Fast Training of Pairwise or Higher-order CRFs

Nikos Komodakis
(University of Crete)

Introduction

Conditional Random Fields (CRFs)

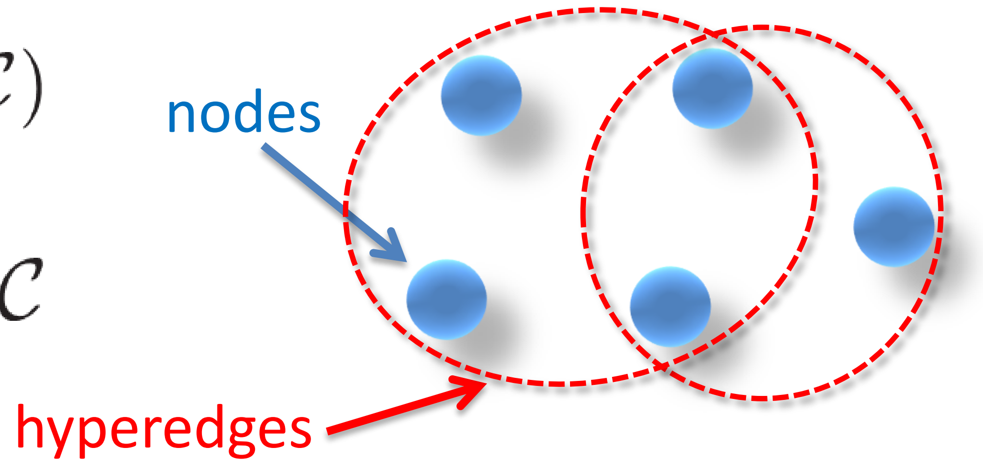
- Ubiquitous in computer vision
  - segmentation / stereo matching
    optical flow / image restoration
    image completion / object detection/localization
    ...
- and beyond
  - medical imaging, computer graphics, digital communications, physics…
- Really powerful formulation

Conditional Random Fields (CRFs)

- Key task: inference/optimization for CRFs/MRFs
- Extensive research for more than 20 years
- Lots of progress
- Many state-of-the-art methods:
  - Graph-cut based algorithms
  - Message-passing methods
  - LP relaxations
  - Dual Decomposition
  - ….

MAP inference for CRFs/MRFs

- Hypergraph $G = (\mathcal{V}, \mathcal{C})$
  - Nodes $\mathcal{V}$
  - Hyperedges/cliques $\mathcal{C}$



- High-order MRF energy minimization problem

$$\mathrm{MRF}_G(\mathbf{U}, \mathbf{H}) \equiv \min_{\mathbf{x}} \sum_{q \in \mathcal{V}} U_q(x_q) + \sum_{c \in \mathcal{C}} H_c(\mathbf{x}_c)$$

unary potential (one per node)

high-order potential (one per clique)

CRF training

- But how do we choose the CRF potentials?
- Through training
  - Parameterize potentials by **w**
  - Use training data to learn correct **w**
- Characteristic example of structured output learning [Taskar], [Tsochantaridis, Joachims]

$$f : Z \rightarrow X$$

how to determine f ?

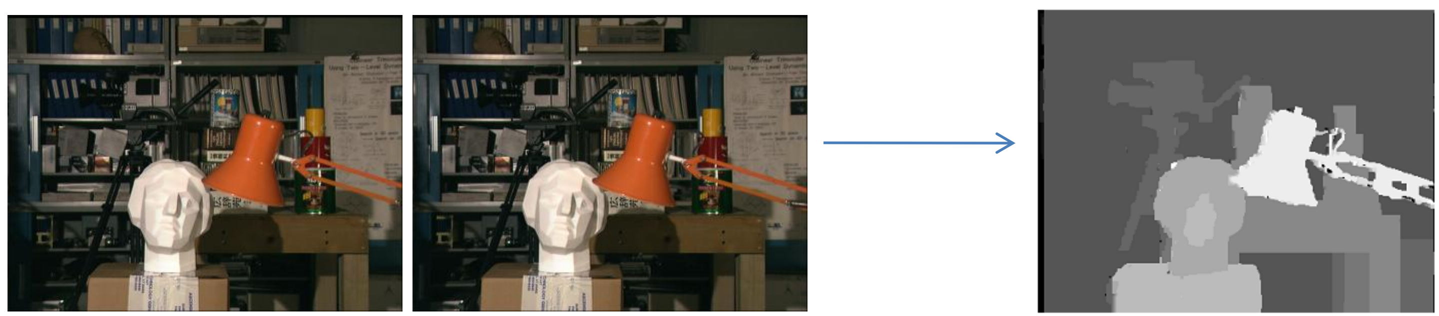
Z: can contain any kind of data

X: CRF variables (structured object)

CRF training

- Stereo matching:
  - Z: left, right image
  - X: disparity map

$f:$



Z X

$$f = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{u}, \mathbf{h})$$

parameterized by **w**

CRF training

- Denoising:
  - Z: noisy input image
  - X: denoised output image

$f:$

Z

X

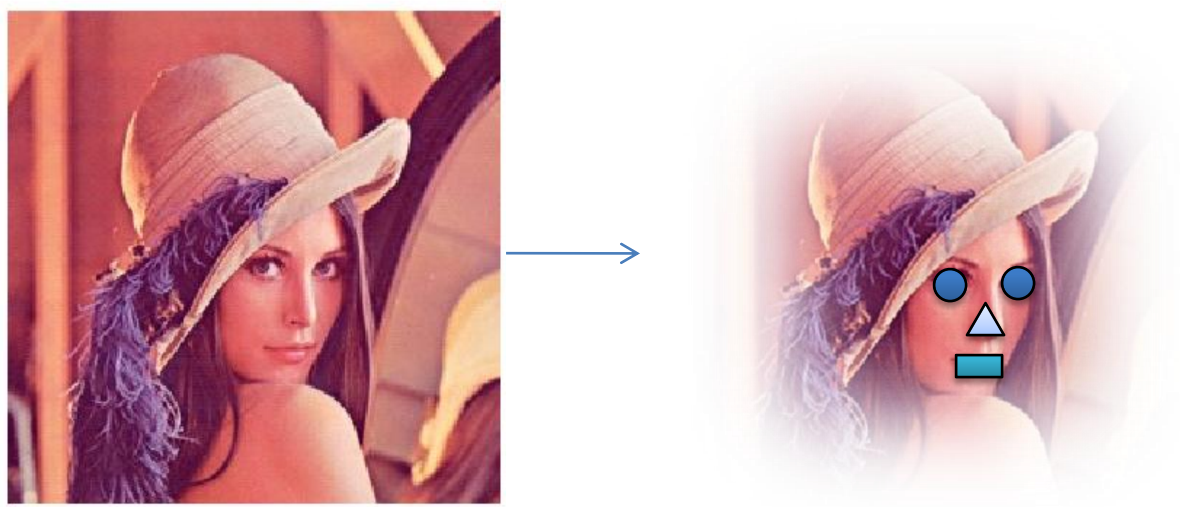
$$f = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{u}, \mathbf{h})$$

parameterized by **w**

CRF training

- Object detection:
  - Z: input image
  - X: position of object parts

$f:$



$$f = \arg\min_{\mathbf{x}} \mathrm{MRF}_G(\mathbf{x}; \mathbf{u}, \mathbf{h})$$

parameterized by **w**

CRF training

- Equally, if not more, important than MAP inference
  - Better optimize correct energy (even approximately)
  - Than optimize wrong energy exactly

- Becomes even more important as we move towards:
  - complex models
  - high-order potentials
  - lots of parameters
  - lots of training data

Contributions of this work

CRF Training via Dual Decomposition

- A very efficient max-margin learning framework for general CRFs

CRF Training via Dual Decomposition

- A very efficient max-margin learning framework for general CRFs
- Key issue: how to properly exploit CRF structure during learning?

