Pyramidal Model for Image Semantic Segmentation

Giuseppe Passino, Ioannis Patras and Ebroul Izquierdo
Queen Mary, University of London, UK
\{giuseppe.passino,ioannis.patras,ebroul.izquierdo\}@eecs.qmul.ac.uk

Abstract—We present a new hierarchical model applied to the problem of image semantic segmentation, that is, the association to each pixel in an image with a category label (e.g. tree, cow, building, ...). This problem is usually addressed with a combination of an appearance-based pixel classification and a pixel context model. In our proposal, the images are initially over-segmented in dense patches. The proposed pyramidal model naturally embeds the compositional nature of a scene to achieve a multi-scale contextualisation of patches. This is obtained by imposing an order on the patches aggregation operations towards the final scene. The nodes of the pyramid (that is, a dendrogram) thus represent patch clusters, or super-patches. The probabilistic model favours the homogeneous labelling of super-patches that are likely to contain a single object instance, modelling the uncertainty in identifying such super-patches. The proposed model has several advantages, including the computational efficiency, as well as the expandability. Initial results place the model in line with other works in the recent literature.

Keywords—semantic segmentation; hierarchical models; probabilistic graphical models;

I. INTRODUCTION

Semantic segmentation of images, that is, the classification of each pixel as belonging to a given category, is becoming a mainstream problem in the computer vision community in the last years. Local classification of pixels is not enough to obtain satisfactory results, due to local appearance ambiguities. The challenge is therefore to model the correlation of the different image parts in a scene to exploit their context.

Several approaches are based on a structural analysis based on probabilistic graphical models such as Conditional Random Field (CRF) [1], that work jointly with an appearance-based model [2] or more often on the result of an independently trained pixel classification component [3], [4]. CRFs have proved very effective to impose short-range constraints between pixels, such as label smoothing through propagation of confidence to ambiguous regions. Even more complex relationships have been modelled, such as mutual position between categories [5] or scene layouts [6]. The extent of the dependencies handled by the CRF is usually augmented by applying it to dense clusters of pixels, called parts or patches, rather than single pixels. The model however clearly shows problems in accounting for long-range dependences of the scale of the whole image. This drawback has been addressed for example by including in the CRF a term depending on global features [7], but the model rigidity does not allow for more advanced representations.

The proposed graphical model is an alternative to favour patch contextualisation and smooth labelling. However, it does so by respecting the compositional nature of the image. This is done by using a representation that is the result of a hierarchical clustering. This iterative clustering approach generates patch sets (super-patches) that represent the scene locally at different scales, and are therefore analysable separately. An example of such structure is given in Fig. 1. A probabilistic model over this graphical structure favours the homogeneous labelling of these clusters that are likely to represent a single object instance.

Few approaches to semantic segmentation resort to hierarchical structures. The multi-scale CRF [6] represents an attempt to consider dependences at different scales. This model does not consider the scene as a structured composition of parts. An attempt of merging a generative object model with random fields is proposed by Larlus et al. [8]. The work is based on a hidden layer of object hypotheses connected in a random field, and therefore not considered in a hierarchy.

II. PROPOSED METHOD

The method for image analysis can be summarised as follows. At first, pixels are clustered in patches (300 per image...
age), obtained through oversegmentation. This is performed using a spectral clustering algorithm [9]. Secondly, patch features are extracted and a probability distribution over the label field is associated to each one of them, by independent classification (Sec. II-A). The hierarchical representation of the image is obtained (Sec. II-B), and used to impose a probabilistic model to correct the initial label distributions (Sec. II-C). The final labels are estimated from the resulting distribution.

A. Patch Visual Classification

The first step in the appearance-based analysis is the feature extraction. In this work, features based on texture, colour and position are used. For texture, a Gaussian Mixture Model (GMM) is employed. A Gaussian Derivatives filterbank [9] associates a frequency response vector to each pixel. A GMM with 40 components is fitted to the frequency response vectors of the entire dataset, obtaining a compact effective representation. The texture descriptor at patch-level is obtained by averaging the mixture component over all the patch pixels. For colour, we use a normalised hue histogram that compensates for the loss of accuracy due to low saturation [10]. The positional features are the normalised coordinates of the patch centre.

