# Fingerspelling recognition with two-steps cascade process of spotting and classification

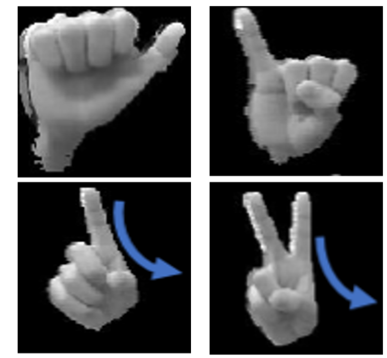
Masanori Muroi[1], Naoya Sogi[1], Nobuko Kato[2] and Kazuhiro Fukui[1]

[1] Graduate School of Systems and Information Engineering,
University of Tsukuba, Japan

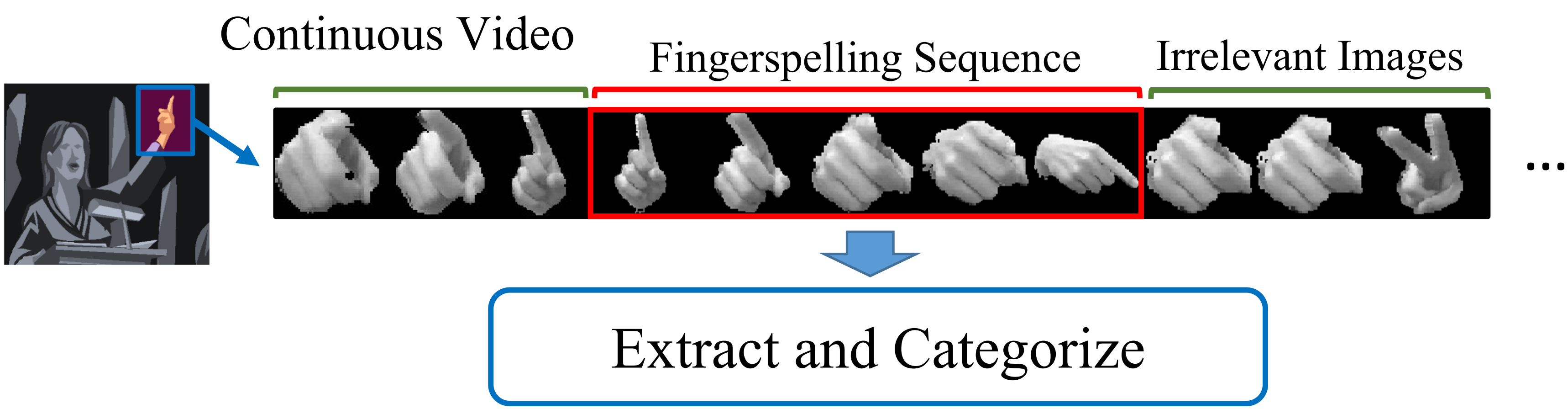
[2] Faculty of Industrial Technology,
Tsukuba University of Technology, Japan

# Motivation

- Fingerspelling is a tool to express a certain letter by a hand shape.
  - Used in conjunction with sign language



- **Main goal: Extract and categorize fingerspelling sequences in a continuous video.**



# Basic Idea

- Divide a whole process into two-steps: **Spotting** and **Classification**
- **Step 1. Spotting**: <u>Segment and extract a fingerspelling sequence</u>


- Identify a given sequence to <u>fingerspelling</u> or <u>irrelevant</u> (= Binary Classification)

- Extract a fingerspelling sequence

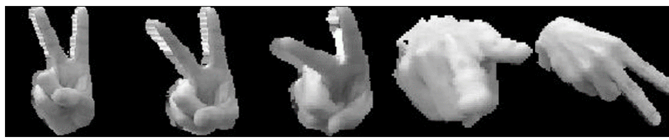
**Step2: Classification Process**

We should consider **<u>temporal dynamic information</u>**.

# Basic Idea

- **Step 2. Classification**: Classify the spotted fingerspelling sequence

Spotted Sequence



3D Hand Shape

3D Object Classification

Class 1 | Class 2 | Class 3 | … | Class *C*

We should consider **3D hand shape information**.

# Solution to realize the Basic Idea

- Propose a fingerspelling recognition framework based on the two types of methods:

Spotting: **Temporal Regularized CCA (TRCCA)**[2]

➢ The smoothness on the temporal domain

Classification: **Orthogonal Mutual Subspace Method (OMSM)**[3] with **CNN features**[4]

➢ The subspace representation of multiple images

[2]S. Tanaka, A. Okazaki, N. Kato, H. Hino and K. Fukui, Spotting ngerspelled words from sign language video by temporally regularized canonical component analysis, *2016 IEEE International Conference on Identity, Security and Behavior Analysis*, 2016, pp. 1-7.
[3] K. Fukui and O. Yamaguchi, The kernel orthogonal mutual Subspace method and its application to 3D object recognition, *in Asian Conference on Computer Vision*, 2007, pp. 467-476.
[4] N. Sogi, T. Nakayama, and K. Fukui, A method based on convex cone model for image-set classication with cnn features, *in 2018 International Joint Conference on Neural Networks*, 2018, pp. 1-8.