CRF Training via Dual Decomposition

- A very efficient max-margin learning framework for general CRFs
- Key issue: how to properly exploit CRF structure during learning?
  - Existing max-margin methods:
    - use MAP inference of an **equally complex CRF** as subroutine
    - have to call subroutine **many times** during learning

CRF Training via Dual Decomposition

- A very efficient max-margin learning framework for general CRFs
- Key issue: how to properly exploit CRF structure during learning?
  - Existing max-margin methods:
    - use MAP inference of an **equally complex CRF** as subroutine
    - have to call subroutine **many times** during learning
  - Suboptimal

CRF Training via Dual Decomposition

- A very efficient max-margin learning framework for general CRFs
- Key issue: how to properly exploit CRF structure during learning?
  - Existing max-margin methods:
    - use MAP inference of an **equally complex CRF** as subroutine
    - have to call subroutine **many times** during learning
  - Suboptimal
    - computational efficiency ???
    - accuracy ???
    - theoretical properties ???

CRF Training via Dual Decomposition

- Reduces training of complex CRF to **parallel training of a series of easy-to-handle slave CRFs**

CRF Training via Dual Decomposition

- Reduces training of complex CRF to **parallel training of a series of easy-to-handle slave CRFs**
- Handles arbitrary **pairwise or higher-order** CRFs

CRF Training via Dual Decomposition

- Reduces training of complex CRF to **parallel training of a series of easy-to-handle slave CRFs**
- Handles arbitrary **pairwise or higher-order** CRFs
- Uses **very efficient** projected subgradient learning scheme

CRF Training via Dual Decomposition

- Reduces training of complex CRF to **parallel training of a series of easy-to-handle slave CRFs**
- Handles arbitrary **pairwise or higher-order** CRFs
- Uses **very efficient** projected subgradient learning scheme
- Allows hierarchy of structured prediction learning algorithms of **increasing accuracy**

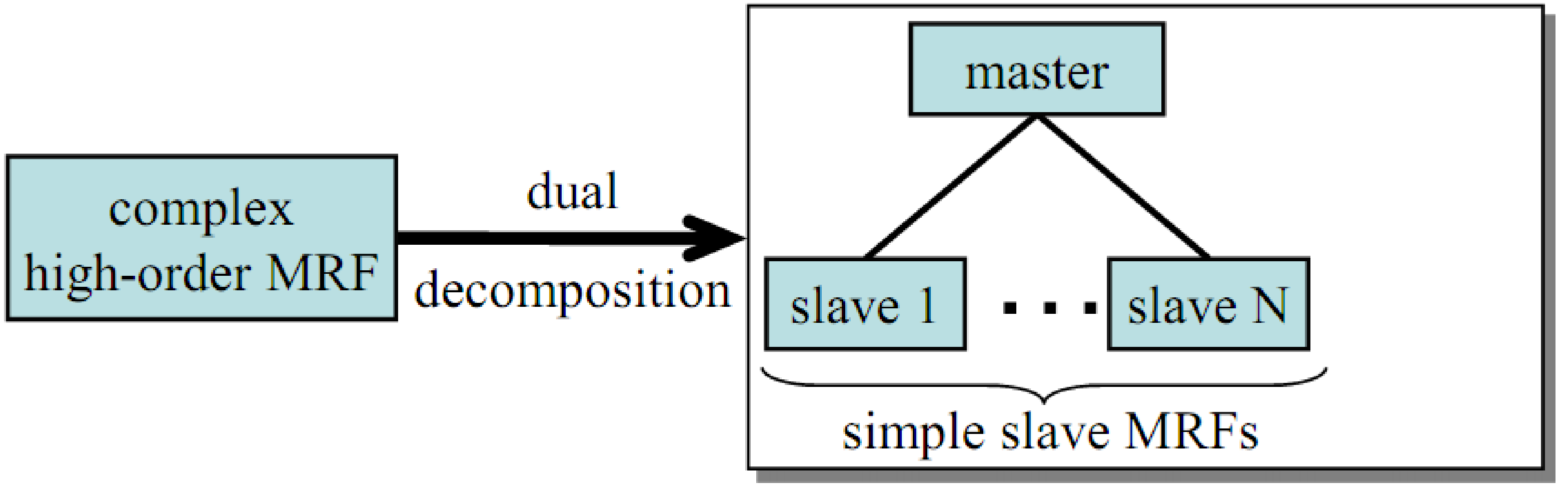
CRF Training via Dual Decomposition

- Reduces training of complex CRF to **parallel training of a series of easy-to-handle slave CRFs**
- Handles arbitrary **pairwise or higher-order** CRFs
- Uses **very efficient** projected subgradient learning scheme
- Allows hierarchy of structured prediction learning algorithms of **increasing accuracy**
- Extremely **flexible and adaptable**
  - Easily adjusted to fully exploit additional structure in any class of CRFs (no matter if they contain very high order cliques)

Dual Decomposition for CRF MAP Inference (brief review)

