Semantic Windows Mining in Sliding Window Based Object Detection

Junge Zhang¹, Xin Zhao¹², Yongzhen Huang¹, Kaiqi Huang¹, Tieniu Tan¹

1. National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
2. Department of Automation, University of Science and Technology of China

{jgzhang, xzhao, yzhuang, kqhuang, tnt}@nlpr.ia.ac.cn

Abstract

This paper studies the problem of end-to-end windows mining directly from detection output. Traditional object detection systems approach this problem in an ad-hoc manner, say, Non-Maximum Suppression (NMS). Beyond NMS, multi-class context modeling has been explored thoroughly recent years. But all these methods put their emphasis on eliminating false positive windows rather than improving recall. To address this problem, we firstly study this problem and propose semantic windows mining. To improve recall, we propose Selective Forward Search (SFS) which keeps most of the semantic windows while substantially reduces the number of false positives. After SFS, to improve precision, we present the end-to-end windows mining by means of similarity refining optimized for mean Average Precision (mAP) and overlap regression. We show a noticeable improvement on the PASCAL VOC datasets in both recall and precision.

1. Introduction

Over the past decade, sliding window approaches to object detection have been demonstrated impressively on many challenging datasets such as PASCAL VOC [12]. Recent progress mainly focuses on representation [5, 8, 16, 17], windows sampling strategy [1] and context modeling [6, 7, 15].

Problem: This paper considers a different problem of end-to-end semantic windows mining directly from raw detection output. The Raw Detection Output (RDO) refers to the windows pooled from detectors without any post-processing. A semantic window refers to a true positive window. Most of existing systems use various heuristic post-processing methods such as NMS. Besides, NMS disconnects the relation between training and testing phases [6]. Multi-class context model [6] tries to address this problem via context learning. But the recent popular context model [6, 7] usually requires multi-class recognition systems’ output.

Hence, it limits its effective application on single class detection. More important, these methods mainly focus on improving precision, while usually failing to improve recall. In [3, 2], they study the post-processing problem in particular multiple object detection.

As a matter of fact, we observe RDO (e.g., RDO from [8]) contains sufficient semantic windows. This indicates they have potential to mine more semantic windows and obtain a system with both high recall and precision. Therefore, it is necessary and possible to study and refine the end-to-end post-processing directly from RDO.

Proposal and contributions: In this work, we propose a semantic windows mining framework. As object detection system is usually evaluated with precision-recall curve, each window is associated with a score. This paper tries to associate each RDO with a score to make the semantic windows rank relative higher than others (this is why we call our method as semantic windows mining). Due to the disadvantages of multi-class context model [6], we focus on utilizing information from single class which is more flexible and applicable. To the best of our knowledge, it is the first time to precisely study the semantic windows mining without multi-class information. The contributions of this paper lie in: 1) we firstly quantized analyze the drawbacks of common used NMS; 2) To overcome the drawbacks of NMS, we propose an effective Selective Forward Search which focuses on high recall while eliminating most of non-semantic windows; 3) Then the semantic windows mining is formalized as similarity refining and overlap regression with the goal of high precision and recall.

2. Motivation

Suppose the sorted bounding boxes in descending order based on their scores in an image are $Q = \{q_1, \ldots, q_n\}$. NMS greedily selects the windows with the highest score while discarding those that are at least 50% covered by a previously selected detection [8]. In
this paper, the RDO is from PBM [8]. Fig. 1 depicts comparison in recall between the final output with NMS and RDO. For all categories, the recall of the system with NMS is reduced much lower than that of RDO. On average, the recall decreases by about 23%. But on positive side, the recall of RDO is as high as 90.5% which means it contains enough semantic windows and indicates the potential to obtain both high recall and precision. The two observations are the basic motivation of this paper.

Then we take a “data-driven” approach to analyze the reasons behind this phenomenon. Fig. 2 gives two representative pairs of examples. The left one is with NMS (we print all detection windows) and the right one is RDO (we only print the true positives for clearness.). Firstly, if two adjacent windows in Q have completely different size (left pair in Fig. 2), they tend to cover two different objects. But NMS is prone to discard these smaller windows following bigger windows in Q. Similarly, if two windows with apparently different aspect ratio also tend to cover two different objects. Thirdly, a window with larger overlap with ground truth (we mention it as overlap later for short.) does not always have a higher score than that with less overlap (see the window covering the person with white T-shirt in the right pair image in Fig. 2). The window with overlap 0.42 has a stronger response of 0.26 than the window with overlap 0.73. We call this situation as non-increasing phenomenon. The two steps mainly tackle the problem of non-increasing phenomenon. Before them, we propose a Selective Forward Search (SFS) to substantially reduce the number of windows while remain high recall.