# Solution to realize the Basic Idea

- Propose a fingerspelling recognition framework based on the two types of methods:

Spotting: **Temporal Regularized CCA (TRCCA)**[2]

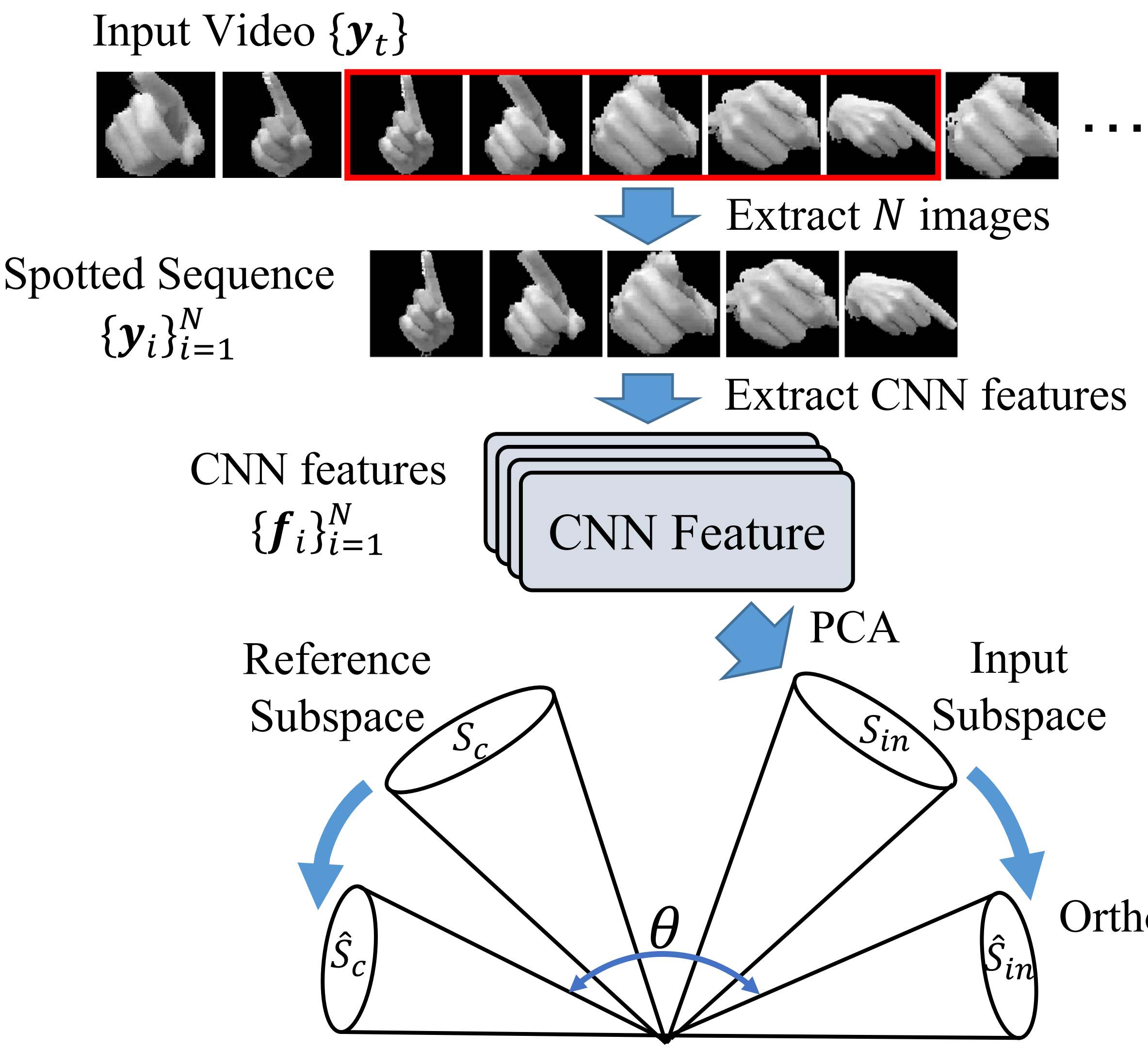
➢ The smoothness on the temporal domain

Classification: **Orthogonal Mutual Subspace Method (OMSM)**[3] with **CNN features**[4]