MRF Optimization via Dual Decomposition

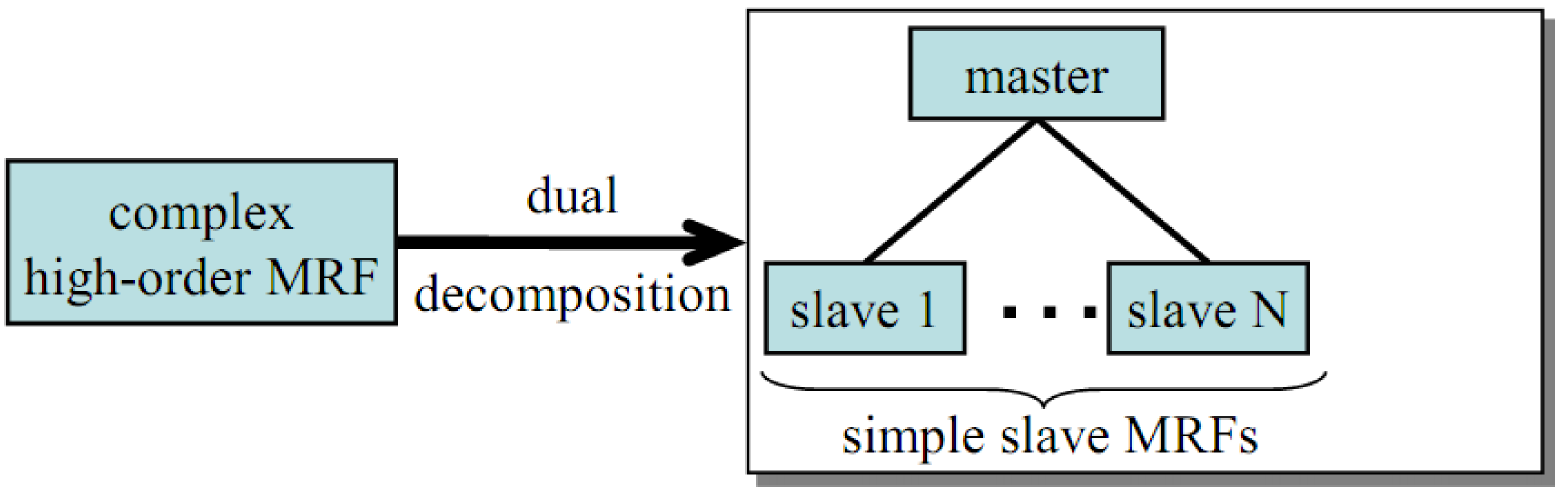
- Very general framework for MAP inference [Komodakis et al. ICCV07, PAMI11]



- Master = coordinator (has global view)
  Slaves = subproblems (have only local view)

MRF Optimization via Dual Decomposition

- Very general framework for MAP inference [Komodakis et al. ICCV07, PAMI11]

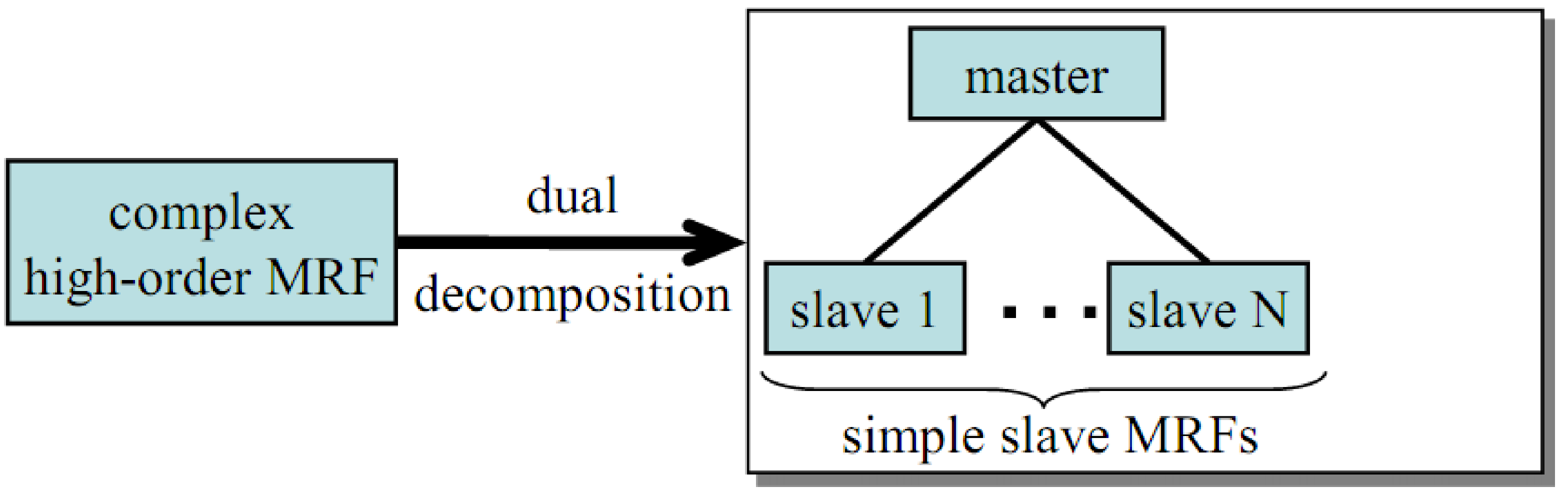


- Master $= \mathrm{MRF}_G(\mathbf{u}, \mathbf{h}) \longleftarrow$ (MAP-MRF on hypergraph *G*)

$$= \min \mathrm{MRF}_G(\mathbf{x}; \mathbf{u}, \mathbf{h}) := \sum_{p \in \mathcal{V}} u_p(x_p) + \sum_{c \in \mathcal{C}} h_c(\mathbf{x}_c)$$

MRF Optimization via Dual Decomposition

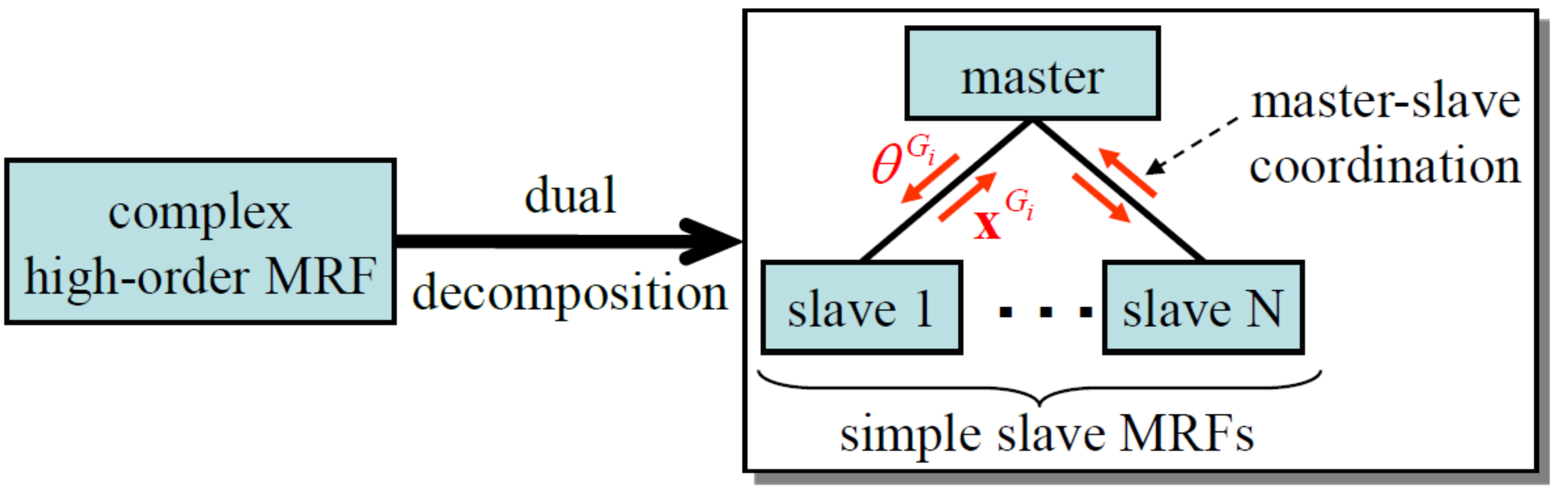
- Very general framework for MAP inference [Komodakis et al. ICCV07, PAMI11]



