Object and Activity Recognition
Grounded on
Midlevel Image Representations

Sinisa Todorovic
joint work with: N. Payet, M. Amer

Oregon State University

September 20, 2011
Marr’s Vision

Visual Representations

Algorithms

Implémentation

Slide by SC Zhu
Open Basic Problems

• Semantic gap between visual features and categories
  • General vs. task-specific representations
Open Basic Problems

- Semantic gap between visual features and categories
- General vs. task-specific representations
- An observation: recent research mostly task-specific
Open Basic Problems

- Semantic gap between visual features and categories
  - General vs. task-specific representations
  - An observation: recent research mostly task-specific

- What are successful general representations?
  - Grammars and logic are back!
  - SIG-11 workshop at ICCV11
    - U Grenander, D Mumford, SC Zhu, A Yuille, L Davis, R Chellapa, N Ahuja, A Leonardis, P Felzenszwalb, S Todorovic ...
Goal

A unified computational framework capable of:

Discovery
Detection
Segmentation
Summarization

... of visual categories
in images and video
What is an Object?

**compositionality**

Spatial arrangement of its parts

input images  

partonomy = SCFG

Todorovic & Ahuja PAMI08
What is an Object?

compositionality

Spatial arrangement of its parts

Parts = Objects in their own right

Part discovery = Suspicious coincidences
What is an Object?

compositionality
Spatial arrangement of its parts

context
Spatial and semantic constraints with other objects

AND-OR graph: a hierarchy of random fields

Todorovic & Ahuja ICCV07
What is an Activity?

**compositionality**
Spatiotemporal arrangement of its parts

and

**context**
Spatiotemporal constraints with other activities
What is an Activity?

**compositionality**

Spatiotemporal arrangement of its parts

and

**context**

Spatiotemporal constraints with other activities

AND-OR graph: a hierarchy of random fields

Brendel & Todorovic CVPR11, ICCV11
Issues

- Human conceptualization
- Hierarchical
- Context sensitive models
- Features
- Images, video

Bottleneck

Grounding grammars on pre-selected low-level features
Example Low-Level Features

STIPs

HOGs

optical flow

video segmentation
Issues

Bottleneck

Grounding requires (probabilistic) repeatability of:
features and their spatiotemporal placement

Satisfied only in particular settings/tasks
(e.g., single-view object recognition)
# Example Issues

<table>
<thead>
<tr>
<th>low-level</th>
<th>mid-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>point-based features</td>
<td>regions, contours</td>
</tr>
<tr>
<td>stable, repeatable</td>
<td>unstable, poor repeatability</td>
</tr>
<tr>
<td>poor informativeness</td>
<td>informative</td>
</tr>
</tbody>
</table>
Hypothesis

Grounding grammars on summaries of low-level features
Hypothesis

Grounding grammars on summaries of low-level features where the summarization is guided top-down.
Objective

Formalize a mid-level feature that will be:
informative like regions (e.g., encode structure)
and
repeatable like points (e.g., view-invariant)
Bags of Right Features/Detections

If the category occurs,
it has to be in the spotlight of many BORDs
so they can jointly support the occurrence hypothesis

Perina & Jojic CVPR11
Example Problem: Activity Recognition

Given a video with noisy people detections
Example Problem

Given a video with noisy people detections

Detect and localize: all activity instances & actors

Amer & Todorovic ICCV11
Example Problem

Grounding the grammar on a space-time grid of Bags of Right Detections (BORDs)

Structure is encoded in the overlap of BORDs
Bags of Right Detections

\[ S_i \cdot x \]

shape context indicator

\# people detections

\# bins

BORD \( i \)
Bags of Right Detections

$x$ - latent variable

constrained by all BORDs

$S_i \cdot x$

shape context indicator
Detection and Localization

BORDs jointly constrain the solution

Amer & Todorovic ICCV11
Example Problem: Object Recognition

Given a set of edges in the image
detect and localize all object instances
and estimate their 3D pose

Payet & Todorovic ICCVII
Bags of Right Detections

BORDs jointly constrain the solution
Our Approach

Initial placement of BORDs on a regular grid

Search for optimal features by warping the grid

Detection & Localization

chains graphical model

Amer & Todorovic ICCV11
Our Approach

Initial placement of BORDs on a regular grid

Search for optimal features by warping the grid

MAP Inference:
1. Warp the grid to expected locations
2. Select MAP BORDs

guided by chains graphical model

video frames
The Chains Model

\[ P(M, O, L_S, L_E, F) \]

Number of BORDs → Ordering of BORDs → Start frame of activity → End frame of activity → Features → Observables

Latent variables

Number of BORDs
The Chains Model

\[ P(M, O, L_S, L_E, F) \]

\[ = P(M, O) P(L_S|M, O, F) P(L_E|M, O, F) \]

\[ \cdot \prod_{i} P(F_{O(i+1)}|F_{O(i)}) \prod_{i \in F - O} P_{bgd}(F_i) \]
MAP Inference

\[ P(M, O, L_S, L_E|F) \]
MAP Inference

\[ P(L_S, L_E|F) \propto \sum_{M,O} P(M, O, L_S, L_E, F) \]
MAP Inference

\[ P(L_S, L_E|F) \propto \sum_{M,O} P(M, O, L_S, L_E, F) \]

\[ = \sum_{M,O} P(M) \left[ \prod_{i,j} P(F_j|F_i) \right] P(L_S, L_E|M, O, F) \]
MAP Inference

\[ P(L_S, L_E | F) \propto \sum_{M,O} P(M, O, L_S, L_E, F) \]

\[ = \sum_{M,O} P(M) \left[ \prod_{i,j} P(F_j | F_i) \right] P(L_S, L_E | M, O, F) \]

\[ = \pi_T^{\text{start}} \left[ \sum_m P(M = m) X^m \right] \pi_{\text{end}} \]
MAP Inference = LP

\[
\text{minimize } \text{tr}\{C_b^T X\} + \alpha \| (I-W) X Q \|_1 + \beta \| (I-X) Q \|_1 \\
\text{subject to } X \geq 0, \ X 1_n = 1_n, \ b \geq 0, \ \| b \|_2^2 = 1.
\]

Searching for optimal features
under non-rigid shape deformations
of the grid of BORDs
Results

Correct detection and localization of:
  kicking and pushing
  and actors involved
Results

Correct detection and localization of:
handshaking and hugging
and actors involved
Results

video frames

Failure example

correct: handshaking and hugging
wrong: actors involved
Example Problem: Object Recognition

MAP inference = LP

Searching for optimal features under non-rigid shape deformations of the grid of BORDs

Payet & Todorovic ICCV'11
Correct detection, localization, and pose estimation
Correct detection, localization, and pose estimation
Conclusion

- Prior work: pre-selected features, typically low-level for repeatability

- Proposed mid-level features:
  - Allow abstraction of low-level features
  - Reduce the semantic gap
  - Enable addressing multiple tasks
  - Repeatable, and jointly encode structure
Acknowledgment

NSF IIS 1018490