Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions

Bangpeng Yao and Li Fei-Fei
Computer Science Department, Stanford University

{bangpeng, feifeili}@cs.stanford.edu
Human-Object Interaction

Playing saxophone

Not playing saxophone
Human-Object Interaction

Robots interact with objects
Automatic sports commentary
“Kobe is dunking the ball.”
Medical care
Background: Human-Object Interaction

- Schneiderman & Kanade, 2000
- Viola & Jones, 2001
- Huang et al, 2007
- Papageorgiou & Poggio, 2000
- Wu & Nevatia, 2005
- Dalal & Triggs, 2005
- Mikolajczyk et al, 2005
- Leibe et al, 2005
- Bourdev & Malik, 2009
- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

To be done

- Lowe, 1999
- Belongie et al, 2002
- Fergus et al, 2003
- Fei-Fei et al, 2004
- Berg & Malik, 2005
- Felzenszwalb et al, 2005
- Grauman & Darrell, 2005
- Sivic et al, 2005
- Lazebnik et al, 2006
- Zhang et al, 2006
- Savarese et al, 2007
- Lampert et al, 2008
- Desai et al, 2009
- Gehler & Nowozin, 2009

context

- Murphy et al, 2003
- Hoiem et al, 2006
- Shotton et al, 2006

- Rabinovich et al, 2007
- Heitz & Koller, 2008
- Divvala et al, 2009

- Gupta et al, 2009
- Yao & Fei-Fei, 2010a
- Yao & Fei-Fei, 2010b
Background: Human-Object Interaction

• Schneiderman & Kanade, 2000
• Viola & Jones, 2001
• Huang et al, 2007
• Papageorgiou & Poggio, 2000
• Wu & Nevatia, 2005
• Dalal & Triggs, 2005
• Mikolajczyk et al, 2005
• Leibe et al, 2005
• Bourdev & Malik, 2009
• Felzenszwalb & Huttenlocher, 2005
• Ren et al, 2005
• Ramanan, 2006
• Ferrari et al, 2008
• Yang & Mori, 2008
• Andriluka et al, 2009
• Eichner & Ferrari, 2009

• Lowe, 1999
• Belongie et al, 2002
• Fergus et al, 2003
• Fei-Fei et al, 2004
• Berg & Malik, 2005
• Felzenszwalb et al, 2005
• Grauman & Darrell, 2005
• Sivic et al, 2005
• Lazebnik et al, 2006
• Zhang et al, 2006
• Savarese et al, 2007
• Lampert et al, 2008
• Desai et al, 2009
• Gehler & Nowozin, 2009

• Lowe, 1999
• Belongie et al, 2002
• Fergus et al, 2003
• Fei-Fei et al, 2004
• Berg & Malik, 2005
• Felzenszwalb et al, 2005
• Grauman & Darrell, 2005
• Sivic et al, 2005
• Lazebnik et al, 2006
• Zhang et al, 2006
• Savarese et al, 2007
• Lampert et al, 2008
• Desai et al, 2009
• Gehler & Nowozin, 2009

• Gupta et al, 2009
• Yao & Fei-Fei, 2010a
• Yao & Fei-Fei, 2010b

context

• Murphy et al, 2003
• Hoiem et al, 2006
• Shotton et al, 2006

• Rabinovich et al, 2007
• Heitz & Koller, 2008
• Divvala et al, 2009

To be done

• Gupta et al, 2009
• Yao & Fei-Fei, 2010a
• Yao & Fei-Fei, 2010b

vs
Outline

• Intuition of Grouplet Representation
• Grouplet Feature Representation
• Using Grouplet for Recognition
• Dataset & Experiments
• Conclusion
Outline

• Intuition of Grouplet Representation
• Grouplet Feature Representation
• Using Grouplet for Recognition
• Dataset & Experiments
• Conclusion
Recognizing Human-Object Interaction is Challenging

Reference image: playing saxophone

Different pose (or viewpoint)

Different lighting

Different background

Different instrument, similar pose

Same object (saxophone), different interactions
Grouplet: our intuition

Bag-of-words  Spatial pyramid  Part-based  Grouplet
Representation:

- Thomas & Malik, 2001
- Csurka et al, 2004
- Fei-Fei & Perona, 2005
- Sivic et al, 2005
- Weber et al, 2000
- Fergus et al, 2003
- Leibe et al, 2004
- Lazebnik et al, 2005
- Felzenszwalb et al, 2005
- Bourdev & Malik, 2009
Grouplet: our intuition