- Set of slaves = $\{\mathrm{MRF}_{G_i}(\boldsymbol{\theta}^i, \mathbf{h})\}$
  (MRFs on sub-hypergraphs G_i whose union covers G)
- Many other choices possible as well

MRF Optimization via Dual Decomposition

- Very general framework for MAP inference [Komodakis et al. ICCV07, PAMI11]



- Optimization proceeds in an iterative fashion via **master-slave coordination**

MRF Optimization via Dual Decomposition

Set of slave MRFs

$$\{\mathrm{MRF}_{G_i}(\boldsymbol{\theta}^i, \mathbf{h})\}$$

convex dual relaxation

$$\mathrm{DUAL}_{\{G_i\}}(\mathbf{u}, \mathbf{h}) = \max_{\{\boldsymbol{\theta}^i\}} \sum\nolimits_i \mathrm{MRF}_{G_i}(\boldsymbol{\theta}^i, \mathbf{h})$$

$$\text{s.t.} \sum_{i \in \mathcal{I}_p} \theta_p^i(\cdot) = u_p(\cdot)$$

For each choice of slaves, master solves (possibly different) dual relaxation

- Sum of slave energies = lower bound on MRF optimum
- Dual relaxation = maximum such bound

MRF Optimization via Dual Decomposition

Set of slave MRFs

$\{\mathrm{MRF}_{G_i}(\boldsymbol{\theta}^i, \mathbf{h})\}$

convex dual relaxation

$$\mathrm{DUAL}_{\{G_i\}}(\mathbf{u}, \mathbf{h}) = \max_{\{\boldsymbol{\theta}^i\}} \sum_i \mathrm{MRF}_{G_i}(\boldsymbol{\theta}^i, \mathbf{h})$$
$$\text{s.t.} \sum_{i \in \mathcal{I}_p} \theta_p^i(\cdot) = u_p(\cdot)$$

Choosing more difficult slaves $\Rightarrow$ tighter lower bounds

$\Rightarrow$ tighter dual relaxations

CRF Training via Dual Decomposition

Max-margin Learning via Dual Decomposition

- **Input:**
  - $\{\bar{\mathbf{z}}^k, \bar{\mathbf{x}}^k\}_{k=1}^K$ (training set of K samples)
  - k-th sample: CRF on $G^k = (\mathcal{V}^k, \mathcal{C}^k)$
  - Feature vectors: $g_p(\cdot, \cdot)$, $g_c(\cdot, \cdot)$

$$u_p^k(x_p) = \mathbf{w}^T g_p(x_p, \bar{\mathbf{z}}^k), \;\; h_c^k(\mathbf{x}_c) = \mathbf{w}^T g_c(\mathbf{x}_c, \bar{\mathbf{z}}^k)$$

- **Constraints:**

$$\mathrm{MRF}_{G^k}(\bar{\mathbf{x}}^k; \mathbf{u}^k, \mathbf{h}^k) \leq \mathrm{MRF}_{G^k}(\mathbf{x}; \mathbf{u}^k, \mathbf{h}^k) - \Delta(\mathbf{x}, \bar{\mathbf{x}}^k)$$

$\Delta(\mathbf{x}, \mathbf{x}')$ = dissimilarity function, ($\Delta(\mathbf{x}, \mathbf{x}) = 0$)

Max-margin Learning via Dual Decomposition

- **Input:**
  - $\{\bar{\mathbf{z}}^k, \bar{\mathbf{x}}^k\}_{k=1}^K$ (training set of K samples)
  - k-th sample: CRF on $G^k = (\mathcal{V}^k, \mathcal{C}^k)$
  - Feature vectors: $g_p(\cdot, \cdot)$, $g_c(\cdot, \cdot)$

$$u_p^k(x_p) = \mathbf{w}^T g_p(x_p, \bar{\mathbf{z}}^k), \;\; h_c^k(\mathbf{x}_c) = \mathbf{w}^T g_c(\mathbf{x}_c, \bar{\mathbf{z}}^k)$$

- **Constraints:**

$$\mathrm{MRF}_{G^k}(\bar{\mathbf{x}}^k; \mathbf{u}^k, \mathbf{h}^k) \leq \mathrm{MRF}_{G^k}(\mathbf{x}; \mathbf{u}^k, \mathbf{h}^k) - \Delta(\mathbf{x}, \bar{\mathbf{x}}^k) + \xi_k$$

$\Delta(\mathbf{x}, \mathbf{x}')$ = dissimilarity function, ($\Delta(\mathbf{x}, \mathbf{x}) = 0$)

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$\min_{\mathbf{w}} \mu R(\mathbf{w}) + \sum_{k=1}^{K} \xi_k$$

$$\xi_k = \mathrm{MRF}_{G^k}(\bar{\mathbf{x}}^k; \mathbf{u}^k, \mathbf{h}^k) - \min_{\mathbf{x}} \left(\mathrm{MRF}_{G^k}(\mathbf{x}; \mathbf{u}^k, \mathbf{h}^k) - \Delta(\mathbf{x}, \bar{\mathbf{x}}^k) \right)$$

$$\Delta(\mathbf{x}, \bar{\mathbf{x}}^k) = \sum_{p \in \mathcal{V}^k} \delta_p(x_p, \bar{x}_p^k) + \sum_{c \in \mathcal{C}^k} \delta_c(\mathbf{x}_c, \bar{\mathbf{x}}_c^k)$$

$$\bar{u}_p^k(\cdot) = u_p^k(\cdot) - \delta_p(\cdot, \bar{x}_p^k)$$

$$\bar{h}_c^k(\cdot) = h_c^k(\cdot) - \delta_c(\cdot, \bar{\mathbf{x}}_c^k)$$

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$\min_{\mathbf{w}} \mu R(\mathbf{w}) + \sum_{k=1}^{K} \xi_k$$

ξ_k ⇒

$$L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w}) \equiv$$
$$\equiv \mathrm{MRF}_{G^k}(\bar{\mathbf{x}}^k; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k) - \min_{\mathbf{x}} \mathrm{MRF}_{G^k}(\mathbf{x}; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k)$$

$$\Delta(\mathbf{x}, \bar{\mathbf{x}}^k) = \sum_{p \in \mathcal{V}^k} \delta_p(x_p, \bar{x}_p^k) + \sum_{c \in \mathcal{C}^k} \delta_c(\mathbf{x}_c, \bar{\mathbf{x}}_c^k)$$

$$\bar{u}_p^k(\cdot) = u_p^k(\cdot) - \delta_p(\cdot, \bar{x}_p^k)$$

$$\bar{h}_c^k(\cdot) = h_c^k(\cdot) - \delta_c(\cdot, \bar{\mathbf{x}}_c^k)$$

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$\min_{\mathbf{w}} \mu R(\mathbf{w}) + \sum_{k=1}^{K} L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w})$$

$$L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w}) \equiv$$
$$\equiv \mathrm{MRF}_{G^k}(\bar{\mathbf{x}}^k; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k) - \min_{\mathbf{x}} \mathrm{MRF}_{G^k}(\mathbf{x}; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k)$$

