Understanding Web Images by Object Relation Network

Na Chen, Qian-Yi Zhou, Viktor K. Prasanna
www2012, Lyon, France

USC University of Southern California
Automatic Image Understanding

- **Motivation**

  - Many useful web applications
    - Auto tagging, image search by image, auto natural language description generation, image clustering, ...

- **Scene**: soccer game
- **Action**: kick, chase
- **Objects**: a group of soccer players, a soccer ball

... Image semantics
Our Solution

- **Object Relation Network (ORN)**
  - ORN is a graph model representing the most probable yet semantically consistent ontology class assignments for image objects and their relations.

Applications
- Auto tagging
- Auto description generation
- Image search by Image

Raw image

ORN

- A Collection of SoccerPlayers
  - SoccerPlayer1
  - SoccerPlayer2
  - SoccerPlayer3

- SoccerBall1
  - kick
  - kick

This is a picture of a soccer game. There are three soccer players and one soccer ball. Soccer player 1 and Soccer player 2 are kicking the soccer ball.
Problem Context

- **Object detection offers computers an eye**
  - What is in the image and where it is

- **Limited to isolated, generic objects**

Detectors cannot tell the differences
Problem Context

- Ontology is a useful source for describing image semantics
  - Semantic hierarchies and constraints

![Ontology Diagram]

- Ontology-aided image annotation
  - Photo annotation (Schreiber et al., 2001)
  - Organizing the structure of image database (ImageNet, 2009)
- Hard to automate ontology-driven image annotation
Methodology

- **Object detection** – preprocess the image to identify generic objects
- **Guide ontology** – semantic source for objects and their relations
- **Energy minimization on a directed graphical model** – transfer ontological semantics to a network of objects connected by relations
System Pipeline

(a) Input
(b) Detection
(c) Guide ontology
(d) Directed graphical model
(e) Energy optimization towards the best labeling
(f) Output: Object Relation Network

Automatic tagging
Automatic description generation
Image search by image
Object Detection

- Discriminatively Trained Part Based Models (Felzenszwalb et al. IEEE TPAMI 2010)
  - 20 categories: person bicycle chair motorbike horse airplane bird boat bottle bus car cat cow table dog potted-plant sheep sofa train tv
- Open Source Computer Vision Library (OpenCV)
  - ball
System Pipeline
A Guide Ontology

- Four types of semantic knowledge are supported
  - Subsumption (subclass of)
  - Domain/range constraints
  - Cardinality constraints
  - Collection
- Any ontology regarding detectable objects and their relations can be adapted into the semantic knowledge layer
System Pipeline

- **Input**
- **Detection**
- **Directed graphical model**
  - Object nodes:
    - $P(o_1 = \text{BasketballPlayer}) = 0.44$
    - $P(o_1 = \text{Person}) = 0.37$
    - $P(o_1 = \text{SoccerPlayer}) = 0.19$
  - Relation nodes:
    - $P(r_{1,2} = \text{Hold}) = 0.10$
    - $P(r_{1,2} = \text{Kick}) = 0.08$
    - $P(r_{1,2} = \text{Interact}) = 0.00$

- **Energy optimization towards the best labeling**
  - $E(L) = E_r(L;\text{Ont}) + E_r(L;\text{Ont}) + E_r(L;\text{Img})$

- **Output**: Object Relation Network
  - Ball
  - BasketballPlayer
  - Person

- **Guide ontology**
- **Applications**
  - Automatic tagging
  - Automatic description generation
  - Image search by image

- **Examples**
  - A basketball player is throwing a basketball.
A feasible labeling
- each node is labeled with a class assignment that is a subclass of its generic class

Goal: find the best labeling $L$ that
- satisfies ontology constraints
- is as informative as possible
- makes the most probable class assignment on each node given its visual appearance

$L$ is predicted by minimizing an energy function

$$E(L) = E_c(L; Ont) + E_i(L; Ont) + E_v(L; Img)$$

where $E_c$, $E_i$, and $E_v$ are the constraint energy, informative energy, and visual energy, respectively.
Constraint Energy

$$E(L) = [E_c(L; Ont)] + E_i(L; Ont) + E_v(L; Img)$$

- **Constraint energy**
- **Informative energy**
- **Visual energy**

$$E_c(L; Ont) = \sum_{(o_i, r_{i,j})} E^D_c + \sum_{(r_{i,j}, o_j)} E^R_c + \sum_{(r_{i,j}, r_{i,k})} E^D_{c-Card} + \sum_{(r_{i,j}, r_{k,j})} E^R_{c-Card}$$

- **Domain/range constraints**

  $$E^D_c(o_i \leadsto C_o, r_{i,j} \leadsto C_r) = \begin{cases} 0 & \text{if } C_o \subseteq \text{domain}(C_r) \\ \infty & \text{otherwise} \end{cases}$$

  $$E^R_c(r_{i,j} \leadsto C_r, o_j \leadsto C_o) = \begin{cases} 0 & \text{if } C_o \subseteq \text{range}(C_r) \\ \infty & \text{otherwise} \end{cases}$$

- **Cardinality constraints**

  $$E^D_{c-Card}(r_{i,j} \leadsto C_1, r_{i,k} \leadsto C_2) = \begin{cases} \infty & \text{if } C_1 = C_2 = C_r \\ 0 & \text{otherwise} \end{cases}$$

  $$E^R_{c-Card}(r_{i,j} \leadsto C_1, r_{k,j} \leadsto C_2) = \begin{cases} \infty & \text{if } C_1 = C_2 = C_r \\ 0 & \text{otherwise} \end{cases}$$
Informative Energy

\[ E(L) = E_c(L;Ont) + E_i(L;Ont) + E_v(L;Img) \]

