Combinatorial Optimization in Computer Vision (IN2245)

Frank R. Schmidt
Csaba Domokos

Winter Semester 2015/2016

12. Multilabel Optimization

Multilabel

Binary Segmentation

Any binary image segmentation can be modeled as $x \in \mathbb{B}^n$.

Usually we minimize an energy of the form

$$E(x) = \sum_{i=1}^{n} f_i x_i + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} f_{ij} x_i \bar{x}_j$$

We can rewrite every pairwise term into a symmetric pairwise term using:

$$\begin{aligned} x\bar{y} =& \frac{1}{2} x\bar{y} + \frac{1}{2}(1-\bar{x})(1-y) \\ =& \frac{1}{2}\left(x\bar{y} + \bar{x}y + x - y\right) \end{aligned}$$

We can therefore assume that $f_{i,j} = f_{j,i} \geqslant 0$ for all $i = 1, \dots, n$ and $j \in \mathcal{N}(i)$.

Binary Segmentation

The binary image segmention energy can be written as

$$E(x) = \sum_{\substack{i=1,\\ x_i=0}}^{n} f_i^{(0)} + \sum_{\substack{i=1,\\ x_i=1}}^{n} f_i^{(1)} + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} f_{ij} \delta_{x_i, x_j}$$

$$\delta_{x_i,x_j} = \begin{cases} 0 & \text{if } x_i = x_j \\ 1 & \text{if } x_i \neq x_j \end{cases}$$

Thus, we can rewrite it as

$$E(x) = \sum_{i=1}^{n} f_i(x_i) + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} f_{ij}(x_i, x_j)$$

Probabilistic Interpretation

Minimizing

$$E(x) = \sum_{i=1}^{n} f_i(x_i) + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} f_{ij}(x_i, x_j)$$

is the same as maximizing the conditional distribution (see Lecture 4)

$$P(x) \propto \prod_{i=1}^{n} \exp\left(-f_i(x_i)\right) \cdot \prod_{i=1}^{n} \prod_{j \in \mathcal{N}(i)} \exp\left(-f_{ij}(x_i, x_j)\right)$$

The idea of multilabel optimization is to replace $x \in \mathbb{B}^n$ by $x \in \mathcal{L}^n$, where $\mathcal{L}$ is called the **label space**.

Data Terms

Unary potentials $\Psi_i(x_i)$ of a graphical model and data terms $f_i(x_i)$ are related to one another via

$$\Psi_i(x_i) \propto \exp\left(-f_i(x_i)\right)$$

The unary potentials are the values of a *probability density function* and hence, we usually have $\Psi(x_i) > 0$.

Therefore, we have $f_i(x_i) \in \mathbb{R}$. In other words, we want to allow positive and negative values alike for the data terms of a multilabeling problem.

If we model the unary potential as a **Gaussian distribution** or **Laplacian distribution**, the data term measures a quadratic resp. linear distance from the parameter μ.

Pairwise Terms

Pairwise potentials $\Psi_{i,j}(x_i, x_j)$ of a graphical model and data terms $f_{ij}(x_i, x_j)$ are related to one another via

$$\Psi_{i,j}(x_i, x_j) \propto \exp\left(-f_{i,j}(x_i, x_j)\right).$$

In order to avoid supermodular terms for binary segmentation, we assumed $f_{i,j} \geqslant 0$ or equivalently $\Psi_{i,j} \leqslant 1$. Thus, we cannot use a *probability density function* and have to model a *discrete probability space*. For that reason, we assume that we only have finite many labels in $\mathcal{L}$.

The *conditional random field* framework assumes that we have

$$f_{i,j}(x_i, x_j) = c_{i,j} \cdot d(x_i, x_j),$$

where $c_{i,j}$ may depend on the observation (image gradient, ...) and $d(\cdot, \cdot)$ is a pairwise prior on the label space.