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$\min_{\mathbf{w}} \mu R(\mathbf{w}) + \sum_{k=1}^{K} L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w})$$

$$L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w}) \equiv$$
$$\equiv \mathrm{MRF}_{G^k}(\bar{\mathbf{x}}^k; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k) - \min_{\mathbf{x}} \mathrm{MRF}_{G^k}(\mathbf{x}; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k)$$

Problem

Learning objective intractable due to this term

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$\min_{\mathbf{w}} \mu R(\mathbf{w}) + \sum_{k=1}^{K} L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w})$$

$$L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w}) \equiv$$
$$\equiv \mathrm{MRF}_{G^k}(\bar{\mathbf{x}}^k; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k) - \min_{\mathbf{x}} \mathrm{MRF}_{G^k}(\mathbf{x}; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k)$$

Solution: approximate it with dual relaxation from decomposition $\{G_i^k = (\mathcal{V}_i^k, \mathcal{C}_i^k)\}$

$$\min_{\mathbf{x}} \mathrm{MRF}_{G^k}(\mathbf{x}; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k) \approx \mathrm{DUAL}_{\{G_i^k\}}(\bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k)$$

Max-margin Learning via Dual Decomposition

$$\min_{\mathbf{x}} \mathrm{MRF}_{G^k}(\mathbf{x}; \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k) \approx \mathrm{DUAL}_{\{G_i^k\}}(\bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k)$$

$$\mathrm{DUAL}_{\{G_i\}}(\mathbf{u}, \mathbf{h}) = \max_{\{\boldsymbol{\theta}^i\}} \sum_i \mathrm{MRF}_{G_i}(\boldsymbol{\theta}^i, \mathbf{h})$$

$$\text{s.t.} \sum_{i \in \mathcal{I}_p} \theta_p^i(\cdot) = u_p(\cdot)$$

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$
\min_{\mathbf{w},\{\boldsymbol{\theta}^{(i,k)}\}} \mu R(\mathbf{w}) + \sum_k \sum_i L_{G_i^k}(\bar{\mathbf{x}}^k, \boldsymbol{\theta}^{(i,k)}, \bar{\mathbf{h}}^k; \mathbf{w})
$$

$$
\text{s.t.} \quad \sum_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot) \quad .
$$

now

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$\min_{\mathbf{w},\{\boldsymbol{\theta}^{(i,k)}\}} \mu R(\mathbf{w}) + \sum_{k} \sum_{i} L_{G_i^k}(\bar{\mathbf{x}}^k, \boldsymbol{\theta}^{(i,k)}, \bar{\mathbf{h}}^k; \mathbf{w})$$

$$\text{s.t.} \quad \sum\nolimits_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot) \quad .$$

now

$$\min_{\mathbf{w}} \mu R(\mathbf{w}) + \sum_{k=1}^{K} L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w})$$

before

Max-margin Learning via Dual Decomposition

- Regularized hinge loss functional:

$$\min_{\mathbf{w},\{\boldsymbol{\theta}^{(i,k)}\}} \mu R(\mathbf{w}) + \sum_k \sum_i L_{G_i^k}(\bar{\mathbf{x}}^k, \boldsymbol{\theta}^{(i,k)}, \bar{\mathbf{h}}^k; \mathbf{w})$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot) \quad .$$

now

$$\min_{\mathbf{w}} \mu R(\mathbf{w}) + \sum_{k=1}^{K} L_{G^k}(\bar{\mathbf{x}}^k, \bar{\mathbf{u}}^k, \bar{\mathbf{h}}^k; \mathbf{w})$$

before

Training of complex CRF was decomposed to parallel training of easy-to-handle slave CRFs !!!

Max-margin Learning via Dual Decomposition

- **Global optimum via projected subgradient learning algorithm:**
  - **Input:**
    - Training samples: $\{\bar{\mathbf{z}}^k, \bar{\mathbf{x}}^k\}_{k=1}^K$
    - Hypergraphs: $\{G^k = (\mathcal{V}^k, \mathcal{C}^k)\}_{k=1}^K$
    - Feature vectors: $\{g_p(\cdot,\cdot)\}, \{g_c(\cdot,\cdot)\}$

Max-margin Learning via Dual Decomposition

- **Global optimum via projected subgradient learning algorithm:**

$\forall k,$ choose decomposition $\{G_i^k = (\mathcal{V}_i^k, \mathcal{C}_i^k)\}$ of hypergraph G^k

$\forall k, i,$ initialize $\boldsymbol{\theta}^{(i,k)}$ so as to satisfy $\sum_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot)$

Max-margin Learning via Dual Decomposition

- **Global optimum via projected subgradient learning algorithm:**

$\forall k,$ choose decomposition $\{G_i^k = (\mathcal{V}_i^k, \mathcal{C}_i^k)\}$ of hypergraph G^k

$\forall k, i,$ initialize $\boldsymbol{\theta}^{(i,k)}$ so as to satisfy $\sum_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot)$

repeat

until convergence

Max-margin Learning via Dual Decomposition

- **Global optimum via projected subgradient learning algorithm:**

$\forall k,$ choose decomposition $\{G_i^k = (\mathcal{V}_i^k, \mathcal{C}_i^k)\}$ of hypergraph G^k

$\forall k, i,$ initialize $\boldsymbol{\theta}^{(i,k)}$ so as to satisfy $\sum_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot)$

repeat

// optimize slave MRFs

$\forall k, i,$ compute minimizer $\hat{\mathbf{x}}^{(i,k)}$ of slave $\mathrm{MRF}_{G_i^k}(\boldsymbol{\theta}^{(i,k)}, \hat{\mathbf{h}}^k)$

until convergence

Max-margin Learning via Dual Decomposition

- **Global optimum via projected subgradient learning algorithm:**

$\forall k,$ choose decomposition $\{G_i^k = (\mathcal{V}_i^k, \mathcal{C}_i^k)\}$ of hypergraph G^k

$\forall k, i,$ initialize $\boldsymbol{\theta}^{(i,k)}$ so as to satisfy $\sum_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot)$

repeat

// optimize slave MRFs

$\forall k, i,$ compute minimizer $\hat{\mathbf{x}}^{(i,k)}$ of slave $\text{MRF}_{G_i^k}(\boldsymbol{\theta}^{(i,k)}, \hat{\mathbf{h}}^k)$

// update $\mathbf{w}$

$\mathbf{w} \leftarrow \mathbf{w} - \alpha_t \cdot d\mathbf{w}$ ← fully specified from $\left\{\hat{x}^{(i,k)}\right\}$

until convergence

Max-margin Learning via Dual Decomposition

- **Global optimum via projected subgradient learning algorithm:**

$\forall k,$ choose decomposition $\{G_i^k = (\mathcal{V}_i^k, \mathcal{C}_i^k)\}$ of hypergraph G^k

