A Multi-Layer Approach to Superpixel-based Higher-order Conditional Random Field for Semantic Image Segmentation

Li Sulimowicz*, Ishfaq Ahmad*, Alexander Aved†

*Department of CSE, University of Texas at Arlington, TX, †Air Force Research Laboratory, Rome, NY

Abstract

Superpixel-based Higher-order Conditional random fields are known for their effectiveness in enforcing both short and long spatial contiguity for pixelwise labelling tasks. However, their higher-order potentials are usually too complex to learn and often incur a high computational cost in performing inference. We propose an new approximation approach to resolve such problems which is a multi-layer pairwise CRFs. Our approach inherits the simplicity from pairwise CRFs by formulating the higher-order and pairwise cues into the pairwise potentials in the first layer. Essentially, this approach provides accuracy enhancement on the basis of pairwise CRFs without training by reusing their pre-trained parameters and/or weights.

1. The Proposed Methods

1.1. Segmentation as Input (SaI)

Conventionally, the higher order potential shown in Eq. 3 is then formulated as a minimization of a secondary pairwise CRFs between an auxiliary variable $y_c$ and the pixel within the higher order clique.

$$
\psi^{SP}(X_S) = \min_{y_c \in \mathcal{L}} (\phi_c(y_c) + \sum_{i \in c} \phi_c(y_c, x_i)) \tag{4}
$$

This conversion makes us believe that the higher-order regulation eventually “flows” down onto pairwise cliques. Inspired by this observation, we pre-process the original RGB images from the dataset with unsupervised segmentation. We use $s_i$ and $s_j$ as the segment indexes of pixel $i$ and $j$, respectively. Then we store a segmented image wherein each pixel $i$ takes the average RGB value $C_{si}$ of the superpixel that it belongs to. We noted such segmented image as $D_s$. The SaI based pairwise potential is denoted as $\psi^{SaI}(x_i, x_j; D_s)$. For example, if we formulate our SaI potential as $\psi^{SaI}(x_i, x_j) = \mu(x_i, x_j)(\theta_p + \theta_v \exp(-\theta_i |C_{si} - C_{sj}|^2))$, we have our intra potentials that locate inside of the segment as $\psi^{SaI}_i(x_i, x_j) = \mu(x_i, x_j)(\theta_p + \theta_v)$, $s_i = s_j$.

Relation to Robust $P^N$ Potts Model. Inside the pairwise graph built on random fields $(X, D_s)$, pairwise edges can be divided into two groups, the intra-edges denoted as $E_{in}$ and the extra-edges as $E_{ex}$, $E = E_{in} \cup E_{ex}$. The sum of intra-potentials are further decomposed as follows:

$$
\sum_{(i,j) \in E_{in}} \psi^{SaI}(x_i, x_j) = \sum_{c \in S} \sum_{l, j \in c} \psi^{SaI}_i(x_i, x_j), \tag{5}
$$

wherein $N_c(x_s)$ is the number of pairwise edges in the clique $c$ that take inconsistent labels. This equation shows the equivalency of the sum of intra-potentials as the Robust $P^N$ Potts Model in Eq. 3. The sum of all extra-potentials functions as a regulator between segments.

SaI based pairwise CRF takes the following Gibbs energy function:

$$
E(X | D_s) = \sum_{i \in V} \psi^{V}(x_i) + \sum_{(i,j) \in E} \psi^{SaI}_i(x_i, x_j) \tag{6}
$$

1.1. Segmentation as Input (SaI)

Conventionally, the higher order potential shown in Eq. 3 is then formulated as a minimization of a secondary pairwise CRFs between an auxiliary variable $y_c$ and the pixel within the higher order clique.

$$
\psi^{SP}_i(x_i, x_j) = \min_{y_c \in \mathcal{L}} (\phi_c(y_c) + \sum_{i \in c} \phi_c(y_c, x_i)) \tag{4}
$$

This conversion makes us believe that the higher-order regulation eventually “flows” down onto pairwise cliques. Inspired by this observation, we pre-process the original RGB images from the dataset with unsupervised segmentation. We use $s_i$ and $s_j$ as the segment indexes of pixel $i$ and $j$, respectively. Then we store a segmented image wherein each pixel $i$ takes the average RGB value $C_{si}$ of the superpixel that it belongs to. We noted such segmented image as $D_s$. The SaI based pairwise potential is denoted as $\psi^{SaI}_i(x_i, x_j; D_s)$. For example, if we formulate our SaI potential as $\psi^{SaI}_i(x_i, x_j) = \mu(x_i, x_j)(\theta_p + \theta_v \exp(-\theta_i |C_{si} - C_{sj}|^2))$, we have our intra potentials that locate inside of the segment as $\psi^{SaI}_i(x_i, x_j) = \mu(x_i, x_j)(\theta_p + \theta_v)$, $s_i = s_j$.

