Contributions
- A superpixel-based top-down saliency model via joint discriminative dictionary and CRF learning
- The use of an objectness potential to include generic object information in saliency estimation

Bottom-up vs. top-down visual saliency
- Bottom-up models only depend on low-level cues such as intensity and color, do not use any class knowledge and try to predict image regions that stand out from their surroundings.
- Top-down approaches are task-driven such as detecting an object instance from a certain category or answering a pre-defined question so they consider high-level information and aim at generating saliency maps for the task at hand.

Our approach

- **Training:** (1) Segment the images into superpixels and represent them with the sigma points descriptor. (2) Extract the objectness maps. (3) Jointly learn the dictionary and the CRF parameters for each object category.
- **Testing:** (1) Segment the images into superpixels and represent them with the sigma points descriptor. (2) Compute the sparse codes of superpixels with dictionaries learned from data. (3) Estimate the objectness map. (4) Use the CRF model to infer the saliency scores.

Superpixel representation
- Each superpixel is represented by 1st and 2nd order statistics of visual features (Tuzel et al., 2006), namely color, edge orientation and spatial information.
- The covariance matrix of feature vectors within $R$:
  \[ C_R = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T \]
  where $z_i$ denotes the $d$-dimensional feature vectors.
- Sigma points (Hong et al., 2009) computed from $C$ by using Cholesky decomposition $C = LL^T$:
  \[
  s_i = \begin{cases} 
  \eta y L^T \mu_i & \text{if } 1 \leq i \leq d \\
  -\eta y L^T \mu_i & \text{if } d + 1 \leq i \leq 2d 
  \end{cases}
  \]
- Final representation of a superpixel is given by:
  \[ x_i(\mu, C) = (\mu, s_1, \ldots, s_d, s_{d+1}, \ldots, s_{2d})^T \]

CRF and dictionary learning
- Construct a CRF model with nodes $Y$ representing the superpixels and edges $E$ describing the connections among them.
- The saliency map is determined by finding the maximum posterior $P(Y | X)$ of labels $Y = \{y_i\}_{i=1}^{M}$ given the set of superpixels $X = \{x_i\}_{i=1}^{n}$.
  \[
  \log P(Y | X, D, \theta) = \sum_{(i,j) \in E} \psi_i(y_i, x_i; D, \theta) + \sum_{i \in V} \gamma_i(y_i, \mu_i; \theta) + \sum_{(i,j) \in E} \phi_{i,j}(y_i, y_j, x_i, x_j; \theta) - \log Z(\theta, D) \]
- **Dictionary potential**

Sparse variables are used to learn a linear classifier and we use the response of this classifier directly as our unary potential:
  \[ \psi_i(y_i, x_i; D, \theta) = -y_i w^T \alpha_i \]
  where $\alpha_i$ denoting the sparse code of superpixel $x_i$.
  \[ \alpha_i(x_i; D) = \arg \min_{\alpha} \frac{1}{2} ||x_i - D\alpha||^2 + \lambda ||\alpha||_1 \]

Objectness potential
This potential measures the likelihood that a superpixel belongs to an object in a class-independent manner as:
  \[ \gamma_i(y_i, \mu_i; \theta) = -\beta y_i (2P(obj | x_i) - 1) \]
  where $P(obj | x_i)$ is the objectness score of superpixel $x_i$ and $\beta$ denotes the parameter of this potential function.

Edge potential
This potential models the interaction between two labels $y_i, y_j$, of two neighboring superpixels as:
  \[ \phi_{i,j}(y_i, y_j, x_i, x_j; \theta) = \rho (1 - \delta(y_i - y_j)) \]
  where $\delta$ denoting unit impulse function.

Learning
We simultaneously learn the CRF parameters $\theta$ and the dictionary $D$ by optimizing:
  \[ (D^*, \theta^*) = \arg \max_{D, \theta} \sum_{r=1}^{M} P(Y^{(m)} | X^{(m)}, D, \theta) \]

Experimental Results
We test the proposed model under three different settings:
- **Setting 1:** Set the parameter of the objectness potential $\beta = 0$ and learn the CRF parameters and the superpixel based dictionary $D$ accordingly.
- **Setting 2:** Learn all the CRF parameters and the dictionary simultaneously.
- **Setting 3:** Extend the first setting by determining the parameter of the objectness potential $\beta$ later via cross-validation, while keeping the learned dictionary $D$ and the other CRF parameters fixed.

Graz-02

<table>
<thead>
<tr>
<th>Bike</th>
<th>Car</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>71.9</td>
<td>61.9</td>
<td>65.5</td>
</tr>
<tr>
<td>49.4</td>
<td>46.6</td>
<td>33.7</td>
</tr>
<tr>
<td>43.7</td>
<td>35.1</td>
<td>41.4</td>
</tr>
<tr>
<td>71.4</td>
<td>13.8</td>
<td>63.8</td>
</tr>
</tbody>
</table>

Table 2: EER results on the Graz-02 dataset.

PASCAL VOC 2007

<table>
<thead>
<tr>
<th>Robot</th>
<th>Bus</th>
<th>Cat</th>
<th>Dog</th>
<th>Horse</th>
<th>Motorbke</th>
<th>Person</th>
<th>Potted plant</th>
<th>Sheep</th>
<th>Sofa</th>
<th>Train</th>
<th>TV-monitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.4</td>
<td>39.0</td>
<td>9.4</td>
<td>5.7</td>
<td>3.4</td>
<td>22.0</td>
<td>30.5</td>
<td>15.8</td>
<td>5.7</td>
<td>31.5</td>
<td>64.9</td>
<td>21.1</td>
</tr>
</tbody>
</table>

Table 2: EER results on the PASCAL VOC 2007 dataset.

Conclusion
- Performing the computations at superpixel level allows us to improve the accuracy of object localizations.
- Generic objectness prior reduces the discriminative power of the dictionary but considering this prior after the joint learning process boosts the performance (setting 3).