- **Constraint energy**
- **Informative energy**
- **Visual energy**

\[
E_i(L;Ont) = \sum_{o_i} E_i^O + \sum_{r_{i,j}} E_i^R + \sum_{(o_i,o_j)} E_i^{Col}
\]

- **Depth information**

\[
E_i^O(o_i \leadsto C_o) = -\omega_{dep} \cdot \text{depth}(C_o)
\]

\[
E_i^R(r_{i,j} \leadsto C_r) = -\omega_{dep} \cdot \text{depth}(C_r)
\]

- **Collection**

\[
E_i^{Col}(o_i \leadsto C_1, o_j \leadsto C_2) = \begin{cases} 
-\omega_{col} \frac{2}{N-1} \text{depth(collection}(C_o)) & \text{if } C_1 = C_2 = C_o \\
0 & \text{otherwise}
\end{cases}
\]
Visual Energy

\[ E(L) = E_c(L; Ont) + E_i(L; Ont) + \boxed{E_v(L; Img)} \]

- Constraint energy
- Informative energy
- Visual energy

\[ E_v(L; Img) = \sum_{o_i} E^O_v + \sum_{r_{i,j}} E^R_v \]

- \( P(\text{basketball}) > P(\text{ball}) \)
- \( P(\text{interact}) > P(\text{non-interact}) \)

- **On object nodes**

\[ E^O_v(o_i \rightsquigarrow C_o) = -\omega_{\text{obj}} P(o_i \rightsquigarrow C_o | \mathcal{F}_o(O_i)) \]

\[ \mathcal{F}_o(O_i) = \{ \text{width}_{bbox}, \text{height}_{bbox}, \bar{H}, \bar{S}, \bar{V}, \sigma_H, \sigma_S, \sigma_V \} \]

- **On relation nodes**

\[ E^R_v(r_{i,j} \rightsquigarrow C_r) = -\omega_{\text{rel}} P(r_{i,j} \rightsquigarrow C_r | \mathcal{F}_r(O_i, O_j)) \]

\[ \mathcal{F}_r(O_i, O_j) = \{ \text{width}_{bbox_i}, \text{height}_{bbox_i}, x_{\text{center}_i}, y_{\text{center}_i}, \text{width}_{bbox_j}, \text{height}_{bbox_j}, x_{\text{center}_j}, y_{\text{center}_j} \} \]
ORN Generation

- Search for the best labeling $\mathcal{L}$ that has the minimum energy

$$E(L) = E_c(L;\text{Ont}) + E_i(L;\text{Ont}) + E_v(L;\text{Img})$$

- Apply $\mathcal{L}$ over the directed graph model
- Create collection nodes
- Remove meaningless relation nodes
Experiments

- **Dataset**
  - over 28,000 Flickr images from VOC2011 and ImageNet
  - 2000 images for training
  - The guide ontology we used

- 6 categories of detectable objects
- Simplified semantic hierarchies from WordNet
Some Good ORNs
More Good ORNs

Our ORN can be correct even if detection goes wrong!
Some Bad ORNs

False detection:  

Missing detection:
Application of ORNs - I

- **Automatic Image Tagging**
  - Ontological class assignments in the ORN as tags

- Using simple inference rules to infer more semantics

```
∃x CyclistCollection(x) ∧ ∃y Cyclist(y) ∧ ∃z Bicycle(z) ∧ ride(y, z) ∧ Tag(t) → t = “bicycle race”
```

person
soccer ball
ball holding

bicycle race
cyclists
bicycles
bicycle riding
More Auto-tagging Results

Tags from good ORNs:

- motorcyclists, motorbike, motorbike riding
- soccer game, soccer players, soccer ball, ball kicking

Tags from bad ORNs:

- chairs
- soccer ball, person
Application of ORNs - II

- **Automatic Image Description Generation**
  - Extend automatic tagging by employing a simple template-based model

This is a picture of a **soccer game**. There are **five soccer players** and **one soccer ball**. Soccer player 3 is **kicking** the soccer ball.

This is a picture of a **bicycle race**. There are two **cyclists** and two **bicycles**. Cyclist 1 is riding Bicycle 1. Cyclist 2 is riding Bicycle 2.

There are one person and one soccer ball. The person is holding the soccer ball.
Application of ORNs - III

- **Image Search by Image**
  - Measure image similarity by graph distance between ORNs
  - Test image library contains over 30,000 images

<table>
<thead>
<tr>
<th>Query image &amp; its ORN</th>
<th>Top 5 results ranked by ORN distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Query Image" /></td>
<td><img src="image2.png" alt="Top Result 1" /> <img src="image3.png" alt="Top Result 2" /> <img src="image4.png" alt="Top Result 3" /> <img src="image5.png" alt="Top Result 4" /> <img src="image6.png" alt="Top Result 5" /></td>
</tr>
</tbody>
</table>

*ORN representations: HorseRider1, Horse1, ride*
More Search Results

Query image & its ORN

Top 5 results ranked by ORN distance

A Collection of Chairs

Chair1

Chair2

Chair3

Chair4

A Collection of SoccerPlayers

SoccerPlayer1

SoccerPlayer2

SoccerPlayer3

SoccerBall1

kick

kick
Conclusion & Future Work

Contributions

- We propose and exploit Object Relation Network towards automatic web image understanding.
- ORN adds semantic structure to image recognition
- We propose and demonstrate three application scenarios that can benefit from ORNs.

Future work

- Explore a larger ontology
- Incorporate more detection/segmentation categories
Thank You!

- **Contact:**
  - Na Chen  nchen@usc.edu
  - Qian-Yi Zhou  qianyizh@usc.edu
  - Viktor K. Prasanna  prasanna@usc.edu