Error Factor Analysis for Wild Scene Image-Labelling

Peng Wang, Alan Yuille
University of California, Los Angeles
jerryking234@gmail.com yuille@stat.ucla.edu

Abstract

PASCAL VOC Segmentation Challenge [10] is currently considered as one of the datasets that reflect the image segmentation difficulties for real world scenarios [29]. However, current evaluation is simply based on a single Intersection Over Union (IOU) score. In this paper, we try to discover the error factors under the IOU, which makes the results more informative to understand rather than a black box. Specifically, we decompose the error into three error types in terms of object characteristics, i.e. general, appearance and shape. Each error type is composed of respective factors, e.g. size and aspect ratio for general, appearance distinctiveness for appearance, etc. Finally, for each factor and error type, we perform analysis over its impact on and correlation with the final IOU through robust regression. Our experiments show that these error factors have significant relationship with the given IOU accuracy, and the analysis provides practical guidance on further improvement of the given algorithm.

1. Introduction

Image labelling is one of the center goal in computer vision which embeds many individual tasks including segmentation, detection, scene recovery and recognition, etc. This area has been remaining a centre attraction for researchers these years [4, 19, 1, 5, 11, 36, 14]. In addition, there has been a lot of dataset proposed in order to measure the performance of algorithms [10]. Targeting at handling real world labelling problem, PASCAL VOC Segmentation Challenge [10] provided a wild scene labelling bench mark for 20 object classes. To evaluate the performance of different algorithms, it provides the Intersection Over Union (IOU) for an overall measurement. Recently, researchers have made dramatic progress [5, 14] on achieving good results on the IOU criteria. Most of them claim to combine multiple types of information for handling certain variations, and several types of difficulties through robust features [5], context [39] or more flexible models [38]. As the results, a general IOU scores and several illustrative qualitative images are reported. In the picked failure cases, not accurate segment candidates, rare perspective and occlusion of the targeting objects are often blamed. We show a typical measurement on the left of Fig. 1. The IOU might be sufficient to measure an algorithm on some simple datasets. However, in wild scene segments, the complexity and variance of semantic classes are largely increased. Thus, if we want to handle these issues in order to improve the accuracy, it is essential for us to understand the amount of difficulties and how much these difficulties yield failures of the algorithms. Luckily, thanks to a lot of recent available additionally label on this large wild scene dataset including the parts, backgrounds [7, 20] and 3D objects representation [34, 31], we are able to perform a better analysis of the error factors to understand the difficulties.

As indicated by examples in Fig. 2, under the same overall IOU, the numerical range of these error factors varies a lot with respect to semantic classes and algorithms. In other words, two algorithms could have same IOU, but with very different properties in handling various issues. This makes it crucial to evaluate algorithm along various dimension of errors. Thus, through the analysis of the detailed impaction of error factors underlying the IOU score, on one hand, we can have a better interpretation over the properties of the algorithms, on the other hand, we could discover the potential complementary aspects between different approaches to make our system better developed. In addition, by speculating the real world difficulties that affecting the performance, we are able to target at the major issues we need to handle and make suggestions on the most promising direction. For example, we can analyse out whether it lacks of training images for particular perspective of an object, and make suggestions on data collection or generation. Last but not the least, based on the distribution of the error factors from the target dataset, one can pick the most suitable algorithm to perform on this dataset.

1.1. Related works

For image labelling, various approaches have been proposed on exploring particular cues such as object shape constrains [40, 17], scene context [28, 22, 11], joint models [37, 18] to handle object variations and occlusion, etc. However, under an overall IOU measurements, biased or

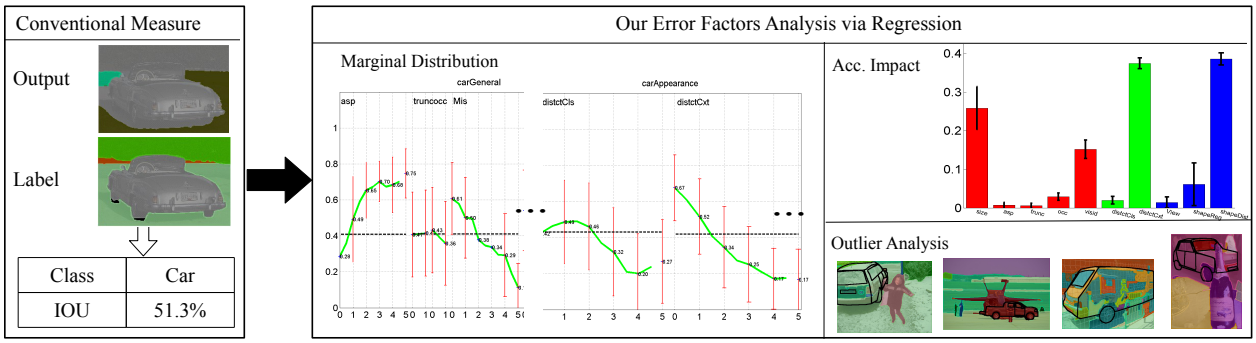


Figure 1. Illustration of our error factor analysis for image-labelling of the car category. On the left, the conventional IOU only gives a number, while in our measurements, we find the factors that cause errors. Specifically, the analysis gives the marginal relationship between each error factor (Size, Occlusion etc.) and the IOU, the joint impaction of error factors and outlier cases that not well explained by the regression model.

home-brewed data, we hardly know how much specific motivated challenges are solved in real world.

