Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction

Ľúbor Ladický, Paul Sturgess, Christopher Russell, Sunando Sengupta, Yalin Bastanlar, William Clocksin, Philip H.S. Torr

Oxford Brookes University

http://cms.brookes.ac.uk/research/visiongroup/
Joint Object Class Segmentation and Dense Stereo Reconstruction

Black Box Solver

Left Camera Image → Black Box Solver → Object Class Segmentation

Right Camera Image → Black Box Solver → Dense Stereo Reconstruction
Joint Object Class Segmentation and Dense Stereo Reconstruction

Objective: Joint Estimation

Left Camera Image

Right Camera Image

Black Box Solver

Object Class Segmentation

Dense Stereo Reconstruction
Dense Stereo Reconstruction

- For each pixel assigns a disparity label: \( y \)
- Disparities from the discrete set \( \{0, 1, \ldots, D\} \)

Left Camera Image  
Right Camera Image  
Dense Stereo Result
Dense Stereo Reconstruction

\[ E^D(y) = \sum_{i \in \mathcal{V}} \psi^D_i(y_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi^D_{ij}(y_i, y_j) \]

 Unary Potential

\[ \psi^D_i(0) = d(P(x_i, y_i), P(x_i + 0, y_i)) \]

Unary Cost dependent on the similarity of patches, e.g. cross correlation
Dense Stereo Reconstruction

\[ E_D(y) = \sum_{i \in \mathcal{V}} \psi_i^D(y_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^D(y_i, y_j) \]

Unary Potential

\[ \psi_i^D(5) = d(P(x_i, y_i), P(x_i + 5, y_i)) \]

Unary Cost dependent on the similarity of patches, e.g. cross correlation

Disparity = 5
Dense Stereo Reconstruction

\[ E^D(y) = \sum_{i \in \mathcal{V}} \psi_i^D(y_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^D(y_i, y_j) \]

Unary Potential

\[ \psi_i^D(10) = d \left( P(x_i, y_i), P(x_i + 10, y_i) \right) \]

Unary Cost dependent on the similarity of patches, e.g. cross correlation
Dense Stereo Reconstruction

\[ E^D(y) = \sum_{i \in \mathcal{V}} \psi^D_i(y_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi^D_{ij}(y_i, y_j) \]

Unary Potential

\[ \psi^D_i(15) = d(P(x_i, y_i), P(x_i + 15, y_i)) \]

Unary Cost dependent on the similarity of patches, e.g. cross correlation

Disparity = 15
Dense Stereo Reconstruction

\[ E^D(y) = \sum_{i \in \mathcal{V}} \psi_i^D(y_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^D(y_i, y_j) \]

- Encourages label consistency in adjacent pixels
- Cost based on the distance of labels

Linear Truncated

Quadratic Truncated
Dense Stereo Reconstruction

- Graph-Cut based Range-move inference
  (Kumar et al. NIPS09, Veksler et al. CVPR09)
Dense Stereo Reconstruction

- Graph-Cut based Range-move inference
  (Kumar et al. NIPS09, Veksler et al. CVPR09)

Original Image  Initial Solution  After 1\textsuperscript{st} expansion
Dense Stereo Reconstruction

- Graph-Cut based Range-move inference
  (Kumar et al. NIPS09, Veksler et al. CVPR09)

Original Image

Initial Solution

After 1\textsuperscript{st} expansion

After 2\textsuperscript{nd} expansion
Dense Stereo Reconstruction

- Graph-Cut based Range-move inference
  (Kumar et al. NIPS09, Veksler et al. CVPR09)
Dense Stereo Reconstruction

- Graph-Cut based Range-move inference
  (Kumar et al. NIPS09, Veksler et al. CVPR09)
Dense Stereo Reconstruction

Does not work for Road Scenes!

Original Image

Dense Stereo Reconstruction
Dense Stereo Reconstruction

Does not work for Road Scenes!

Different brightness in cameras

Patches can be matched to any other patch for flat surfaces
Dense Stereo Reconstruction

Does not work for Road Scenes!

Different brightness in cameras

Patches can be matched to any other patch for flat surfaces

Could object recognition for road scenes help?

Recognition of road scenes is relatively easy (Sturgess et al., BMVC09)
Object Class Segmentation

- Aims to assign a class label for each pixel of an image
- Classifier trained on the training set
- Evaluated on never seen test images
Object Class Segmentation

\[ E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{(i, j) \in E} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{L}} \psi_c(x_c) \]

- Likelihood of a pixel taking a label

(Shotton et al. ECCV06, He et al, CVPR04, Ladický et al. ICCV 09)
Object Class Segmentation

\[ E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{(i,j) \in E} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{L}} \psi_c(x_c) \]

- Contrast sensitive Potts model
- Encourages label consistency in adjacent pixels
Object Class Segmentation

\[ E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{(i, j) \in E} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{C}} \psi_c(x_c) \]