The independent patch classification model is discriminative. To obtain the distribution over the label set $\mathcal{L}$ for the $i$-th patch, we use a multi-class logistic regression (LR),

$$p(y_i | \theta; \mathbf{x}) = \frac{\exp(\theta_{y_i} \cdot \mathbf{x})}{\sum_{y' \in \mathcal{L}} \exp(\theta_{y'} \cdot \mathbf{x})},$$

where $\theta$ are the parameters of the model. The feature vector $\mathbf{x}$ is actually obtained by weighted average of the patch features with the ones of the neighbours in the four upper levels in the pyramid, to improve contextualisation.

B. Binary Partition Trees

The image representation used in this work is a Binary Partition Tree (BPT) [11], which is a dendrogram resulting from an agglomerative clustering algorithm. This forms the basis for the proposed pyramidal model. The dendrogram or BPT is also referred as image pyramid in the following. In order to obtain the pyramid, the patches are initially considered as nodes in a graph whose edges represent the patch connectivity. Using a boundary probability map derived for the spectral clustering [9], a Minimum Spanning Tree (MST) for the graph is obtained. The BPT is obtained by recursively merging patch clusters, selecting at each step the merging operation between the ones involving the edges of the MST. Connections are chosen in order of increasing merging cost, thus obtaining a dendrogram. The merging cost for two super-patches $r_i, r_j$, similarly to the original BPT algorithm, is

$$f(r_i, r_j) = \alpha_p f_c(r_i, r_j) + (1 - \alpha_p) f_p(r_i, r_j),$$

where $\alpha_p = 0.5$ is a mixing parameter. The components

$$f_c = N_i \| w(\bar{x}_i - \bar{x}_k) \|_2 + N_j \| w(\bar{x}_j - \bar{x}_k) \|_2$$

$$f_p = \max\{0, \min\{P_i, P_j\} - 2P(r_i \cup r_j)\}$$

are related to the visual features $x_i, x_j$ and the perimeters $P_i, P_j$ respectively. $N_i, N_j$ are the areas of the two super-patches, $w$ is a weighting function that divides the feature difference by the dynamic range of the feature components in the image, and finally $P(r_i \cup r_j)$ is the perimeter of the super-patch resulting from the merging. The $f_c$ term favours the merging of visually similar super-patches, while the $f_p$ term favours patches that are compact. The measure in Eq. (2) is the cost measure for the agglomerative clustering algorithm. An example of dendrogram (image pyramid, or BPT) is in Fig. 1.

C. Pyramidal Model

At first, a probability distribution over the labels is associated to each node in the pyramid. For patches, the independent classification is used. For each super-patch, the distribution is obtained by averaging the initial distributions of all the composing patches. The average is weighted with the area of the single patches, in order to compensate for patches of different dimensions. This is consistent with the fact that the descriptors used at patch-level are histograms, and therefore they have an additive nature. A benefit of the proposed model is that it is easy to integrate further classifiers in the super-patch analysis to enable multi-scale classification that accounts for emergent visual properties.

The pyramidal model is then applied. The underlying idea is that super-patches are recursively merged in the pyramid, so that a certain number of nodes will incorporate object instances in the image (object nodes or super-patches). The hypothesis is that patches belonging to the same object instance are merged before patches belonging to different objects. Given a object node, all the underlying patches will necessarily have the same label distribution, corresponding to the one associated to the object super-patch. It is however difficult to decide which super-patches represent entire object instances, and a static threshold on either the plain or cumulative [11] merging cost in the agglomerative clustering step is bound to errors.