In recent years, researchers are moving towards dataset that is from wild scene scenario to prevent machine learning from being distracted by biased images or datasets that are lack of generality. Pinto et al. [24] presented that a natural dataset includes more variations rather than just number of classes. Torralba et al. [29] proposed cross dataset validation which concluded that large real world datasets, like PASCAL [10], ImageNet [8] and SUN [35], are well sampled respecting real world scenario for recognition. These works, on one hand, indicate that the datasets composed from real world samples seem to be the one we should deal with. On the other hand, the high mixture of variations from these data makes it unclear that how well the new proposed algorithms handle what they claimed, which can hardly explained by the overall IOU scores and several hand picked examples.

Despite of collecting unbiased datasets, many works also focus on improving the evaluation criteria to better understand and compare the performance of algorithms as fair and complete as possible. In one aspect, researchers are trying to design multiple general summary scores for evaluation [30]. Pont-Tuset et al. [25] argued that under the situation of image labelling or detection, criteria with the precision-recall curves for object and parts are better measurements reflecting the quality of semantic segments. Nevertheless, they did not try to understand and model the errors that fail the semantic labelling algorithms. In another aspect, some works aimed at modelling a special type of the errors. For example, Divvala et al. [9] presented an evaluation for analysing the role of context beyond the IOU score with respect to the object properties such as size and occlusion. The most relevant work to this paper is proposed by Hoiem et.al [16], they performed a more complete diagnosing for object detection. Specifically, they separated the various of object into several types of hardness such as occlusion, size, aspect ratio, visibility of parts, viewpoint etc., and the analysed impact is done on removing the instances with certain hardness or modelling the false positive examples individually. However, in our case, firstly, the error patterns in image labelling are different from detection, and we give more useful error proposal depend on object characteristics. In addition, we perform a joint regression that considers the interaction between these error factors.

To our best knowledge, this work is the first one that formally analyses the errors in image semantic labelling tasks. In particular, our contributions can be summarized as follows: (1) We provide many useful error factors and the methods to quantize them for evaluation by organizing and taking use of the most recent published object parts, background [7] and 3D pose label [34] of PASCAL. (2) We propose to do error factor analysis for both objects and backgrounding regions, and model the factor impact to the final accuracy by robust regression. Such analysis helps us understand deeper about how much each error factor affects the accuracy given an algorithm. (3) We will share our software to the community in order to help researchers to explore their results. In the experiments, we use the state-of-the-art algorithms for illustration through identifying their issues and suggesting promising directions.

2. Decompose the errors in image labelling

What makes an object or a semantic region like a dog distinctive from other classes? Intuitively, starting from the ground truth information human visually observed from image, we can decompose the description of an object into its shape, appearance and respective context. For a background classes like sky, we can describe the region based on its general appearance. As we know, the confusion would happen if these information is unclear, not discriminative or missing.We thus try to decompose the error factors from this object information perspective. In [16], they factorized the er-