- Encouraging consistency in superpixels (Kohli et al. CVPR08)
- Merging information at different scales (Ladický et al. ICCV09)
Object Class Segmentation

- Graph-Cut based $\alpha$-Expansion inference (Boykov et al. ICCV99)

Original Image

Initial solution

grass
Object Class Segmentation

- Graph-Cut based $\alpha$-Expansion inference
  (Boykov et al. ICCV99)

Original Image
Initial solution
Building expansion
Object Class Segmentation

- Graph-Cut based $\alpha$-Expansion inference (Boykov et al. ICCV99)

Original Image

Initial solution

Building expansion

Sky expansion
Object Class Segmentation

- Graph-Cut based $\alpha$-Expansion inference (Boykov et al. ICCV99)

Original Image

Initial solution

Sky expansion

Tree expansion

Building expansion
Object Class Segmentation

- Graph-Cut based $\alpha$-Expansion inference
  (Boykov et al. ICCV99)

![Original Image](image1)

**Initial solution**

- grass

**Building expansion**

- building

**Sky expansion**

- sky

**Tree expansion**

- tree

**Final Solution**

- aeroplane
- building
- grass
Object class and 3D location are mutually informative.

- Sky always in infinity (disparity = 0)
Object class and 3D location are mutually informative
- Sky always in infinity (disparity = 0)
- Cars, buildings & pedestrians have their typical height
Object class and 3D location are mutually informative

- Sky always in infinity (disparity = 0)
- Cars, buses & pedestrians have their typical height
- Road and pavement on the ground plane
Object class and 3D location are mutually informative

- Sky always in infinity (disparity = 0)
- Cars, buses & pedestrians have their typical height
- Road and pavement on the ground plane
- Buildings and pavement on the sides
Object class and 3D location are mutually informative
- Sky always in infinity (disparity = 0)
- Cars, buses & pedestrians have their typical height
- Road and pavement on the ground plane
- Buildings and pavement on the sides

Both problems formulated as CRF
- Joint approach possible?
Joint Formulation

\[ E(z) = \sum_{i \in \mathcal{V}} \psi_i^I(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}^I(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^J(z_c) \]

- Each pixel takes label \( z_i = [x_i \ y_i] \in L_1 \times L_2 \)
- Dependency of \( x_i \) and \( y_i \) encoded as a unary and pairwise potential, e.g.
  - strong correlation between \( x = \text{road}, y = \text{near ground plane} \)
  - strong correlation between \( x = \text{sky}, y = 0 \)
  - Correlation of edge in object class and disparity domain
Joint formulation

\[ E(z) = \sum_{i \in \mathcal{V}} \psi_i^L(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^J(z_c) \]

- Weighted sum of object class, depth and joint potential

\[ \psi_i^J([x_i, y_i]) = w_O^u \psi_i^O(x_i) + w_D^u \psi_i^D(y_i) + w_C^u \psi_i^C(x_i, y_i) \]

- Joint unary potential based on histograms of height
Joint Formulation

\[ E(z) = \sum_{i \in \mathcal{V}} \psi_i^L(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(z_i, z_j) + \sum_{c \in \mathcal{C}} \psi_c^J(z_c) \]

Pairwise Potential

- Object class and depth edges correlated
- Transitions in depth occur often at the object boundaries

\[ \psi_{ij}^J([x_i, y_i], [x_j, y_j]) = w^p_O \psi_{ij}^O(x_i, x_j) + w^p_D \psi_{ij}^D(y_i, y_j) + w^p_C \psi_{ij}^O(x_i, x_j) \psi_{ij}^D(y_i, y_j) \]
Joint Formulation
Standard α-expansion

- Each node in each expansion move keeps its old label or takes a new label $[x_{L1}, y_{L2}]$,
- Possible in case of metric pairwise potentials
• Standard $\alpha$-expansion
  • Each node in each expansion move keeps its old label or takes a new label $[x_{L1}, y_{L2}]$,
  • Possible in case of metric pairwise potentials

Too many moves! ($|L1| |L2|$)
Impractical!
Inference

- Projected move for product label space
  - One / Some of the label components remain(s) constant after the move

- Set of projected moves
  - $\alpha$-expansion in the object class projection
  - Range-expansion in the depth projection
• Leuven Road Scene dataset
  • Contained
    • 3 sequences
    • 643 pairs of images
  • We labelled
    • 50 training + 20 test images
    • Object class (7 labels)
    • Disparity (100 labels)
  • Available on our website
    • http://cms.brookes.ac.uk/research/visingroup/files/Leuven.zip
Qualitative results

- Large improvement for dense stereo estimation
- Minor improvement in object class segmentation
Quantitative disparity results

Dependency of the ratio of correctly labelled pixels within the maximum allowed error delta
• Application to monocular sequences
• Making method (close to) real time
• Application to multi-view problems
• Optical flow / motion estimation
• First dataset with both object class and disparity labels
• Joint estimation improves significantly disparity results
• Projected moves make inference much faster

• Questions ?