$\forall k, i,$ initialize $\boldsymbol{\theta}^{(i,k)}$ so as to satisfy $\sum_{i \in \mathcal{I}_p^k} \theta_p^{(i,k)}(\cdot) = \bar{u}_p^k(\cdot)$

repeat

// optimize slave MRFs

$\forall k, i,$ compute minimizer $\hat{\mathbf{x}}^{(i,k)}$ of slave $\text{MRF}_{G_i^k}(\boldsymbol{\theta}^{(i,k)}, \hat{\mathbf{h}}^k)$

// update $\mathbf{w}$

$\mathbf{w} \leftarrow \mathbf{w} - \alpha_t \cdot d\mathbf{w}$ ← fully specified from $\left\{\hat{x}^{(i,k)}\right\}$

// update $\boldsymbol{\theta}^{(i,k)}$

$$\boldsymbol{\theta}^{(i,k)}(\cdot) \mathrel{+}= \alpha_t \cdot \left(\left[\hat{x}_p^{(i,k)} = \cdot\right] - \frac{\sum_{j \in \mathcal{I}_p^k} \left[\hat{x}_p^{(j,k)} = \cdot\right]}{\mathcal{I}_p^k} \right)$$

until convergence

Max-margin Learning via Dual Decomposition

- **Incremental subgradient** version:
  - Same as before but considers subset of slaves per iteration
  - Subset chosen
    - deterministically or
    - randomly (**stochastic subgradient**)
  - Further improves computational efficiency
  - Same optimality guarantees & theoretical properties

Max-margin Learning via Dual Decomposition

- Resulting learning scheme:
  - ✓ Very efficient and very flexible
  - ✓ Requires from the user only to provide an optimizer for the slave MRFs
  - ✓ Slave problems freely chosen by the user
  - ✓ Easily adaptable to further exploit special structure of any class of CRFs

Choice of decompositions $\{G_i^k\}$

$\mathcal{F}_0$ = true loss (intractable)

$\mathcal{F}_{\{G_i^k\}}$ = loss from decomposition $\{G_i^k\}$

- $\mathcal{F}_0 \leq \mathcal{F}_{\{G_i^k\}}$

  (upper bound property)

- $\{G_i^k\} < \{\tilde{G}_j^k\} \Longrightarrow \mathcal{F}_0 \leq \mathcal{F}_{\{\tilde{G}_j^k\}} < \mathcal{F}_{\{G_i^k\}}$

  (hierarchy of learning algorithms)

Choice of decompositions $\{G_i^k\}$

- $G^k_{\text{single}} = \{G^k_c\}$ denotes following decomposition:
  - One slave per clique $c \in \mathcal{C}$
  - Corresponding sub-hypergraph $G^k_c = (\mathcal{V}^k_c, \mathcal{C}^k_c)$
    $\mathcal{V}^k_c = \{p | p \in c\}$, $\mathcal{C}^k_c = \{c\}$
- Resulting slaves often easy (or even trivial) to solve even if global problem is complex and NP-hard
  - leads to widely applicable learning algorithm
- Corresponding dual relaxation is an LP
  - Generalizes well known LP relaxation for pairwise MRFs (at the core of most state-of-the-art methods)

Choice of decompositions $\{G_i^k\}$

- But we can do better if CRFs have special structure…
- Structure means:
  - More efficient optimizer for slaves **(speed)**
  - Optimizer that handles more complex slaves **(accuracy)**

  (Almost all known examples fall in one of above two cases)

- We adapt decomposition to problem at hand to exploit its structure

Choice of decompositions $\{G_i^k\}$

- But we can do better if CRFs have special structure…

- E.g., **pattern-based** high-order potentials (for a clique c) [Komodakis & Paragios CVPR09]

$$H_c(\mathbf{x}) = \begin{cases} \psi_c(\mathbf{x}) & \text{if } \mathbf{x} \in \mathcal{P} \\ \psi_c^{\max} & \text{otherwise} \end{cases}$$

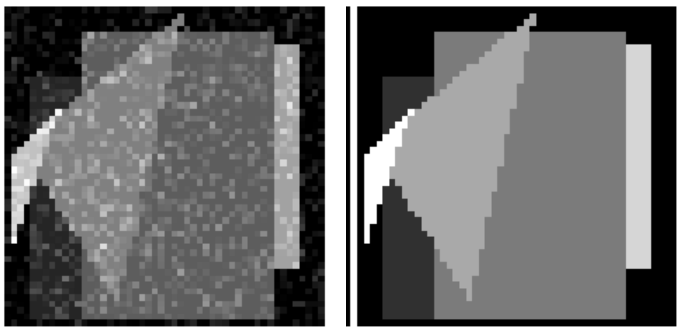
  $\mathcal{P}$ subset of $\mathcal{L}^{|c|}$ (its vectors called **patterns**)

- We only assume:
  - Set $\mathcal{P}$ is sparse
  - It holds $\psi_c(\mathbf{x}) \leq \psi_c^{\max}, \ \forall \mathbf{x} \in \mathcal{P}$
  - No other restriction