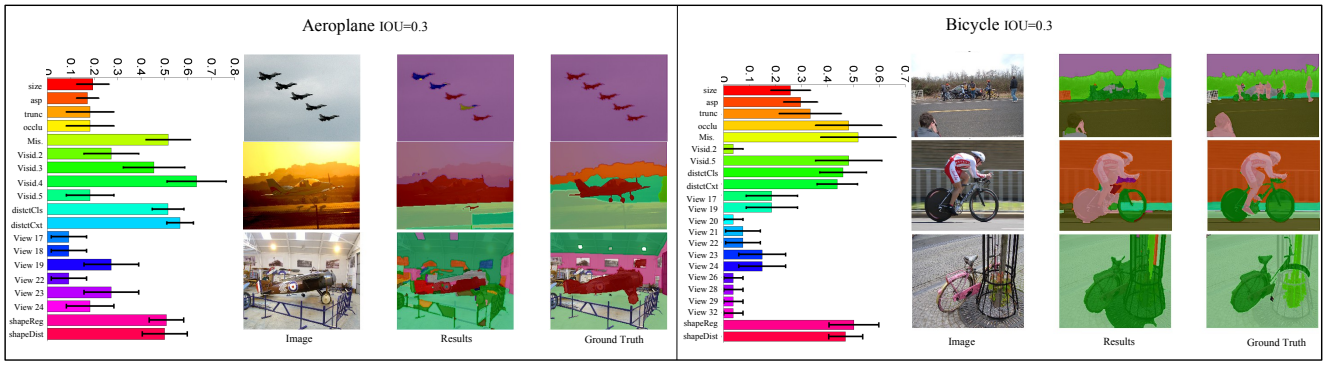


Figure 2. From the figure above, all the images have same IOU score. In each column, within a certain class, it shows that given the same IOU, the value of the different error factors could be very different, which is shown on the left. The y-axis represents the mean value of the corresponding error factors. Specifically, the error of truncation gives much less impact for aeroplane than for bicycle. This indicates different algorithms can have same IOU but very different properties. Between different type of classes, the impact of these error factors can also vary a lot.

rors depended only on object general characteristics, while for semantic segmentation, our summary could provide a more complete analysis of errors for better understanding.

2.1. Error factors proposal

In image labelling, [26] presented an ideal case for semantically segment images, which needs perfect segmentation and well distinctive features. In the following, given a particular class and an algorithm, we will summarize those error factors that fail the algorithm, by analysing each labelled ground truth instance. In addition, to quantify those error factors, the larger the quantized value, the more difficult a case is.

General characteristic. As stated in [16], some characteristics of objects or semantic regions in an image would affect the recognition in general no matter detection and segmentation. We can summarize these character as:

1. Clearness: Size, and aspect ratio, which are computed as the pixel number in the image and region height over width.

2. Visibility of an object: Object truncation and occlusion in the image, which describe whether partial of the object information is missed from the image. In order to quantify this information, for rigid objects, we have such information labelled by [34] as a binary indicator, while we do not have such information for non-rigid objects. In addition, for all objects, we take use of the parts dataset from [7], in which objects are separated into parts, e.g. animals are separated into head, torso and legs. For each part, we use a binary variable to indicate its present (1 for missing), namely "visid i" where i is the id of the part. This is useful to measure the importance of object parts. Last, we measure the level of occlusion and truncation through number of parts missing, and we name it as "Mis" in the experiments.

Appearance error factors. Appearance is considered as the first cues for image labelling [5, 4] as it is practically useful to distinguish regions [7] and handle shape variation or occlusion of non-rigid objects [13].

In order to obtain certain pattern of appearance, most algorithms start their process based on either superpixels [4] within which appearance is often homogeneous or a set of proposed object segments [36]. Then, a discriminative model is built to label the respect regions. However, such a processing suffers from certain difficulties when the object appearance is close to background or not discriminative. To separately analysis and measure these difficulties from appearance perspective, we take use of the ground truth masks and superpixel segmentation. Formally, our appearance error factors are written as,

1. Appearance distinctiveness from the given classes for classification: appearance could be similar between different semantic classes like cat and dog, sky and water. To quantize such distinctiveness, we first build appearance model through support vector regression (SVR) for each object from superpixel oversegmentation [32] of the objects. We use the O2P [5] feature $\mathbf{f}$, yielding a model $s(\mathbf{f}|c) = sigmoid(f_c(\mathbf{f}))$ where c indicates the class, f_c is the SVR score. Then, given a ground truth mask $\mathbf{G}$, we decompose the mask using a set of superpixels $\{\mathbf{S}_i\}_1^{N_o}$. In addition, we compute the distinctiveness of one segment by the posterior $p(c|\mathbf{f}) = \frac{s(\mathbf{f}|c)}{\sum_c s(\mathbf{f}|c)}$. Then the overall distinctiveness could be measure as weighted average by $p(c|\mathbf{f_G}) = \sum_i a_i p(c|\mathbf{f}_i)$ where a_i is the portion of the

area in the ground truth. We call this as "distctCls" for simplicity in experiments and set it as $1 - p(c|\mathbf{f_G})$.

2. Distinctiveness from surrounding context: similarity between the appearance of the target semantic regions and the surrounding regions would induce under-segmentation or over-segmentation errors at the beginning. We quantify this similarity based on the superpixel segmentation upper bound for each semantic region. Formally, given a set of ground truth instance masks $\mathcal{G} = \{\mathbf{G}_i\}_1^{N_g}$, and a set of proposed segments $\mathcal{S} = \{\mathbf{S}_j\}_1^{N_s}$ generated from the state-of-the-art segmentation method [2], we select the best subset ${\mathcal{S}_s}^*$ from $\mathcal{S}$, such that $IOU^* = \max_{\mathcal{S}_s \subseteq \mathcal{S}} IOU(\mathcal{G}, \mathcal{S}_s)$, and we quantify the distinctiveness from surrounding context by the max $1 - IOU^*$. We call this as "distctCxt" in experiments.