Modeling the Pairwise Term

A straightforward extension of the length term for binary segmentation is the **Potts Model**

$$d\colon \mathcal{L}\times\mathcal{L} \to \mathbb{R}_0^+ \qquad (\ell_1,\ell_2) \mapsto \begin{cases} 1 & \text{if } \ell_1 \neq \ell_2 \\ 0 & \text{if } \ell_1 = \ell_2 \end{cases}$$

If we assume that $\mathcal{L} \subset \mathbb{R}$, we can also use the **Linear Model** or L^1 **Model**

$$d\colon \mathcal{L}\times\mathcal{L} \to \mathbb{R}_0^+ \qquad (\ell_1,\ell_2) \mapsto |\ell_1 - \ell_2|$$

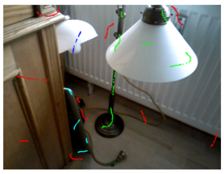
For $p > 0$, we can define the L^p **Model** as

$$d\colon \mathcal{L}\times\mathcal{L} \to \mathbb{R}_0^+ \qquad (\ell_1,\ell_2) \mapsto |\ell_1 - \ell_2|^p$$

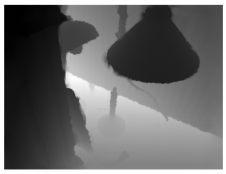
Note that the Potts model can be seen as the L^p model for $p = 0$.

In addition, we observe that the L^p model is convex iff $p \geqslant 1$.

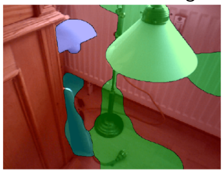
Multi-object Segmentation



Annotated RGB Image



Depth Image



RGB-Based Segmentation

RGB-D-Based Segmentation

Multilabel on Forests

If the graphical model on which we want to solve the multilabel problem is a **tree**, we can apply the *Belief Propagation* approach.

We can still solve the multilabel problem if the graphical model is a **forest**, *i.e.*, a disjoint union of trees. In this case, each tree can be optimized independently of the other trees.

One example of a forest is the lack of any pairwise potentials. In that case, each variable can be optimized independently of the other variables. This is a similar behavior to the **modular functions** in the binary case.

Since we usually use a graph model that does not form a tree (or forest), we have to study when the derived energy can be globally optimized.

NP-hardness

Multiway Cut

Given an undirected graph $G = (V, \mathcal{E}, c)$ with vertex set V, edge set $\mathcal{E}$ and weighting function $c\colon \mathcal{E} \to \mathbb{R}_0^+$, one can define the **multiway cut problem**, which generalizes the graph cut problem.

Let $s_0, \ldots, s_{k-1} \in V$ be **terminal nodes**. We call $C \subset \mathcal{E}$ a multiway cut, iff any two nodes s_i and s_j are disconnected in $(V, \mathcal{E} - C)$.

The cut value of a multiway cut is

$$\mathrm{Cut}(C) = \sum_{(i,j)\in C} c(i,j).$$

This coincides with the graph cut problem if $k = 2$ by setting $C := \mathcal{E} \cap S \times T$ if (S, T) is the cut of the graph.

NP-hardness of the Potts Model

It was shown that the multiway cut problem is NP-hard if we use $k \geqslant 3$ terminal nodes. Nonetheless, one can find a $(2 - \frac{2}{k})$ approximation.

Interestingly, every multiway cut problem can be translated into an MRF problem using the Potts model. In other words, any polynomial time algorithm of the Potts model would also solve the multiway cut problem. Hence, the Potts model is NP hard for $|\mathcal{L}| \geqslant 3$.

To see this, let $G = (V, \mathcal{E}, c)$ be an undirected graph and $K := 1 + \sum_{e\in\mathcal{E}} c(e)$ an upper bound for any multiway cut. Further let $s_0, \ldots, s_{k-1} \in V$ be the k terminal nodes. Then solving the multiway cut problem is equivalent to minimizing

$$E(x) = \sum_{i=0}^{k-1} -K[x_{s_i} = i] + \sum_{(i,j)\in\mathcal{E}} c_{i,j}[x_i \neq x_j]$$

Data Term Optimization

If we have a multi-label problem without pairwise terms, we can transform it into a graph cut problem. This is not surprising, since we could solve this problem by a mere tresholding approach.