Experimental results

Image denoising

- Piecewise constant images

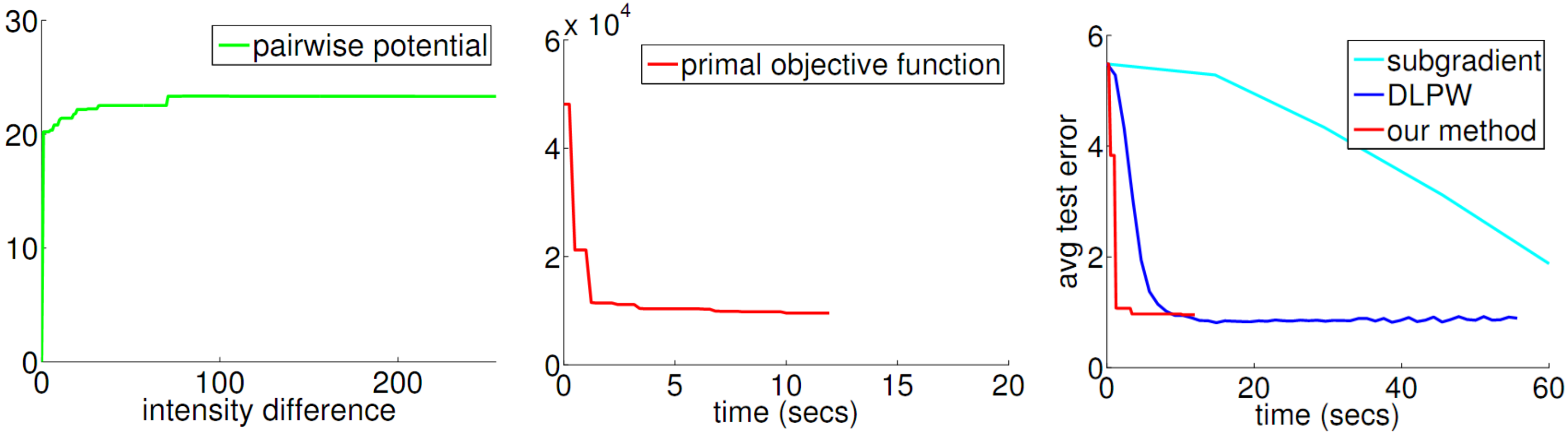


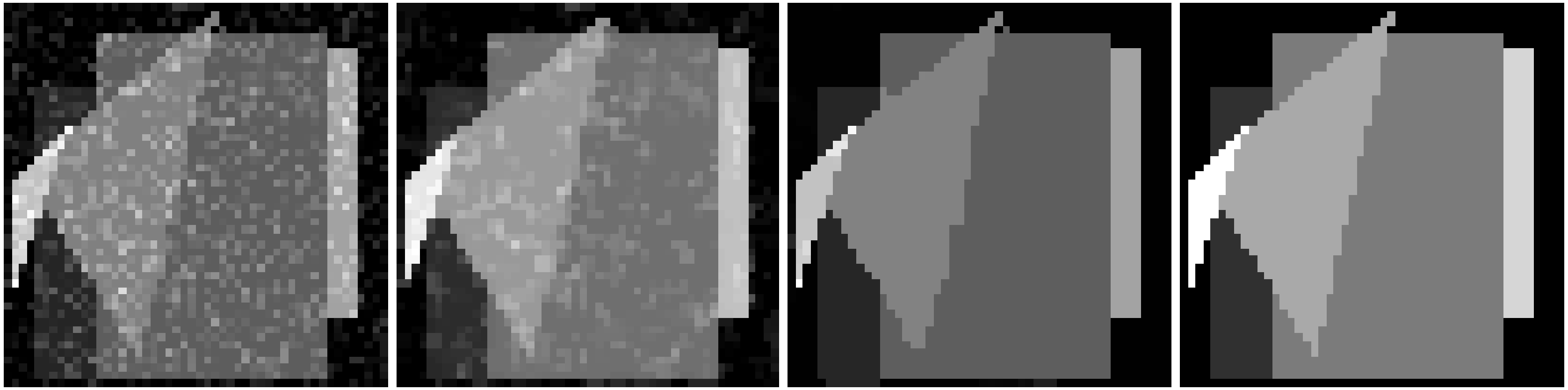
Z X

- Potentials: $u_p^k(x_p) = |x_p - z_p|$ $h_{pq}^k(x_p, x_q) = V(|x_p - x_q|)$

- Goal: learn pairwise potential $V(\cdot)$

Image denoising



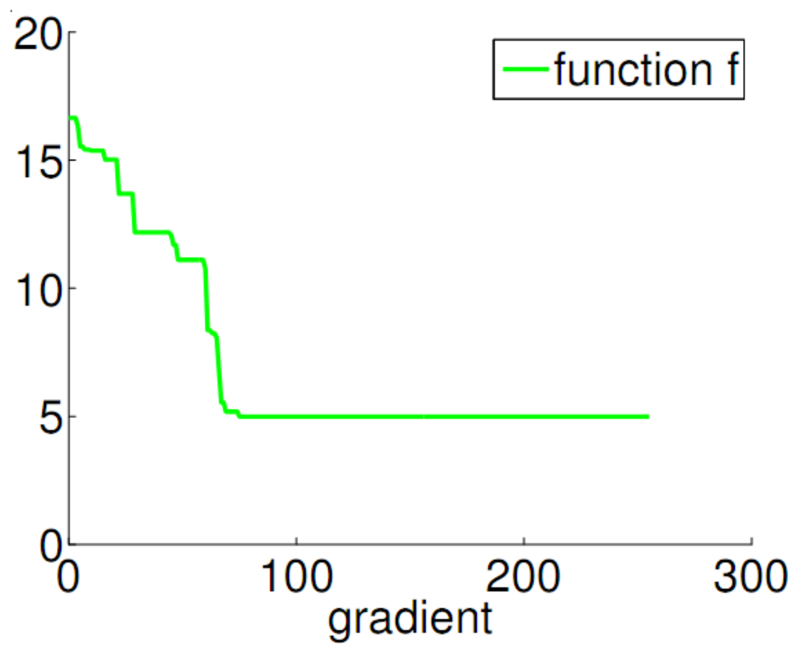


Stereo matching

- Potentials: $u_p^k(x_p) = \left| I^{left}(p) - I^{right}(p - x_p) \right|$

$$h_{pq}^k(x_p, x_q) = f\left(\left|\nabla I^{left}(p)\right|\right)\left[x_p \neq x_q\right]$$

- Goal: learn function $f(\cdot)$ for gradient-modulated Potts model

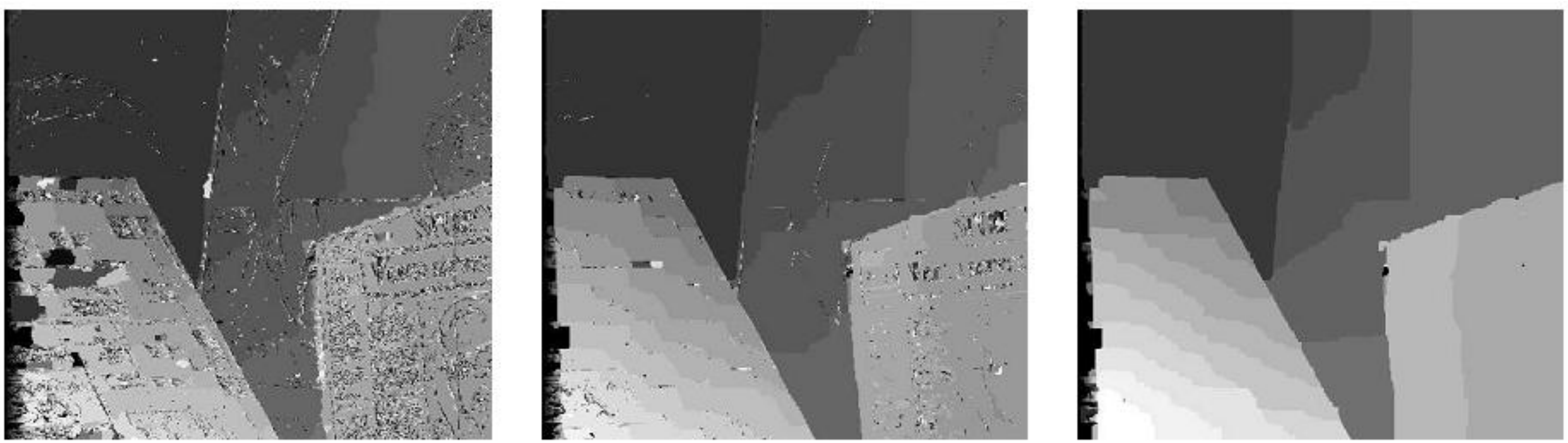


Stereo matching

- Potentials: $u_p^k\left(x_p\right)=\left|I^{left}\left(p\right)-I^{right}\left(p-x_p\right)\right|$

$$h_{pq}^k\left(x_p,x_q\right)=f\left(\left|\nabla I^{left}(p)\right|\right)\left[x_p\neq x_q\right]$$