Shape error factors. Addition from the appearance, shape is also regarded as the key characteristic of objects that researchers are focusing on [12] and recently using to combine in image labelling [38, 33]. Particular, an object with certain shapes or special distinctive structures such as dog and cat faces [1] would make it easy for identifying the target regions. However, misleading shape cues may also yield error classification and segmentation. In our case, we summarize the confusion from object shape as follows,

1. Object viewpoint: For a rigid object, we obtain the viewpoint through the object orientation provided by PASCAL 3D data [34]. In order to obtain the impact of certain viewpoints, we quantify the perspective by the given two rotation annotations, i.e. azimuth and elevation. We quantize the rotation angle with bin of 45 degree, yielding 8 bins for azimuth in the range of $[0, 360]$ and 5 bins for elevation in the range of $[-90, 90]$. Thus, each object is arranged into one of the 40 bins by first quantizing in azimuth. For example, if the 11_{th} bin of an object is one, then the bin id for azimuth and elevation are the the 3_{nd} (90) and 2_{nd} (-45) respectively. For non-rigid objects, as there is no 3D label available, we do not consider the geometry view angle.

2. Shape regularity: Certain object like chair in side view might have highly various contour that effects the segmentation accuracy, while other object like monitor in a convex shape could be less various. Within an object, different view points produce different shape regularity. To quantize the shape complexity, we take the entropy of curvature [23] to be the numerical evaluation method.

3. Shape distinctiveness: Objects with similar shapes could have similar gradient statistics yielding confusion. To measure the shape distinctiveness of a certain object, we build a mixture of shape model for each object through deformable part based model (DPM) [12]. In particular, for rigid objects, we directly train a shape model $s_s(\mathbf{s}|c) = sigmoid(g_c(\mathbf{s}))$ where s_s is the output score, and g_c is the DPM responding score given the bounding box of the ground truth segment $\mathbf{s}$. Then the posterior is written as $p(c|\mathbf{f}) = \frac{s_s(\mathbf{f}|c)}{\sum_c s_s(\mathbf{f}|c)}$. However, for non-rigid object, the variation of the object could be large. Thus, additional to the mixture of whole model, we train parts model such as the head model and torso model $\{s_s(\mathbf{s}_i|c)\}$ for human and animals, where i is the part id. Then, given the ground truth shape $\mathbf{s}$, we measure the distinctiveness as its posterior $p(c|\mathbf{s}, \{\mathbf{s}_i\}) = \frac{1}{2}p(c|\mathbf{s}) + \frac{1}{2}\sum_i a_i p(c|\mathbf{s}_i)$ where a_i is the area portion of part i. We call this as "ShapeDist" in our experiments and set it as $1-p(c|\mathbf{s}, \{\mathbf{s}_i\})$.

2.2. Discover the relationship

From the proposed error factors, we wish to get several useful information. First, in each factor, we want to know in which numerical range the factor impacts the accuracy most. Second, we want to know how much of each factor affects the final results respectively. For example, by quantify object perspective, we want to know which viewpoint impact the IOU and how much perspective affects the IOU overall. Targeting at the impact, we can respectively increase the lacked training images or design algorithms, which makes it effective in improving the current system.

To get the information of which numerical range affects the accuracy, we first perform a non-parametric regression to the marginal distribution of each error factor. This can also help us to get an intuition of how the error factors' values are related with the final accuracy. For each marginal data, we perform a kernel regression with Laplace kernel due to the outlier points we have. Formally, the regressed marginal distribution is written as,

$$E_r(r|e_i, c, a) = \frac{\sum_j r_j \kappa(e_{ij}, e_i|c, a)}{\sum_j \kappa(e_{ij}, e_i|c, a)},$$
$$\kappa(e_{ij}, e_i|c, a) = exp(-\frac{|e_{ij} - e_i|}{b_i})$$

The regressed curve illustrate the relationship of the accuracy and the error factors. Additionally, we can obtain the prediction interval from the distribution through performing a bootstrapping over the data points.

To get the impact of each error factors. Formally, given the k proposed error factors, i.e. $\mathbf{e} = [e_1, e_2, \ldots, e_k]$, we model their relation with the computed results accuracy r (IOU in our case) by formulating it as the expectation of the