➢ The subspace representation of multiple images

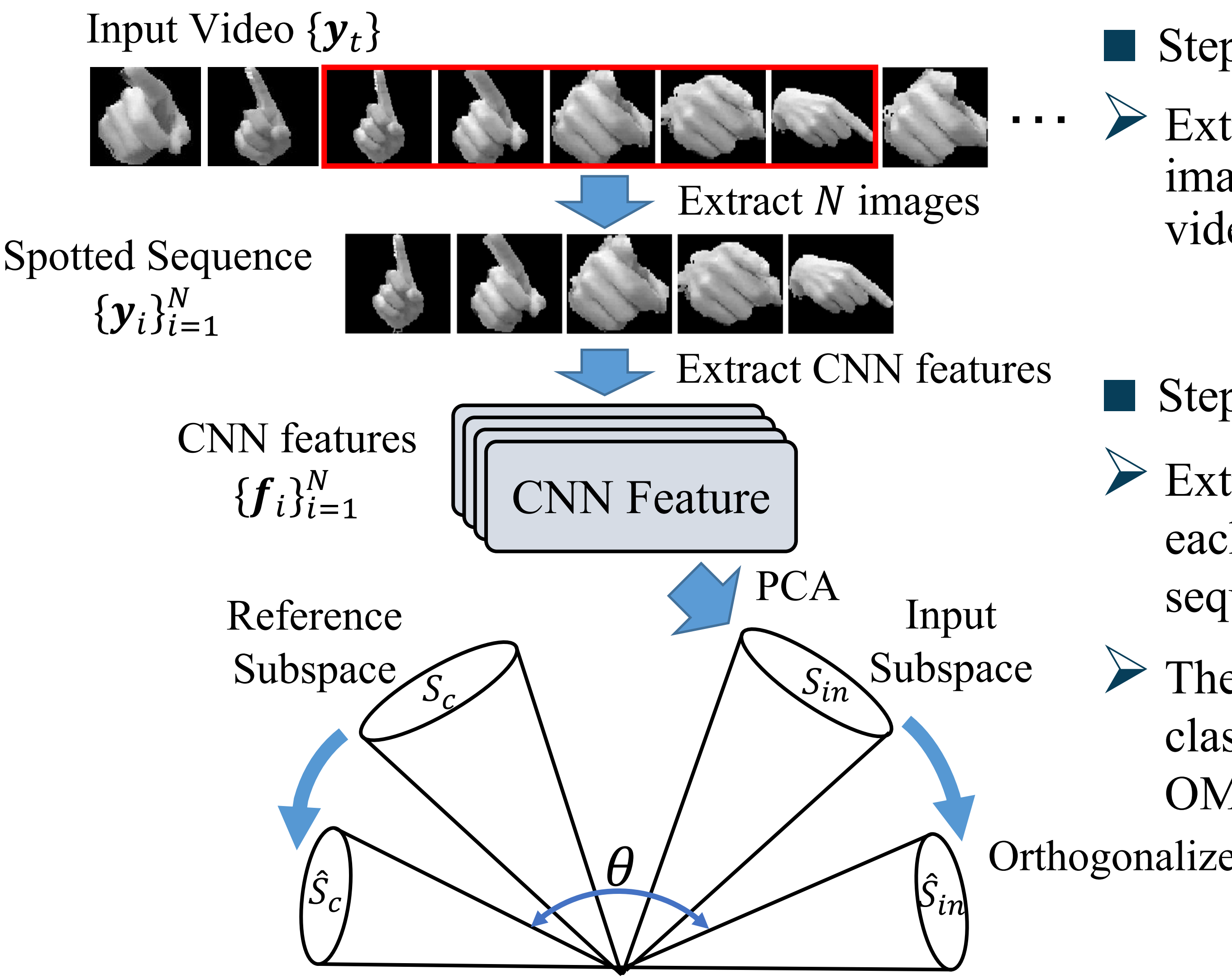
[2]S. Tanaka, A. Okazaki, N. Kato, H. Hino and K. Fukui, Spotting ngerspelled words from sign language video by temporally regularized canonical component analysis, *2016 IEEE International Conference on Identity, Security and Behavior Analysis*, 2016, pp. 1-7.
[3] K. Fukui and O. Yamaguchi, The kernel orthogonal mutual Subspace method and its application to 3D object recognition, *in Asian Conference on Computer Vision*, 2007, pp. 467-476.
[4] N. Sogi, T. Nakayama, and K. Fukui, A method based on convex cone model for image-set classication with cnn features, *in 2018 International Joint Conference on Neural Networks*, 2018, pp. 1-8.

# Proposed Framework for Fingerspelling Recognition




- **Step 1:**
  - Extract $N$ fingerspelling images from an input video using TRCCA
- **Step 2:**
  - Extract CNN features from each frame of the spotted sequence
  - The set of CNN features is classified by applying OMSM

# Proposed Framework for Fingerspelling Recognition




- **Step 1:**
  - Extract $N$ fingerspelling images from an input video using TRCCA
- **Step 2:**
  - Extract CNN features from each frame of the spotted sequence
  - The set of CNN features is classified by applying OMSM

# Classification accuracy and recognition time

- Accuracies and recognition times of different frameworks.

| Framework | Accuracy | Recognition Time |
|---|---|---|
| TRCCA [1] | 64.1% | **39.7 ms** |
| CNN feat- OMSM | 68.9% | 52.7 ms |
| KOTRCCA [1] | 79.0% | 169.0 ms |
| TRCCA-CNN(softmax) | 80.7% | 56.9 ms |
| TRCCA-KOMSM[2] | 86.9% | 187.3 ms |
| **TRCCA-CNN feat-OMSM(Proposed)** | **88.2%** | 91.2 ms |