To do this end, we take $|\mathcal{L}| - 1$ different copies of our variables. In other words, we have for each variable $i = 1, \ldots, n$ exactly $k - 1$ different nodes $v_{i,0}, \ldots, v_{i,k-2}$ and define the following capacities

$$\begin{aligned} c(s, v_{i,0}) &= f_i(0) \\ c(v_{i,\ell-1}, v_{i,\ell}) &= f_i(\ell) \qquad c(v_{i,\ell}, v_{i,\ell-1}) = \infty \qquad \text{for } \ell = 1, \ldots, k-2 \\ c(v_{i,k-2}, t) &= f_i(k-1) \end{aligned}$$

Lower Ideals of Totally Ordered Labels

This means that if $v_{i,\ell}$ is connected with the source s, also all nodes $v_{i,\ell'}$ for $\ell' < \ell$ are connected with the source s as well.

Thus, the variables $\xi_{i,\ell} := [s \text{ is connected with } v_{i,\ell}]$ have one of the following constellations:

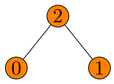
$$\begin{aligned} &\xi_i = (\xi_{i,0}, \ldots, \xi_{i,k-2}) = (0, \ldots, 0) \\ \text{or } &\xi_i = (\xi_{i,0}, \ldots, \xi_{i,k-2}) = (1, \ldots, 1, 0, \ldots, 0) \\ \text{or } &\xi_i = (\xi_{i,0}, \ldots, \xi_{i,k-2}) = (1, \ldots, 1) \end{aligned}$$

In other words, ξ_i is a representation of the lower ideal with respect to $x_i \in \mathcal{L}$ assuming that $\mathcal{L}$ is a totally ordered label set.

Note that for the path $(s, v_{i,0}, \ldots, v_{i,k-2}, t)$ there is only one transition from the source set S to the sink set T and the cost that contributes to the cut value is exactly $f_i(x_i)$.

Lower Ideal of Partially Ordered Labels

$\leq$	0	1	2
0	X		X
1		X	X
2			X



It is also possible to use the lower ideals of partially ordered sets:

$$\mathcal{I}_{\mathcal{L}} = \{0_{\leq}, 1_{\leq}, 2_{\leq}, 0_{\leq} \cup 1_{\leq}\}$$

Since the set of lower ideals contains not only join-irreducible elements, we cannot use the same approach. In fact, enforcing join-irreducibility would lead to super-modular terms.

Therefore, we will focus on totally ordered label sets $\mathcal{L}$.

Convex Prior

Linear Distance Prior

So far, we only transported the data term into the graph cut framework. This was done by introducing auxiliary nodes. This means that for neigboring pixels i and j we have the binary variables $\xi_{i,0}, \dots, \xi_{i,k-2}$ and $\xi_{j,0}, \dots, \xi_{j,k-2}$ with

$$x_i = \sum_{\ell=0}^{k-2} \xi_{i,\ell} \qquad\qquad x_j = \sum_{\ell=0}^{k-2} \xi_{j,\ell}$$

If we introduce pairwise terms between $\xi_{i,\ell}$ and $\xi_{j,\ell}$, we will add a penalty term if x_i and x_j do not agree.

In fact, we obtain the L^1 model for the multilabeling problem

$$\sum_{\ell=0}^{k-2} \xi_{i,\ell}\bar{\xi}_{j,\ell} + \xi_{j,\ell}\bar{\xi}_{i,\ell} = \sum_{\ell=0}^{k-2} [\xi_{i,\ell} \neq \xi_{j,\ell}] = |x_i - x_j|\,.$$

Quadratic Distance Prior

Also the quadratic model or L^2 model can be transformed into a binary graph cut problem by adding extra edges:

$$\sum_{\ell_1=0}^{k-2} [\xi_{i,\ell_1} \neq \xi_{j,\ell_1}] + 2 \sum_{\ell_1=0}^{k-2} \sum_{\ell_2=0}^{\ell_1-1} \xi_{i,\ell_1}\bar{\xi}_{j,\ell_2} + \xi_{j,\ell_1}\bar{\xi}_{i,\ell_2} = (x_i - x_j)^2$$

For $|x_i - x_j| \leqslant 1$, this is obviously true. Let us assume the relationship is proven for $d = x_i - x_j > 0$. For $x_i + 1$, we have to cut the edge between ξ_{i,x_i+1} and ξ_{j,x_i+1} and the d different edges between ξ_{i,x_i+1} and $\xi_{j,x_i+1-\delta}$.