Grouplet Representation:

- Part-based configuration
- Co-occurrence
- Discriminative
- Dense

Capture the subtle difference in human-object interactions.
Outline

• Intuition of Grouplet Representation
• Grouplet Feature Representation
• Using Grouplet for Recognition
• Dataset & Experiments
• Conclusion
Grouplet representation (e.g. 2-Grouplet)

Notations

- $I$: Image.
- $P$: Reference point in the image.
- $\Lambda$: Grouplet.
- $\lambda_i$: Feature unit.
  - $A_i$: Visual codeword;
  - $x_i$: Image location;
  - $\sigma_i$: Variance of spatial distribution.
Grouplet representation (e.g. 2-Grouplet)

Notations

- \( I \): Image.
- \( P \): Reference point in the image.
- \( \Lambda \): Grouplet.
- \( \lambda_i \): Feature unit.
  - \( A_i \): Visual codeword;
  - \( x_i \): Image location;
  - \( \sigma_i \): Variance of spatial distribution.
- \( \nu(\Lambda, I) \): Matching score of \( \Lambda \) and \( I \).
- \( \nu(\lambda_i, I) \): Matching score of \( \lambda_i \) and \( I \).

Visual codewords

Gaussian distribution

\[ \nu(\Lambda, I) \]
Matching score between \( \Lambda \) and \( I \)
Grouplet representation (e.g. 2-Grouplet)

$I$: Image.
$P$: Reference point in the image.
$\Lambda$: Grouplet.
$\lambda_i$: Feature unit.
- $A_i$: Visual codeword;
- $x_i$: Image location;
- $\sigma_i$: Variance of spatial distribution.

$v(\Lambda, I)$: Matching score of $\Lambda$ and $I$.
$v(\lambda_i, I)$: Matching score of $\lambda_i$ and $I$.

For an image patch:
- $a'$: Its visual appearance;
- $x'$: Its image location.
$\Omega(x)$: Image neighborhood of $x$.

$$v(\Lambda, I) = \min_i \left\{ v(\lambda_i, I) \right\} = \min_i \left\{ \sum_{x' \in \Omega(x_i)} \left[ p(A_i \mid a') \cdot N(x' \mid x_i, \sigma_i) \right] \right\}$$

Matching score between $\Lambda$ and $I$
Matching score between $\lambda_i$ and $I$
Codeword assignment score
Gaussian density value
Grouplet representation (e.g. 2-Grouplet)

Notations

- $I$: Image.
- $P$: Reference point in the image.
- $\Lambda$: Grouplet.
- $\lambda_i$: Feature unit.
  - $A_i$: Visual codeword;
  - $x_i$: Image location;
  - $\sigma_i$: Variance of spatial distribution.
- $\nu(\Lambda, I)$: Matching score of $\Lambda$ and $I$.
- $\nu(\lambda_i, I)$: Matching score of $\lambda_i$ and $I$.
- For an image patch:
  - $a'$: Its visual appearance;
  - $x'$: Its image location.
- $\Omega(x)$: Image neighborhood of $x$.
- $\Delta$: A small shift of the location.

Matching score between $\Lambda$ and $I$

$\nu(\Lambda, I) = \min_i \left\{ \nu(\lambda_i, I) \right\}$

Matching score between $\lambda_i$ and $I$

$\nu(\lambda_i, I) = \min_j \left\{ \max_{a'} \left\{ \sum_{x' \in \Omega(x')} \left[ p(\sum_{i} a' \cdot a_i | \lambda_i) \cdot p(N(x' | \lambda_i, \sigma_i)) \right] \right\} \right\}$
Grouplet representation

- Part-based configuration
- Co-occurrence
- Discriminative

Playing saxophone

Other interactions

(matching score: 0.6  matching score: 0.4  matching score: 0.0  matching score: 0.1)
Grouplet representation

- Part-based configuration
- Co-occurrence
- Discriminative
- Dense

All possible Codewords

Densely sample image locations

Many possible spatial distributions

All possible combinations of feature units

1-grouplet 2-grouplet 3-grouplet
Outline

• Intuition of Grouplet Representation
• Grouplet Feature Representation
• Using Grouplet for Recognition
• Dataset & Experiments
• Conclusion
A “Space” of Grouplets
A “Space” of Grouplets

Playing violin
Other interactions
A “Space” of Grouplets

Playing saxophone  Other interactions

Playing violin  Other interactions
A “Space” of Grouplets

Playing saxophone  Other interactions

Playing violin  Other interactions

On background

Shared by different interactions
We only need discriminative Grouplets

Number of feature units: $N$.
$N$ is large (192200)

Number of Grouplets: $2^N$
very large space
Obtaining discriminative grouplets for a class

Obtain grouplets with large \( \nu(\Lambda, I) \) on the class.