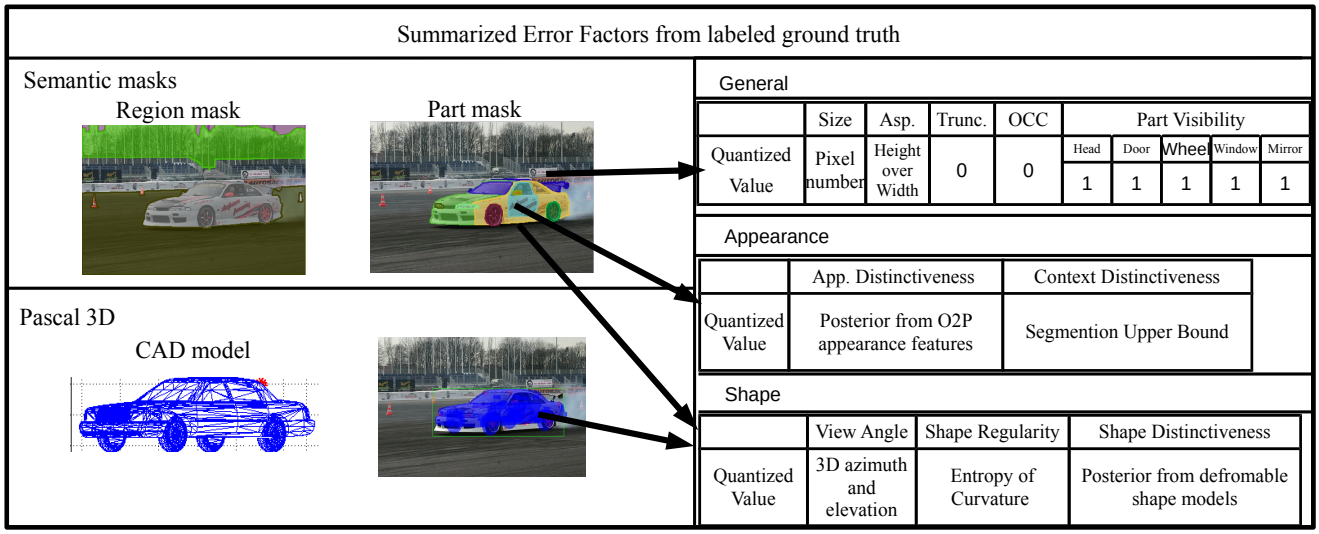


Figure 3. The summary of how we generate the quantized value for each the error factor and the ground truth data we have for performing the analysis.

posterior distribution,

$$E_r(r|\mathbf{e},c,a) = \int_r rp(r|\mathbf{e},c,a)dr, \quad (1)$$

$$p(r|\mathbf{e},c,a) = Laplace(\mathbf{w}^T\mathbf{e},b) = \frac{1}{2b}exp(-\frac{|r-\mathbf{w}^T\mathbf{e}|}{b})$$

where c is a specific class, and a is the given algorithm. As real data has many outliers, we chose the robust linear regression [21] with the Laplace distribution to formulate the posterior. The weights of factors $\mathbf{w}$ and variance b can be estimated by maximum likelihood estimation (MLE). Through such probability perspective, on one hand, we can get the significance level and confidence interval of the estimated impact of each factors, i.e. the absolute value of $\mathbf{w}$, which allows us to target at the important factors with large absolute impact value w_k. On the other hand, we are able to estimate the predict interval given the error factors through $p(r|\mathbf{e},c,a)$, that allows us understand how well the algorithm performs given certain difficulties.

Nevertheless, in our factor model, the relation between our numeric value of the error factors and accuracy might not be monotonic. For instance, the the accuracy might be high in certain perspective while low on others. This is also shown in the experiments of [16]. Thus, in order to utilize the robust regression, we try to make the factor's value monotonic with the regressing value. Here we take the perspective error factor as an example. The same strategy is applied for other factors such as aspect ratio. Practically, we first separate the numeric range of into multiple bins $\{b_i\}_{i=1}^n$ and regard each bin as a dummy variable, depended on which we do robust regression given to find out the impact of each dummy variable $\{m_i\}_{i=1}^n$. Intuitively, the regressed impact m_i can be regarded as the difficulty level of b_i, i.e. the more negative the m_i is, the more difficult when the numeric bin $b_i = 1$ is for the algorithm. In this formulation, the impact of an error factor is defined as the drop of accuracy with per-unit increasing of respected difficulty level. Then, for perspective, we can reunion the binary variables $\{b_i\}_{i=1}^n$ into a single numerical variable in taking the value of $\{i = 1, \cdots, n\}$ representing difficulty level from the descend sorting of $\{m_i\}_{i=1}^n$. This makes the value of the error factor monotonously related with the accuracy. Finally, we can generate the impact of each factor by redoing the robust regression.

To ensure the regressed impact value of all error factors are comparable, after reunion, we need to normalize the integral value into the range of $[0, 1]$.

3. Experiments and insights

In this session, we present the analysis given the state-of-the-art segmentation results over the PASCAL VOC 2010 dataset, and we will publish the toolbox for researchers to better analyse their algorithms.