In the proposed model an initial guess on the number of object instances (either a fixed number or depending on the merging costs) is first made. Starting from this hypothesis, the model contemplates the probability that a node which is chosen as object node is effectively the composition of multiple objects, and therefore can be split into two candidate object nodes. An analogous probability is then assumed for the children, in a recursive fashion towards the bottom of the pyramid. On the other side, a node initially chosen as an object node can also be part of a bigger object instance, in which case a merge operation with its sibling will lead to a new candidate object node. Again,
the merge mechanism is applied recursively towards the top of the pyramid. As a result, distributions associated to each super-patch are mixed according to the split and merge probabilities to obtain the final distribution for each patch.

For a super-patch $sp$ we indicate the split and merge probabilities as $p_s(sp), p_m(sp)$ respectively. The independent label distribution is $p_{IND}^N$, and $p_{IND}^i$ for the $i$-th patch. The final patch probability is obtained recursively. Considering the split model in isolation, and indicating as $\{sp_1, \ldots, sp_{N-1}\}$ the super-patches in the path between the patch $i$ and the object node $sp_N$, the modified distribution is $p_i^s = (N_p)p_i^s$,

$$(N)p_i^s = p_{IND}^N(1-p_s(sp_N)) + p_s(sp_N)(N-1)p_i^s. \tag{4}$$

The quantity $(n)p_i^s$ is the label distribution when considering $sp_n$ as object node. The $p_i^s$ is obtained recursively, and the recursion step is

$$(n)p_i^s = p_{IND}^N(1-p_s(sp_n)) + p_s(sp_n)(n-1)p_i^s, \tag{5}$$

for $n \in [1, N]$. The base of the recursion is $(0)p_i^s = p_{IND}^N$.

The merge probability for the nodes above the initial object nodes is integrated in a similar way. Therefore, indicating the super-patches in the path between the object node and the root $sp_N$ as $\{sp_1, \ldots, sp_{N-1}\}$,

$$p_i^m = (0)p_i^m = p_i^s(1-p_m(sp_1)) + p_m(sp_1)(1)p_i^m. \tag{6}$$

The recursion step and the base of the recursion are

$$(n)p_i^m = p_{IND}^N(1-p_m(sp_{n+1})) + \sum_{m} p_m(sp_{n+1})(n+1)p_i^m, \tag{7}$$

for $n \in [1, N-1]$.

The split and merge probabilities $p_s(sp), p_m(sp)$ are obtained as

$$p_{(s,m)}(sp) = (1+\exp(-\theta_{(s,m)} \cdot x_{sp}))^{-1}, \tag{8}$$

where $\theta_{(s,m)}$ are the parameters of the models and $x_{sp}$ are the features used for this estimation. The feature vector used to this end is composed by four component. A constant component accounts as an offset. The second feature is the depth of the super-patch in the pyramid (from the top). The third feature is the symmetric Kullback-Leibler divergence between the distributions of its children. The fourth is the difference between the number of patches in the two children.

### III. Model Fitting

The model fitting is performed by likelihood maximisation over a pixel-level labelled training set. The appearance model is trained separately with a standard iterative maximisation of the patch labelling log-likelihood with a gradient-based optimisation strategy.

For the pyramidal model, the probability distribution for each patch is a summation rather than a product of terms. Therefore the joint maximum a posteriori (MAP) log-likelihood cannot be evaluated efficiently without numerical problems due to the small-scale of the terms. We use an approximation based on Maximum Likelihood (ML), assuming patches as independent given the super-patches distributions. The working principle of the pyramidal model, that is, the spreading of the classification error between all the patches in a super-patch, presents an optimisation challenge. The labelling for ambiguous patches is rectified at the expense of a lower confidence for the well-classified ones. This loss of confidence is not enough to modify the labelling for the latter patches, but the global likelihood is reduced. We therefore introduce in the target function a term accounting for the label field entropy, to favour more homogeneous labellings. The resulting likelihood for the $i$-th image is

$$\log L_i(\theta_{(s,m)}) = \sum_{j=1}^{m} \log(p_i^m(y_{ij})) + \alpha \sum_{j=1}^{m} \sum_{y} p_i^m(y') \log(p_i^m(y')) ,\tag{9}$$

where $\alpha$ is a mixing parameter. The first derivatives of the model with respect to the components $\theta_{(s,m)}$ can be calculated recursively by deriving Eq. (5) and Eq. (7). The
maximum of $\log L(\theta_{(s,m)}) = \prod_{i} \log L_{i}(\theta_{(s,m)})$ is then obtained with iterative, gradient-based optimisation.