Remove grouplets with large \( \nu(\Lambda, I) \) from other classes.

Apriori Mining

Selected 1-grouplets

Candidate 2-grouplets

Mine 1000~2000 grouplets, only need to evaluate \((2~100) \times N\) grouplets

Number of feature units: \( N \).

\( N \) is large (192200)

**Number of Grouplets:** \( 2^N \)

very large space

[Agrawal & Srikant, 1994]
Using Grouplets for Classification

Discriminative grouplets $[\Lambda_1, \ldots, \Lambda_N]$ → $I$ → $[\nu(\Lambda_1, I), \ldots, \nu(\Lambda_N, I)]$ → SVM
Outline

• Intuition of Grouplet Representation
• Grouplet Feature Representation
• Using Grouplet for Recognition
• Dataset & Experiments
• Conclusion
People-Playing-Musical-Instruments (PPMI) Dataset

http://vision.stanford.edu/resources_links.html

PPMI+

# Image: (172) (191) (177) (179) (200) (198) (185)

PPMI-

# Image: (164) (148) (133) (149) (188) (169) (167)

Original image

Normalized image
(200 images each interaction)
Recognition Tasks on People-Playing-Musical-Instruments (PPMI) Dataset

**Classification**

- Playing different instruments
  - Playing French horn
  - Playing violin
  vs.
  - Not playing violin

- Playing vs. Not playing
  - Playing violin
  vs.
  - Not playing violin

**Detection**

For each interaction, 100 training and 100 testing images.
Classification: Playing Different Instruments

- 7-class classification on PPMI+ images

![Bar chart showing classification accuracy for various methods]

![Graph showing the number of mined grouplets vs grouplet size]

SPM: [Lazebnik et al, 2006]
DPM: [Felzenszwalb et al, 2008]
Constellation: [Fergus et al, 2003]
[Niebles & Fei-Fei, 2007]
Classifying Playing vs. Not playing

- Seven 2-class classification problem; PPMI+ vs. PPMI- for each instrument.
Classifying Playing vs. Not playing

- Seven 2-class classification problem; PPMI+ vs. PPMI- for each instrument.

---

![Graph showing classification accuracies for different instruments.

**Axes:**
- Y-axis: Accuracy
- X-axis: Instruments (Bassoon, Erhu, Flute, French horn, Saxophone, Violin)

**Legend:**
- BoW
- DPM
- SPM
- Grouplet+SVM

**Observations:**
- Flute and Saxophone have the highest accuracy for Grouplet+SVM.
- French horn shows a notable peak for DPM.
- Violin has a consistent performance across all methods.

---

**Average PPMI+ images:**
- Visual representation of average PPMI+ images for each instrument.

**Average PPMI- images:**
- Visual representation of average PPMI- images for each instrument.
Detecting people playing musical instruments

Procedure:
- Face detection with a low threshold;
- Crop and normalize image regions;
- 8-class classification
  - 7 classes of playing instruments;
  - Another class of not playing any instrument.

Playing saxophone  No playing  No playing
Detecting people playing musical instruments

Area under the precision-recall curve:

- Out method: 45.7%;
- Spatial pyramid: 37.3%.
Detecting people playing musical instruments

Area under the precision-recall curve:
  - Out method: 45.7%;
  - Spatial pyramid: 37.3%.
Examples of Mined Grouplets

Playing bassoon:

Playing saxophone:

Playing violin:

Playing guitar:
Conclusion

• Holistic image-based classification

Vs.

• Detailed understanding and reasoning

Pose estimation & object detection


Thanks to
Juan Carlos Niebles, Jia Deng, Jia Li, Hao Su, Silvio Savarese, and anonymous reviewers.

And You