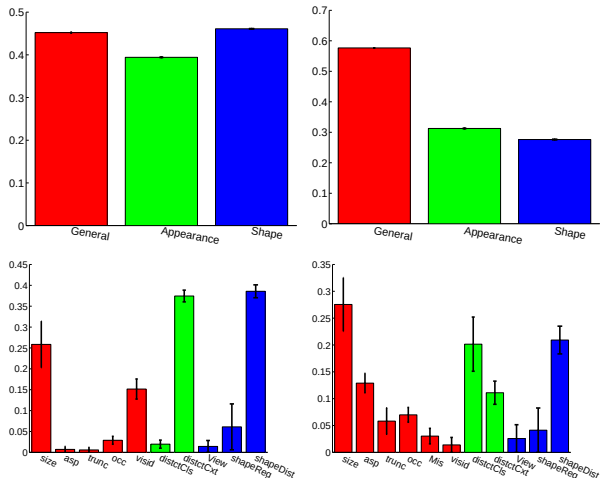
Implementation details. We train our appearance and shape models on the training set of PASCAL VOC 2010 detection train set, and validate those model on the validation set. We perform the factor analysis over the corresponding train-validation set. The ground truth segmentations of the objects and background regions for training our models are provided by [15, 7]. For balancing the weights of different error factors, we normalized all the value into the scale of [0,1] through the 98% and 2% decile value over all categories, and we regard the points at the tail to be outliers. To avoid too much correlation between size, part missing and our appearance, shape error factors, we manually sup-

press the appearance, shape error factors to be the minimum quantized value when the value of the size factor is smaller than certain threshold, i.e. 0.1, or the value of part missing is high, i.e. larger than 0.8.

Measured algorithm. Second order pooling (O2P) for image-labelling [5] is currently the state-of-the-art segmentation method on PASCAL VOC segmentation challenges almost every year. In detail, it first use CPMC segmentation [6] to generate object segment candidates and label them based on the features from second order pooling. The feature embedded LBP for describing the appearance and SIFT for describing the gradient shape information. To produce a full label of an image, we first have the objects labelled exactly by [5], then we label the rest background regions by replacing CPMC segments to superpixels [32].

3.1. Error factor analysis

To illustrate our idea of analysing the characteristic of an algorithm, Fig. 4 and Fig. 5 gives an example of detailed error factor analysis for two categories, i.e. the factor analysis for car and bottle. We include other analysed results in the supplementary material. By checking the left column of Fig. 4, the impact summary graph in the first row tells us the three types of factors give almost the same influence on the accuracy for the car category. The figure in the second row tells us the expected impact of each error factor and its variance. The number on the y-axis gives the value of expected accuracy decrease given per-unit increase of corresponding error factors. We can see that for general factors, the size



(a) (b)

Figure 4. Impact analysis results. The impaction of each error factor for (a) car and (b) bottle. The first row gives the impact summary of three error types and the second row gives the impact and variance of each error factors. In the car's detailed impact figure, we drop the plot bar of "Mis" because its regressed significance is less than a plotting threshold.

and parts missing gives relative strong influence on the accuracy (25% and 17%). For appearance factors, the major influence is from the factor of distinctiveness from context, where we know the O2P algorithm is major infected (0.38) because the appearance of car is close to its surrounding background. For shape error factors, we find strong correlation between the O2P segmentation accuracy with our shape distinctiveness, which indicates that when the shape of car is not distinctive, e.g. similar with bus, the O2P would also be negatively influenced. Upon comparing between different classes in Fig. 4, we can observe that the impaction of error factor varies between different classes. For example, in appearance error type, the IOU accuracy of bottle is major infected due to appearance similarity between bottle and other classes rather than the difficulties from appearance similarity to surrounding background. From such information, we suggest one may need improvement on the segmentation algorithm for car, but additional information, e.g. context for bottle to handle appearance similarity.

In addition, from Fig. 5, we visualize the information of marginal distribution for each error factor to illustrate the correlation and impact of each numeric range of the error factors. The impact is large if the change between two consequential interval points is large. By checking the marginal distribution of car category, we can obtain the information that for the error factor of size, the impact is stronger when the size becomes smaller while the impact becomes weaker when the size is relatively large. In appearance, the "distctCxt" is strongly linear related with IOU, and in shape, certain view points gives positive impact such as view 18, which is within $[45, 0]$ and $[90, 0]$ for azimuth and elevation. This indicates the side view of car is better modelled than the others. From such information, we can target at increasing the training data in the respective sizes or perspectives. Notice rather than conditional distributions as shown in [16], the marginal distribution may give a wider interval, but statistically provides a better expected impact value for practical usage.