### 3.1. Selective Forward Search

Sliding window detector will produce many similar windows around an object. Hence, clustering of windows can be performed to determine which windows belong to the same object roughly. K-means is used in this work. In contrast to NMS, SFS avoids the bias to different size and aspect ratio. Thus SFS is capable of high recall. Algorithm 1 explains the details of SFS.

In SFS, each window content is represented by spatial feature, appearance feature and score feature. The spatial feature is defined as: \( (x_1, y_1, x_2, y_2, w, h, ar) \). \( ar = \frac{1}{1 + \exp(-1.5 w/h)} \) denotes the aspect ratio. The coordinates \( x_1, y_1, x_2, y_2 \in [0, 1] \) are normalized by the image width and height. The score feature is defined as \( \phi(s) = \frac{1}{1 + \exp(-0.5s)} \). LAB color histogram \( h_{\text{lab}}(6+6+6) \) is used as appearance feature which is reasonable in the same image. Thus, a 26-dimensional feature vector is built for the window \( B \):

\[
g = (x_1, y_1, x_2, y_2, w, h, ar, \phi(s), h_{\text{lab}})
\]

### 3.2. Similarity Refining

If the similarity between a window and positive training exemplars is larger, the confidence of being a semantic window will be larger. Thus, the exemplar similarity is used to refine the detection’s confidence. The distance we use is defined as

\[
d_W(q, p) = \|q - p\|_W = (q - p)^T W (q - p)
\]

The positive semi-definite matrix \( W \), on one hand, is expected to be discriminatively trained, and on the other

\[\text{Figure 1: The quantized results of recall for final output with NMS and RDO.}\]

\[\text{Figure 2: Window size, aspect ratio and non-increasing phenomenon are the main reasons why NMS fails. (Best view in color).}\]
hand, the evaluation criteria (e.g., mAP) for ranking quality is desired to be reflected. Although mAP is non-differentiable, Metric Learning to Rank (MLR) [13] based on Structural SVM can tackle this problem and is used in this work via solving:

\[
\arg \min_{W, \xi} f(W, \xi) = tr(W) + \frac{C}{n} \sum_{x \in \chi} \xi_x
\]

s.t. \( W \succeq 0, \xi \geq 0; \forall x \in \chi, y \in \gamma : \langle W, \psi(x, y) \rangle \geq \langle W, \psi(x, y) \rangle + \Delta(y^*_x; y) - \xi_x \)

where \( \chi \) is the training sets with \( n \) samples and \( \gamma \) is the set of all possible rankings. \( y^*_x \) is the best or true ranking for \( \chi \). \( \Delta(y^*_x, y) \) is the loss produced by predicting \( y \) rather than \( y^*_x \). \( \xi_x \) is the slack variable and \( C \) controls the slack trade-off.

The partial order feature [10] is adopted to encode input-output pairs \( \psi(x, y) \). As mAP (the score for a perfect ranking is 1) is used as the criteria for detection system, the loss function based on mAP becomes:

\[
\Delta(y^*_x, y) = score(y^*_x) - score(y) = 1 - score(y)
\]

\[
= 1 - \frac{1}{\sum_{k=1}^{n} Pr\text{ec}@k(y) \mathbb{1}_{[k \in \chi^+_q]}} \sum_{k=1}^{n} Pr\text{ec}@k(y) \mathbb{1}_{[k \in \chi^+_q]}
\]

(5)

Pr\text{ec}@k(y) means the precision out of the first k returned of a ranking \( y \).