### IV. Experiments and Conclusions

The model has been tested with the Microsoft Cambridge dataset MSRC-9 of 9 object classes. The dataset is a collection of natural scenes, mostly outdoor. The contained categories are listed in the results together with their relative occurrence. We tested the pyramidal model, comparing it with the independent patch classification and a basic CRF model. The results are reported in Table I, where they are compared with another state of the art model on the same dataset. Two examples of labelling are shown in Fig. 2. The mixing parameter $\alpha$ has been obtained by validation on the training set.

The results prove promising, also in comparison with the state of the art. The potentiality of the model is not completely exploited. We found that the optimisation fitness function in Eq. (9) and the quality of the final labelling are not tightly coupled, for the reasons mentioned in the previous section. This is the main reason for the introduction of the entropy term in the fitness function. The reason for us to believe that a better model fitting is possible is that for a simplified version of the model with no split and merge features, a manual model fitting led to results by far superior than what obtained by automatic fitting of the same model. Additionally, a significant dependence of the training output on the value of the parameter $\alpha$ has been experienced.

As feature work, improvements in the training, as well as integration of super-patch independent classifiers, are expected to surmount the mentioned shortcomings. Labelling of small-scale, isolated object instances would also benefit from the latter. In general, at present the absence of autonomous super-patch classification prevents the model to account for total failure of the appearance model in correctly classify an entire region.

Finally, an important advantage of the proposal compared to the CRF is the time efficiency. The training time for the CRF is over 40 hours, compared with only 20 minutes for the appearance and pyramidal models.

### Acknowledgements

The research leading to this paper was supported by the European Commission under contract FP7-216444, Peer-to-peer Tagged Media, PetaMedia.

### References


---

Table I

Obtained labelling precision on the MSRC-9 dataset comparing an independent patch classification model (IND), a basic CRF model, and the proposed pyramidal model (PYR), as well as other literature work (LIT).

<table>
<thead>
<tr>
<th></th>
<th>Build. (14.5%)</th>
<th>Grass (30.1%)</th>
<th>Tree (14.1%)</th>
<th>Cow (7.2%)</th>
<th>Sky (13.4%)</th>
<th>Plane (2.8%)</th>
<th>Face (3.2%)</th>
<th>Car (7.5%)</th>
<th>Cycle (7.3%)</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IND</td>
<td>68.2</td>
<td>94.4</td>
<td>72.4</td>
<td>60.5</td>
<td>98.3</td>
<td>41.4</td>
<td>58.9</td>
<td>57.3</td>
<td>61.3</td>
<td>77.1</td>
</tr>
<tr>
<td>CRF [2]</td>
<td>63.0</td>
<td>94.2</td>
<td>68.9</td>
<td>84.4</td>
<td>93.7</td>
<td>75.8</td>
<td>92.9</td>
<td>76.4</td>
<td>86.5</td>
<td>82.9</td>
</tr>
<tr>
<td>PYR</td>
<td>89.5</td>
<td>95.1</td>
<td>79.5</td>
<td>75.9</td>
<td>95.8</td>
<td>57.1</td>
<td>77.1</td>
<td>76.6</td>
<td>73.9</td>
<td>85.8</td>
</tr>
<tr>
<td>LITglob[7]</td>
<td>73.6</td>
<td>91.1</td>
<td>82.1</td>
<td>73.6</td>
<td>95.7</td>
<td>78.3</td>
<td>89.5</td>
<td>84.5</td>
<td>81.4</td>
<td>84.9</td>
</tr>
</tbody>
</table>

---

1 http://research.microsoft.com/vision/cambridge/recognition/default.htm