Finally, in Fig. 6 we show several examples of outlier instances detected, that include surprising misses and surprising hits. These instances are out of 90% prediction interval by our model. We aimed to discover additional factors that may cause these outliers. However, we find these cases can be explained by our proposed error factors. By checking the calculated error factors' value of these outliers, the surprising cases are mostly due to the non-linear properties of the data, i.e. one factor gives too much impact than expected. For the surprising misses of car and bottle, we have low difficulties for most factors such as size and perspective, but they suffer mostly from cluttered segmentation (high distctCxt). On the opposite, for the surprising hits, though we have heavy occlusion or truncation, the good segmentation (low distctCxt) gives it a surprising improvement on

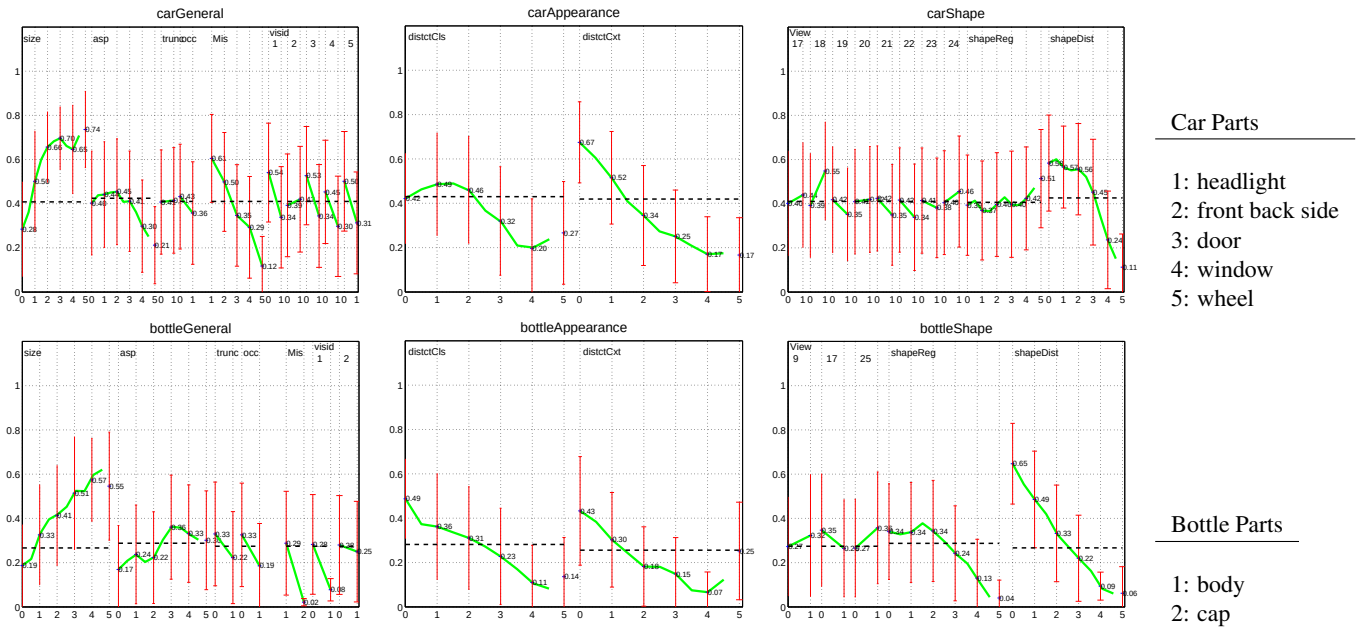


Figure 5. Two examples of marginal analysis from kernel regression and 95% interval, i.e. car and bottle, for illustrating. On the right, we show our corresponding part names for each id in visid section of general error types. We mark out the regressed value and perdition interval on several internal points of each factor.

the accuracy.

4. Issues and Suggestions

From the above analysis, we can see, for each class, the impacts vary a lot, which means not an algorithm can handle all the errors in a single framework. It is important for us to target at respective errors and move forward step by step. In the following, we give promising future suggestions on image labelling tasks, which can also be useful for detection to some extend.

Data augmentation. As demonstrated in algorithms from large data [14], sufficient large data would help us solving the cases such as small objects, rare perspective or shape for detection, which is also possible to largely help segmentation. Though currently it might be difficult to collect large human labelled segmented data, it is promising that we can learn from synthesis. In our opinion, recent approaches [27, 3, 31] using 3D data for solving the 2D to 3D problem are good beginnings.

Figure 6. Outlier analysis of the two illustrating classes. We show the instances with their IOU calculated from the given results significantly different from the prediction using our model. The ground truth is line out with black contour and the prediction is visualized in VOC color mapping [10].

Algorithm improvement. From our analysis, e.g. the error factor analysis of car, we see the image-labelling problem is not a single segmentation issue, but should be solved jointly with detection, object composition estimation and scene context. Scene layout and surrounding context could help small objects [7], shape and composition provide instruction to improve the raw segments, such as constraint the under segmentation, etc. We encourage the joint works that have been proposed in [38, 33]. However, many of them suffer from learning with small amount of data or sometimes a heuristic model which is not general enough for usage. We would like to see more general models that can be possible to learn from a large amount of synthesised data in the future.