Given a set \( \{ex_1, ..., ex_i, ..., ex_m\} \) of exemplars from training sets (including ground truth and detection results) with overlap larger than 0.7), the similarity between each query exemplar and detection window can be easily computed according to Eq. 3. Finally, the confidence of a window conditional on exemplar similarity is obtained in average form:

\[
P(s|x_{ex}) = \frac{1}{m} \sum_{i=1}^{m} h \left( (x_{ex} - ex_i)^TW(x_{ex} - ex_i) \right)
\]

(6)

where \( h(x) = \exp(-0.05x) \) for normalisation in the range [0,1]. The spatial feature and 960-dim Gist [14] is used as the window representation. To speedup the metric learning process and similarity computation, we project the feature vector onto 30-dimensional subspace by Partial Least Squares (PLS). Moreover, we use a compact exemplar dictionary of size 400 by clustering the complete exemplars with K-means.

3.3. Overlap Regression

After the confidence refining from the aspect of exemplar similarity, now we precisely and explicitly study the relation between window and overlap.

In contrast to traditional structural SVM formulation [4], this paper re-formulates it as a problem of Support Vector Regression (SVR) against overlap. The objective of SVR is:

\[
\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)
\]

s.t. \( \xi_i \geq 0, \xi_i^* \geq 0; \forall i : y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \)

\( \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \)

where \( C \) is the trade-off constant between the margin and slacks. Spatial feature, score feature and Gist are used in regression. We differ from [11] in that [11] focuses on segment categorization which is completely different.

We argue that our formulation is superior to structural SVM [4] or related methods in that 1) The feature space by Eq. 7 is much less than structural SVM; 2) Our method explicitly enforces a desired overlap on the training samples which guarantees the window with higher overlap has higher score. For structural SVM, only the margin between the best and the most violated point is enforced. Ordering among examples beyond the margin is arbitrary.

With the learnt regressor, we can directly predict each window’s overlap. Then we define the final score of a window as:

\[
P(s|B) = w_dP(s|x_d) + w_{ex}P(s|x_{ex}) + w_{ov}P(s|x_{ov})
\]

(8)

where \( P(s|x_d) \) and \( P(s|x_{ov}) \) are the confidence from the detector and overlap regression in the form of logistic function \( \sigma(x) = 1/(1 + \exp(-0.5x)) \). \( w \) is the weight for each component. In our system, we simply use an average weight for them.

4. Experiments

We evaluate our method on challenging PASCAL VOC2007 [12] which is widely used as testbed for de-
tection. The max recall rate and mean Average Precision (mAP) are adopted as criteria. In all the experiments, the RDO is from [8].

**Results of SFS.** In our experiment, \( K \) and \( T_{size} \) are set to 10 and 1.5 empirically. On average, SFS reduces 93.27% of the number of the non-semantic windows and still obtains a high recall of 86.0% with a minor reduction of 4.5% in recall compared with RDO. The recall of baseline method NMS is only 67.9% which indicates SFS provides a strong improvement of 26.6% over the baseline. This result, on one hand verifies the analysis of reasons why NMS fails to obtain a high recall, and on the other hand, it proves the effectiveness of SFS. Reduction of 93.27% makes it feasible for following similarity refining and overlap regression.

**Results of Windows Mining.** Table 1 gives the complete results on PASCAL VOC2007. We compare with two other related methods for fair comparison: NMS and multi-class layout [6] which also focuses on end-to-end learning NMS but based on multi-class spatial interaction. We didn’t compare with the bounding box prediction in [8] because they focus on modifying window position based on parts which is completely different task. The proposed system stably outperforms the two methods noticeably both in precision and recall on all categories. Our system obtains an mAP improvement by 1.7% and 8.2% over baseline and [6]. Compared with other progress in literature, we believe the improvement is in fact very promising on this dataset. Moreover, the proposed method improves recall largely of 13% and 17% over the baseline and [6].

---

**5. Conclusion**

This paper has presented an effective end-to-end learning based windows mining system which firstly proposes and addresses the problem of directly learning output from detectors with single class’s information. Our method addresses the shortcomings of common used NMS, then this paper explores the similarity between training exemplars and detection window and also precisely study the explicit connection between window content and its spatial overlap with regression model. The resulting system achieves promising performance in terms of recall and precision.

**Acknowledgement** This work is funded by National Natural Science Foundation of China (Grant No.61175007, 61175002), the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDA06030300), the National Basic Research Program of China (Grant No.2012CB316302), the Tsinghua National Lab for Information Science and Technology Cross-discipline Foundation (Grant No.Y2U1011MC1).

**References**


