Grasping Pattern Estimation Based on Co-occurrence of Object and Hand Shape

Takuya Kawakami, Tadashi Matsuo, Yoko Ogawa, Nobutaka Shimada,
Ritsumeikan University
Introduction

• An object as a tool has its own function. The function is closely related to how a human grasp it [1].

Can we estimate how to grasp an object from the object itself?

It will be useful for object recognition and robot manipulation.

Related work


They achieved highly accurate classification by utilizing how to grasp an object, but...

• It estimates only an object label (not how to grasp it).
• All teacher labels must be given manually.
Our goal

Learn human interactions without teacher labels.

Training

Recall how a human grasp it.

Make action to grasp it.
Proposed method

• We generate an interaction descriptor, a numeral representation of a human grasping method.

• And then we make an inference model to learn the relation between object and grasping method.
Grasping method is represented as a grasping image. It consists of a depth image, hand mask and object mask. It is paired with the corresponding object image.
Capture of human’s grasping scene

Observing human’s grasping scene

RGB-D sensor

Grasping procedure

Only object

Time series

Grasping
Put up

Captured point cloud

Remove unnecessary points

Hand and object points
Alignment based on object

Tracking of point clouds of all frames by ICP

Object only scene

input

Coordinate transformation by ICP

Align
Segmentation of a hand and an object

Just before changing the number of regions, we segment hand points / object points.

Hand point cloud

Object point cloud

Using hand point

Using object point

Depth

Hand mask

Object mask
Interaction descriptor

\[ D(E(I_{\text{grasp}})) \]

\( I_{\text{grasp}} \)

Grasping image

Encoder

Decoder

Interaction descriptor

Restored grasping image

\( E \) and \( D \) are trained by minimizing restoration error without teacher labels.

A descriptor represents essence of an input.
Shift invariant auto-encoder

An ordinary auto-encoder encodes shape and position. But spatial shift in grasping images is not important. We use shift invariant auto-encoder to encode a shape itself. (descriptor includes shape information only)

Structure of auto-encoder

Encoder $E$

Grasping image
$32 \times 32 \times 3$ ch

Convolution layer

Max Pooling

Full connection

Decoder $D$

Interaction descriptor

Restored grasping image
$32 \times 32 \times 3$ ch

Parameters

Input size: $32 \times 32 \times 3$

Descriptor dimension: 30

Kernel size: $9 \times 9$
Objects and grasping types

Mug | Cup | Ball | Spray
---|---|---|---
Training
---|---|---|---
Test

Grasping types

Mug | Cup | Ball | Spray
---|---|---|---
Training images: 80scenes × 12kinds = 960
Test images: 80scenes × 4kinds = 320
Restored grasping images

Input grasping images

<table>
<thead>
<tr>
<th>mug</th>
<th>cup</th>
<th>ball</th>
<th>spray</th>
</tr>
</thead>
</table>

Images restored from interaction descriptor

Interaction descriptor has approximate shape information.
Distribution of interaction descriptors

- Non-frame: Object for training
- Red frame: Object for test

1st principal component

2nd principal component
**Inference model**

- **Known interaction descriptor (teacher)**
  \[ P_{teacher} \]

- **Train** \( R \) **to minimize**
  \[ \| P_{recall} - P_{teacher} \|^2 \]

- **Recalled interaction descriptor**
  \[ P_{recall} = R(I_{obj}) \]

- **Inference model:** \( R \)

- **Paired for the same object**

- **Unknown object image**

- **Restored grasping image**
Structure of the inference model

Patch image $32 \times 32$

Convolution

$14 \times 14 \times 16$ch

$L_2$ Pooling

Convolution

$5 \times 5 \times 256$ch

$L_2$ Pooling

Full connection

$128$ nodes

$30$ nodes

Interaction descriptor

Object images

*mug*  
*cup*  
*ball*  
*spray*  
*Train*  
*Test*
Recalled grasping images (train)

The inference model successfully recalls grasping images.
Recalled grasping images (test)

The model approximately recalls hand region masks.
The model approximately recalls hand region masks.
Recall from images with/without an important part

Mug

with handle

without handle

Cup

Recalled grasping image

Depth

Hand

Object

Similar!
Recall from images with/without an important part

Object image 64×64

Patch image 32×32

Correct grasping image

Recalled grasping image

Restored grasping image

?
Integration of recalled hand region masks

Recalled image
(32 × 32)

Integrated hand region masks (64 × 64)

Integrate descriptors in the same cluster

Interaction descriptor space
Multiple grasping types for object

To see part-specific inference, we train auto-encoder and inference model with below grasping types.

- **mug**: Grasping type 1, Grasping type 2
- **cup**: Grasping type 2
- **ball**: Grasping type 1
- **spray**: Grasping type 2
Integrated hand region mask

Integrated hand region mask

Cluster 1

Cluster 2

A real example of grasping

The integrated hand mask for cluster $i$ is defined as:

$$P_i(x, y) = \frac{S_i(x, y)}{N_i(x, y)}$$

$S_i(x, y)$: Sum of recalled hand mask in the $i$-th cluster

$N_i(x, y)$: Number of non-zero at $(x, y)$ of recalled hand mask in the $i$-th cluster

Integrated hand region mask indicates hand region when human grasps the object.
Integrated hand region mask

A real example of grasping

The integrated hand mask for cluster $i$ is defined as:

$$S_i(x, y) = \frac{S_i(x, y)}{N_i(x, y)}$$

$S_i(x, y)$: Sum of recalled hand mask in the $i$-th cluster

$N_i(x, y)$: Number of non-zero values ($y$) of recalled hand mask in the $i$-th cluster

Integrated hand region mask indicate hand region when a human grasps the object.
Conclusion

• We proposed a method to recall grasping method from an object. It is based on:
  • Interaction descriptor by shift invariant auto-encoder
    We can generate numeral representation of grasping method without teacher labels.
  • Inference model by CNN
    The relation between object shape and grasping method can be modeled by utilizing interaction descriptor.

• The proposed method can estimate hand region for grasping an object from the object itself.

• The proposed method will be useful for robot manipulator.
We trained auto-encoders with shifted MNIST training images.

Distributions of descriptors from shifted test images such as :

Input: $32 \times 32$
Descriptor dim: 30
Max shift width: 8

Descriptors from common shapes automatically aggregate.
Example for hand-object interaction

Train AEs with 2-channel images consisting of hand/object masks.

Restored images

Ordinary AE  Shift invariant AE
Example for hand-object interaction

Train AEs with 2-channel images consisting of hand/object masks.

Depending on positions

Representing typical situation of grasping scissors

Situation

Hand mask

Object mask

Input

Ordinary AE

Shift invariant AE
Images restored by an ordinary auto-encoder

Images restored by a shift invariant auto-encoder