, no. 4, pp. 861–873, 2009.

[8] A. Borji, L. Itti, J. Liu, P. Musialski, P. Wonka, J. Ye, S. Ji, W. Xu, M. Yang, K. Yu, et al., "State-of-the-art in visual attention modeling.," *Transactions on Pattern Analysis and Machine Intelligence*.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," *Vision research*, vol. 39, no. 19, pp. 3157–3163, 1999.

[11] M. Cerf, E.P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of vision*, vol. 9, no. 12, 2009.

[12] C. Hickey and J. Theeuwes, "Context and competition in the capture of visual attention," *Attention, Perception, & Psychophysics*, vol. 73, no. 7, pp. 2053–2064, 2011.

[13] S.C. Mack and M.P. Eckstein, "Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment," *Journal of Vision*, vol. 11, no. 9, 2011.

[14] E.P. Simoncelli and W.T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Image Processing, 1995. Proceedings., International Conference on*. IEEE, 1995, vol. 3, pp. 444–447.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[16] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 194–201, 2012.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, 2007.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2001, vol. 1, pp. I–511.

[19] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[20] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2963–2970.

[21] J. Cohen and P. Cohen, *Applied multiple regression/correlation analysis for the behavioral sciences.*, Lawrence Erlbaum, 1975.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and intelligent laboratory systems*, vol. 58, no. 2, pp. 109–130, 2001.