Limitations and future improvements. Though we can obtain a lot information from such analysis, admittedly, this analysis suffers from some limitations on factorization and modelling the error factors. Firstly, the PASCAL VOC set for us to use might not be large enough to cover all the numeric range of the proposed errors. Secondly, we do not have the whole information of all the objects such as the 3D pose information for non-rigid objects like cats, and it is also not easy to quantize the pose through viewpoints for

a non-rigid object even given its 3D information. Thirdly, some of our error factors might be correlated, such as the occlusion and part missing, or the view points and shape distinctiveness. This may yield that the impact of one factor might be absorbed by the more correlated one, while we can partially compensate this limitation by looking at the conditional or marginal distribution. In the future, we would try to improve the data and labelling.

5. Conclusion

In this paper, we propose an error factor analysis evaluation approach for wild scene image labelling tasks. Our evaluation decompose the IOU accuracy measurement into error factors from the object composition perspective, and we formulate the relationship between the factors and accuracy through robust regression from joint and marginal distributions. From the analysis, we find significant relationship between some of our proposed error factors and the accuracy, which gives researchers a better understanding of the popular PASCAL VOC segmentation data we are dealing with and the characteristics of a state-of-the-art algorithms. Based on such analysis, we proposed several suggestions which are promising directions to improve the performance of current algorithms.

Acknowledgements

The research work of Peng Wang, Alan Yuille is supported by ONR N00014-12-1-0883 and NIH Grant 5R01EY022247-03.

References

[1] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. D. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, pages 3378–3385, 2012.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. In *TPAMI*, pages 898–916, 2011.

[3] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

[4] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. Gonzlez. Harmony potentials - fusing global and local scale for semantic image segmentation. In *IJCV*, pages 83–102, 2012.

[5] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV (7)*, pages 430–443, 2012.

[6] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1312–1328, 2012.

[7] X. Chen, R. Mottaghi, X. Liu, N.-G. Cho, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[9] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, pages 1271–1278, 2009.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. In *TPAMI*, pages 1915–1929, 2013.

[12] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester. Cascade object detection with deformable part models. In *CVPR*, pages 2241–2248, 2010.

[13] S. Fidler, R. Mottaghi, A. L. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *CVPR*, pages 3294–3301, 2013.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[15] B. Hariharan, P. Arbelaez, L. D. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, pages 991–998, 2011.

[16] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV (3)*, pages 340–353, 2012.

[17] A. Kae, K. Sohn, H. Lee, and E. G. Learned-Miller. Augmenting crfs with boltzmann machine shape priors for image labeling. In *CVPR*, pages 2019–2026, 2013.

[18] F. Li, J. Carreira, G. Lebanon, and C. Sminchisescu. Composite statistical inference for semantic segmentation. In *CVPR*, pages 3302–3309, 2013.

[19] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, pages 1712–1719, 2010.

[20] R. Mottaghi, X. Chen, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[21] K. P. Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT Press, Aug. 2012.

[22] H. Myeong and K. M. Lee. Tensor-based high-order semantic relation transfer for semantic scene segmentation. In *CVPR*, pages 3073–3080, 2013.

[23] D. L. Page, A. Koschan, S. R. Sukumar, B. Roui-Abidi, and M. A. Abidi. Shape analysis algorithm based on information theory. In *ICIP (1)*, pages 229–232, 2003.

[24] N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1), 2008.

[25] J. Pont-Tuset and F. Marqués. Measures and meta-measures for the supervised evaluation of image segmentation. In *CVPR*, pages 2131–2138, 2013.

[26] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, pages 1–8, 2007.

[27] K. Rematas, T. Ritschel, M. Fritz, and T. Tuytelaars. Image-based synthesis and re-synthesis of viewpoints guided by 3d models. In *CVPR*, 2014.

[28] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[29] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011.

[30] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):929–944, 2007.

[31] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *CVPR*, 2014.

[32] P. Wang, G. Zeng, R. Gan, J. Wang, and H. Zha. Structure-sensitive superpixels via geodesic distance. In *International Journal of Computer Vision*, pages 1–21, 2013.

[33] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan. Semantic segmentation without annotating segments. In *ICCV*, pages 2176–2183, 2013.

[34] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

[35] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[36] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. In *CVPR*, pages 1923–1930, 2013.

[37] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes. Layered object models for image segmentation. In *TPAMI*, pages 1731–1743, 2012.

[38] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702–709, 2012.

[39] Y. Zhang and T. Chen. Efficient inference for fully-connected crfs with stationarity. In *CVPR*, pages 582–589, 2012.

[40] T. Zller and J. M. Buhmann. Shape constrained image segmentation by parametric distributional clustering. In *CVPR (1)*, pages 386–393, 2004.