[Supplementary Material] Simple-to-Complex Discriminative Clustering for Hierarchical Image Segmentation

Haw-Shiuan Chang and Yu-Chiang Frank Wang

Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

1 Proof of Probabilistic Interpretation

We now prove and detail why, in our proposed segmentation framework, solving the graph-based optimization at each level l in our hierarchy is equivalent to the maximum likelihood estimation (MLE) of the observed $P^l(X)$ (note that X represent the observed image features). To be more specific, we need to verify that:

(I) Minimization of E^l in (8) (of the main paper) is effectively maximizing the log-likelihood of $P^l(X)$.

(II) The above process is equivalent to the fitting of features observed from superpixels at the bottom level using the clusters determined at level l.

To prove (I), we need to relate the minimization of $E^l()$ to the MLE for the log-likelihood of $P^l(X)$. In other words, the following relationship needs to be verified:

$$-log(P^l(X)) \propto E^l(\mathbf{m}_{sp}) = E_D(\mathbf{m}_{sp}) + \lambda E_S^l(\mathbf{m}_{sp}), \tag{1.1}$$

where $\mathbf{m}_{sp}$ denotes the labeling vector for all superpixels collected at the bottom level. The functions $E^l()$, $E_D()$ and $E_S^l()$ are the total energy, data, and smoothness terms, respectively (see Sect. 2.3).

As noted in Sect. 2.4, $P^l(X)$ is the likelihood of the observed image features given the clustering outputs at level l. For image segmentation, the superpixels are not expected to be independent to each other. Thus, we decompose $P^l(X)$ in (1.1) as

$$P^l(X) = \prod_{p_s} P(x_{p_s}) \prod_{p_s,p_q} P^l_{Contour}(p_s, p_q), \tag{1.2}$$

where p_s and x_{p_s} represent the superpixel and its observed feature, respectively. Note that $P^l_{Contour}(p_s, p_q)$ denotes the contour probability between consecutive superpixels p_s and p_q, and (1.2) is (9) of the main paper.

To calculate the likelihood $P(x_{p_s})$ in (1.2) using the clustering outputs determined at level l, we have $P(x_{p_s}) = \sum_y P(m_{p_s} = y)P(x_{p_s}|m_{p_s} = y)$, where $m_{p_s} = y$ indicates that superpixel p_s belongs to cluster y (among $K \times r$ clustering outputs at level l). For simplicity, we have $P(m_{p_s} = y) = 1$ if cluster y contains superpixel p_s, and

0 otherwise. Recall that we advance color, texture, and locality cues for describing image superpixels. Under the assumption of equal priors, we can rewrite $P(x_{p_s}|m_{p_s})$ as $P(x_{p_s}|m_{p_s}) \propto P_{Color}(m_{p_s}|p_s)^{w_C} P_{Text}(m_{p_s}|p_s)^{w_T} P_{Local}(m_{p_s}|p_s)^{w_L}$, where P_{Color}, P_{Text}, and P_{Local} are the likelihoods observed for each feature, which are weighted by w_C, w_T, and w_L (scaled by λ), respectively. Taking the negative log-likelihood over all superpixels, we see that $-\log(\prod_{p_s} P(x_{p_s}))$ is equal to the data term derived in (8), except for segment s is now replaced by superpixel p_s.

On the other hand, the contour probability in (1.2) is $P^l_{Contour}(p_s, p_q) = P(m_{p_s} = m_{p_q}) P^l_{Contour}(p_s, p_q | m_{p_s} = m_{p_q}) + P(m_{p_s} \neq m_{p_q}) P^l_{Contour}(p_s, p_q | m_{p_s} \neq m_{p_q})$, where $m_{p_s} = m_{p_q}$ and $m_{p_s} \neq m_{p_q}$ indicate the cases when superpixels p_s and p_q belong to the same and different clusters, respectively. Similarly, we have $P(m_{p_s} = m_{p_q}) = 1$ if $m_{p_s} = m_{p_q}$, and 0 otherwise. When $m_{p_s} = m_{p_q}$ (i.e., two superpixels belong to the same cluster/segment), we should neglect the contour evidences between two superpixels. Therefore, we set $P^l_{Contour}(p_s, p_q | m_{p_s} = m_{p_q}) = 1$, and the negative log-likelihood of the $\prod_{(p_s,p_q)} P^l_{Contour}(p_s, p_q)$ will be $-\log(\prod_{(p_s,p_q)} P^l_{Contour}(p_s, p_q)) = \sum_{(p_s,p_q)} \mathbb{1}_{(m_{p_s} \neq m_{p_q})}(-log(P^l_{Contour}(p_s, p_q | m_{p_s} \neq m_{p_q})))$, where $\mathbb{1}()$ is the indicator function. Notice that it has exactly the same form of the smoothness term $E^l_S(\mathbf{m}_{sp})$ in (8) (except for operating at the superpixel level). Based on the above derivations for (1.1), the proof for (I) is complete, and we successfully verify that minimizing the proposed E^l in (8) is effectively maximizing the log-likelihood of $P^l(X)$.