- Goal: learn function $f(\cdot)$ for gradient-modulated Potts model

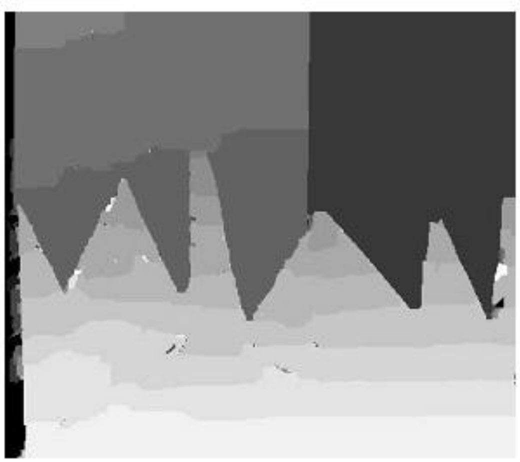


"Venus" disparity using $f(\cdot)$ as estimated at different iterations of learning algorithm

Stereo matching

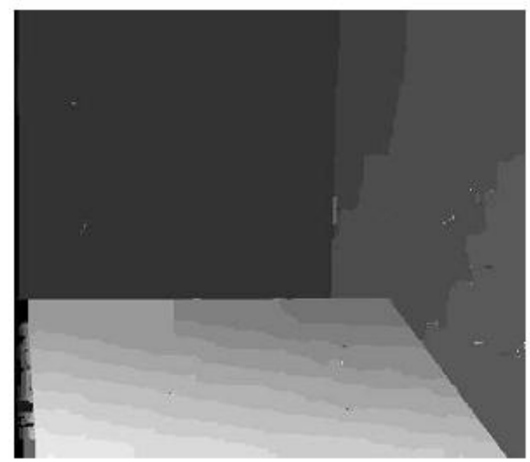
- Potentials:
$$u_p^k\left(x_p\right)=\left|I^{left}\left(p\right)-I^{right}\left(p-x_p\right)\right|$$
$$h_{pq}^k\left(x_p,x_q\right)=f\left(\left|\nabla I^{left}(p)\right|\right)\left[x_p\neq x_q\right]$$

- Goal: learn function $f(\cdot)$ for gradient-modulated Potts model



Sawtooth
4.9%

Poster
3.7%

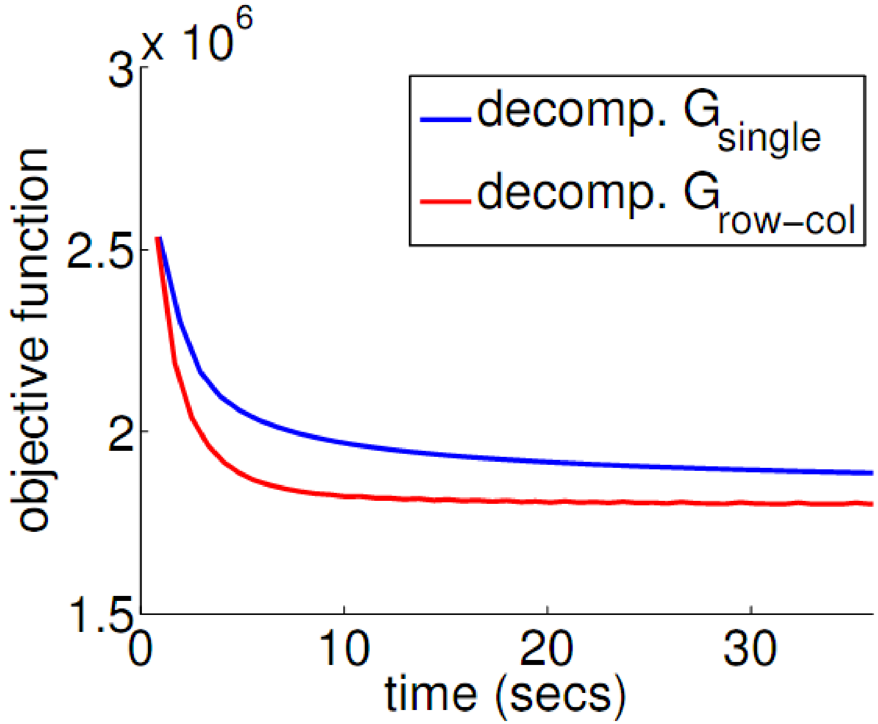


Bull
2.8%

Stereo matching

- Potentials: $$u_p^k\left(x_p\right)=\left|I^{left}\left(p\right)-I^{right}\left(p-x_p\right)\right|$$
$$h_{pq}^k\left(x_p,x_q\right)=f\left(\left|\nabla I^{left}(p)\right|\right)\left[x_p\neq x_q\right]$$

- Goal: learn function $f(\cdot)$ for gradient-modulated Potts model



High-order P^n Potts model

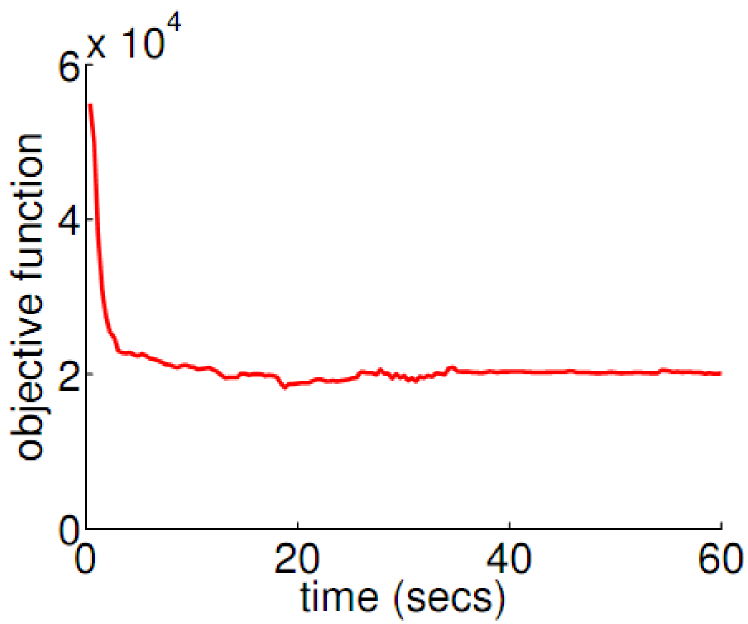
Goal: learn high order CRF with potentials given by

$$h_c(\mathbf{x}) = \begin{cases} \beta_l^c & \text{if } x_p = l,\ \forall p \in c \\ \beta_{\max}^c & \text{otherwise} \end{cases},$$

[Kohli et al. CVPR07]

$$\beta_l^c = \mathbf{w}_l \cdot z_l^c$$

Cost for optimizing slave CRF: O(|L|) ⇨ Fast training



- 100 training samples
- 50x50 grid
- clique size 3x3
- 5 labels (|L|=5)

Clustering

- Goal: distance learning for clustering [ICCV'11]
  - Novel discriminative formulation
  - In this case cliques are of very high order: contain all variables
  - On top of that, there exist unobserved (latent) variables during training
  - Significant extension: dual decomposition for training **high-order CRFs** with **latent variables**