Overall, the costs sum up to

$$(x_i - x_j)^2 + 1 + 2 \cdot d = (x_i - x_j)^2 + 1 + 2(x_i - x_j) = (x_i + 1 - x_j)^2$$

Convex Prior

Lemma 1. *Let us assume that we have a convex function $f : \mathbb{R} \to \mathbb{R}$ that satisfies*

$$f(0) = 0 \qquad\qquad f(-x) = f(x).$$

Then, using $f(x_i - x_j)$ as penalty for neighboring pixels $(i, j) \in \mathcal{E}$ can be globally optimized.

This can be done by using extra edges between ξ_{i,ℓ_1} and ξ_{j,ℓ_2} and assigning the following positive capacity $c_{\ell_1-\ell_2}$ to this edge:

$$c_d = \begin{cases} f(d-1) - 2f(d) + f(d+1) & \text{if } d > 0 \\ f(1) & \text{if } d = 0 \\ 0 & \text{if } d < 0 \end{cases}$$

Convex Prior

Proof. Since f is convex, we have

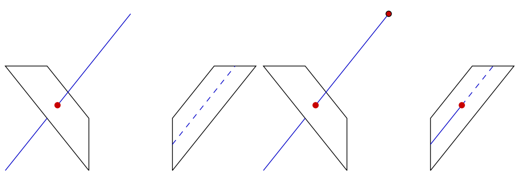
$$f(d) = f\left(\frac{1}{2}(d-1) + \frac{1}{2}(d+1)\right) \leqslant \frac{f(d-1) + f(d)}{2}.$$

Thus, $f(d-1) - 2f(d) + f(d+1) \geqslant 0$.
For the same reason we have $f(1) \geqslant 0$ and thus, c_d is always non-negative.

Without loss of generality, we can assume that $x_j = x_i - d$. The lemma is obviously true for $d = 0$ and $d = 1$. For general d, the cut is

$$\begin{aligned}\sum_{\delta=0}^{d-1} c_\delta \cdot (d - \delta) &= \sum_{\delta=0}^{d-2} c_\delta \cdot ((d-1) - \delta) + \sum_{\delta=0}^{d-1} c_d \\ &= f(d-1) + [f(d) - f(d-1)] = f(d)\end{aligned}$$

Stereo Matching

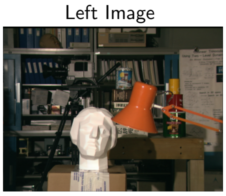


Given two images I_1 and I_2, an observed 2D point $x \in \Omega \subset \mathbb{R}^2$ of I_1 corresponds to a 3D point X that is situated on a line in $\mathbb{R}^3$.
This projective line will be observed as a line on the second image I_2.

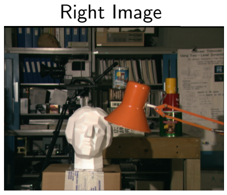
At the projective point $x' \in \Omega$ the image information should be similar to x, *i.e.*, $I_1(x) \approx I_2(x')$. This defines the data term for a depth map estimation. It is common to combine this data term with an L^1 or L^2 regularization.

Stereo Matching

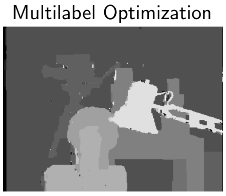
Left Image



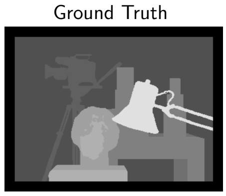
Right Image



Multilabel Optimization



Ground Truth



Multiway Cut

- Dahlhaus, Johnson, Papadimitriou, Seymour, Yannakakis, *The complexity of multiway cuts*, 1992, ACM Symp. on Theory of Comp., 241–251.

Computer Vision

- Veksler, *Efficient Graph-Based Energy Minimization Methods in Computer Vision*, 1999, PhD Thesis, Cornell University.
- Ishikawa, *Exact Optimization for Markov Random Fields with Convex Priors*, 2003, IEEE TPAMI 25(10), 1333–1336.