To prove (II), we need to show that $E^l(\mathbf{m})$ which we minimize over the segments at level l is effectively the energy term $E^l(\mathbf{m}_{sp})$, observed by the corresponding superpixels at the bottom level using our clustering results determined at level l (and scaled by a constant). In other words, our goal is to prove:

$$E^l(\mathbf{m}) \propto E^l(\mathbf{m}_{sp}), \text{s.t. } m_s = m_{p_s}, \forall p_s \in s. \tag{1.3}$$

To verify the above relationship, we need to associate each feature cue observed at segment and superpixel levels. For color cues, assuming that the color features derived for each pixel are independent, the probability of superpixel p_s belonging to cluster m_s will be the product of probabilities that every pixel in p_s belongs to m_s, i.e., $P_{Color}(m_{p_s}|p_s) = \prod_{p \in p_s} \prod_c P_c(i_p|m_{p_s})$ (as determined in (2) of our main paper, and c is the index of color channels).

By multiplying $P_{Color}(m_{p_s}|p_s)$ of all superpixels p_s in segment s, we effectively obtain $P_c(i_p|m_s)$ over all pixels p in that segment, which indicates the pixel-level probability of segment s belonging to cluster m_s. More precisely, with $m_s = m_{p_s}$, we have

$$P_{Color}(m_s|s) = \prod_{p \in s} \prod_c P_c(i_p|m_s) = \prod_{p \in p_s, p_s \in s} \prod_c P_c(i_p|m_{p_s}) = \prod_{p_s \in s} P_{Color}(m_{p_s}|p_s). \tag{1.4}$$

Note that we drop the scaling factor for simplicity. Similar to the color features, we apply our definition for textural features in (4) and derive the same probability to the

textural features as:

$$P_{Text}(m_s|s) = \prod_t \prod_i P_t(i|m_s)^{w_t \sum_{p_s} P(i|p_s) \frac{P(s|p_s)P(p_s)}{P(s)}} = \prod_{p_s} P_{Text}(m_{p_s}|p_s)^{P(s|p_s)}, \tag{1.5}$$

where t and i is the index of textural channels and the index of bins in textural histograms, respectively. As for the locality features, we have $P_{Local}(m_s|s) = \prod_{p_s \in s} P_{Local}(m_{p_s}|p_s)$ (and can be calculated by (6) and (7)).

We note that, $P^l_{Contour}(s,q)$ is the product of contour probabilities $P^l_{Contour}(p)$ of all pixels p along the boundary between segments s and q, and the values of $P^l_{Contour}(p)$ for all pixels along the same boundary will be the same. Thus, we have $P^l_{Contour}(s,q) = P^l_{Contour}(p)^{|(s,q)|}$, where $|(s,q)|$ is the length of the associated boundary between segments s and q. For superpixels p_s (of segment s) and p_q (of segment q), the contour probability is calculated by $P^l_{Contour}(p_s, p_q | m_{p_s} \neq m_{p_q}) = P^l_{Contour}(p)^{|(p_s,p_q)|}$, where $|(p_s,p_q)|$ is the length of boundary between superpixels p_s and p_q. From the above derivations, we have

$$P^l_{Contour}(s,q) = P^l_{Contour}(p_s, p_q | m_{p_s} \neq m_{p_q})^{\frac{|(s,q)|}{|(p_s,p_q)|}}, \tag{1.6}$$

which is corresponding to the smoothness term determined in (8). Note that we ignore the normalization terms in the above derivations, since they could simply viewed as constants in our energies during the optimization.

With the above derivations and (8), we successfully relate the energy terms observed at the segment and superpixel levels (i.e., $E^l(\mathbf{m})$ and $E^l(\mathbf{m_{sp}})$). In other words, by proving (1.1) and (1.3), we verify that our segmentation effectively fits the features observed from superpixels at the bottom level using the clusters determined at level l.