Relation to Robust $P^N$ Potts Model. Inside the pairwise graph built on random fields $(X, D_s)$, pairwise edges can be divided into two groups, the intra-edges denoted as $E_{in}$ and the extra-edges as $E_{ex}$, $E = E_{in} \cup E_{ex}$. The sum of intra-potentials are further decomposed as follows:

$$
\sum_{(i,j) \in E_{in}} \psi^{SaI}_i(x_i, x_j) = \sum_{c \in S} \sum_{l, j \in c} \psi^{SaI}_i(x_i, x_j), \tag{5}
$$

wherein $N_c(x_s)$ is the number of pairwise edges in the clique $c$ that take inconsistent labels. This equation shows the equivalency of the sum of intra-potentials as the Robust $P^N$ Potts Model in Eq. 3. The sum of all extra-potentials functions as a regulator between segments.

SaI based pairwise CRF takes the following Gibbs energy function:

$$
E(X | D_s) = \sum_{i \in V} \psi^{V}_i(x_i) + \sum_{(i,j) \in E} \psi^{SaI}_i(x_i, x_j) \tag{6}
$$
The benefit of this model is that the computational complexity is same as that of the original pairwise CRFs of both the learning and inference processes. However, Sal does slightly suffer from the pairwise regulation loss. This is due to: when \( s_i \neq s_j \), we have \( |C_{s_i} - C_{s_j}| \leq |I_i - I_j| \) compared with \( D \), a gap of the left and the right sides of the equation exists. The proposed Sal-based Multi-layer CRFs Framework (SM-CRF) incorporates another pairwise layer after Sal to compensate the potential drawback of Sal.

### 1.2. Sal-based Multi-layer CRFs Framework

SM-CRF is a two layer pairwise CRFs as shown in Fig. 1 wherein the first layer is Sal and the second layer is a pairwise CRF, which takes the original RGB image as input instead. The initial unary label map is \( U (x) \), the output unary map from Sal layer is denoted as \( U_{sal}^{(1)} (x) \), and the unary map as input to the second layer is denoted as \( U^{(1)} (x) \). We set \( U^{(1)} (x) = \alpha \ast U_{sal}^{(1)} (x) + \beta \ast U (x) \), \( \alpha + \beta = 1 \), which is trained with grid search. This multiple layer structure helps the first Sal layer to incorporate more efficient pairwise regulations which results in better performance at preserving the details of the boundaries.

### 2. Experiments

#### 2.1. Evaluation with DenseCRF

We generated two models based on DenseCRF. 1) Sal, which is the same as DenseCRF, except that our observation is \( D_s \) instead of \( D \). 2) SM-CRF; consists of Sal followed by a second layer which is a pairwise CRF with \( D \).

<table>
<thead>
<tr>
<th>Accurate Ground Truth</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>82.35</td>
</tr>
<tr>
<td>Average</td>
<td>86.63</td>
</tr>
<tr>
<td>IOU</td>
<td>86.76</td>
</tr>
</tbody>
</table>

Table 1: Performance comparison of our methods with DenseCRF \([4]\) and Dense+Potts \([7]\) on MSRC-21 dataset.

<table>
<thead>
<tr>
<th>Accurate Ground Truth</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>87.99</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison of our methods with CRF-RNN \([8]\) and H-CRF-RNN \([1]\) on PASCAL-2012.

From Table 1, the accuracy improvement we gained for

![Figure 2: Examples of qualitative results on MSRC-21 dataset. From left to right, it is DenseCRF, Sal-CRF, SM-CRF, Dense+Potts, and GT.](image)

Sal and SM-CRF compared with DenseCRF is (2.82%, 4.26%) on MeanIoU. The visual comparison is shown in Fig. 2. The model Sal consumed the same amount of time as DenseCRF, and for SM-CRF, the time is slightly more than double of DenseCRF. The proposed scheme decreased the error rate by as much as 16.3%. While, Dense+Potts gained an accuracy equivalent to DenseCRF. This may be due to the trained higher-order term, which can recommend a new label based on its dependent training. This can be both helpful and harmful. Also, we noticed that Dense+Potts performed worse than our models in preserving sharp boundaries of objects.

#### 2.2. Evaluation with CRF-RNN

We evaluated two models based on CRF-RNN \([8]\): Sal-RNN, where we just changed the input to CRF-RNN from \( D \) to \( D_s \) and SM-CRF-RNN where we have two layers, the input of each layer is \( D_s \) and \( D \), respectively. The quantitative results are shown in Table 2 which indicate that we obtained 1% MeanIoU accuracy improvement for SM-CRF-RNN, and for Sal-RNN, we gained 1% Average accuracy improvement. The accuracy boost is equivalent to H-CRF-RNN \([1]\). The accuracy improvement we achieved through our models was on the condition that our models were either not trained or were trained only with a simple grid search. By contrast, H-CRF-RNN requires fine tuning on large dataset of more that 77, 000 images which consists of the PASCAL VOC 2012 and Microsoft COCO dataset \([5]\). The visual results are shown in Fig. 3. The figures shows that our models perform well “filling” the incomplete predicted object.

![Figure 3: Examples of qualitative results on PASCAL VOC 2012 dataset. From left to right, it is Original Image, CRF-RNN, Sal-RNN, SM-CRF-RNN, and Ground Truth.](image)
References


