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

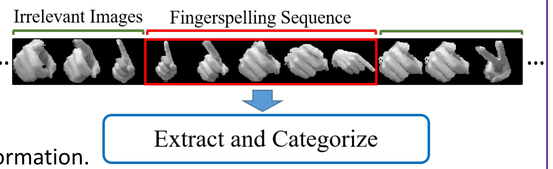
Masanori Muroi[1], Naoya Sogi[1], Nobuko Kato[2], Kazuhiro Fukui[1]

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

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

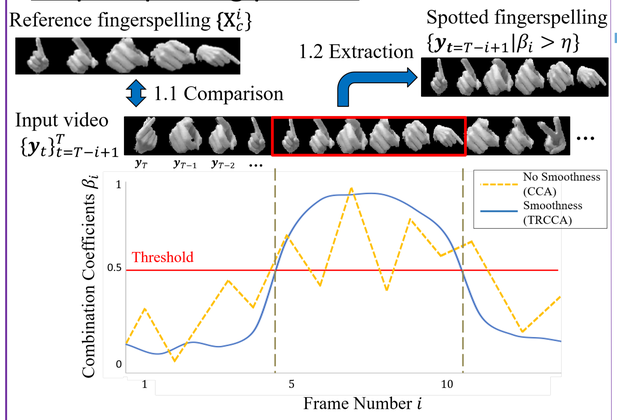
(1) Introduction

- Fingerspelling is a tool to express a letter by a hand shape. ➢ Used in conjunction with sign language
- **Main goal : Detect and categorize fingerspelling in a continuous video as a mixture of fingerspelling sequences and irrelevant images.**
- **Basic idea :** Divide a whole process into two-steps cascade process:
  1. *Spotting* : Segment and extract a fingerspelling sequence in an input video by utilizing temporal dynamic information.
  2. *Classification* : Classify the spotted sequence by utilizing 3D hand shape information.
- We propose a fingerspelling classification framework based on two types of methods:
  1) ***Temporal Regularized CCA (TRCCA)***[1] for spotting
  2) ***Orthogonal Mutual Subspace Method (OMSM)***[2] with ***CNN feature***[3] for classification



(2) Proposed framework

Step1: Spotting process



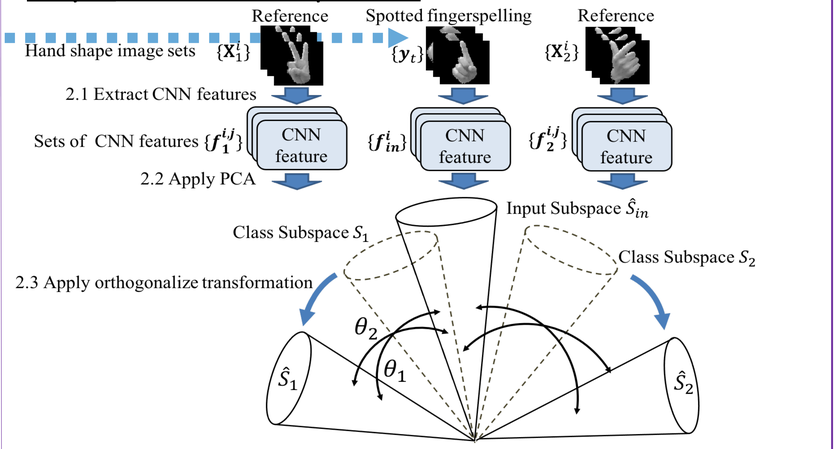
- Fingerspelling images in an input video are segmented and extracted using TRCCA.
- TRCCA: Extension of the *canonical correlation analysis* with the smoothness on the temporal domain
  - ➢ The smoothness efficiently incorporates the temporal information

- The detailed procedure

1.1 The input image sequence $\{\boldsymbol{y}_t\}_{t=T-i+1}^{T}$ is compared with the reference fingerspelling $\{\mathbf{X}_c^i\}$ by TRCCA. If the input sequence has high similarity with the reference fingerspelling, the input sequence is identified to fingerspelling.

1.2 Fingerspelling images $\{\boldsymbol{y}_{t=T-i+1}|\beta_i > \eta\}$ are extracted from the input sequence.

Step2: Classification process



- Spotted fingerspelling is classified by OMSM with CNN features using hand shape image sets.
- OMSM: Represent 3D shape of a hand by subspace.
  - ➢ The subspace representation is effective for 3D object recognition

- The detailed procedure

2.1 CNN features $\{\boldsymbol{f}_{in}^i\}$ and $\{\boldsymbol{f}_c^{i,j}\}$ are extracted from $\{\boldsymbol{y}_{t=T-i+1}|\beta_i > \eta\}$ and $\{\mathbf{X}_c^i\}$.

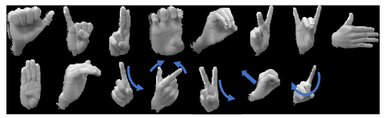
2.2 Each class subspace $\{S_c\}$ and an input subspace S_{in} are generated by applying PCA to the sets of CNN features.

2.3 Orthogonal subspaces $\{\hat{S}_c\}$ and $\hat{S}_{in}$ are generated by applying orthogonalize transformation to $\{S_c\}$ and S_{in}.

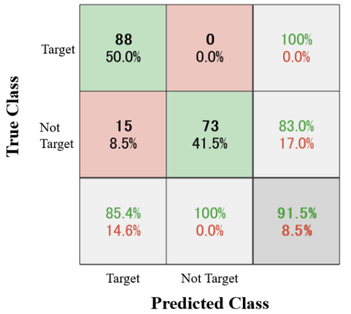
2.4 The spotted fingerspelling is classified based on similarities between the input subspace S_{in} and reference subspaces $\{\hat{S}_c\}$.

(3) Experiments

- Dataset:
  - ➢ We recorded **15 fingerspelling classes** by a depth camera.
  - ☐ The hand region is extracted from the whole input image based on the depth map.
  - ➢ We synthesized an input video, which continuously inputs fingerspelling and not fingerspelling sequences alternately.
- Evaluation index:
  - ➢ Spotting performance, Classification accuracy, and Recognition time.



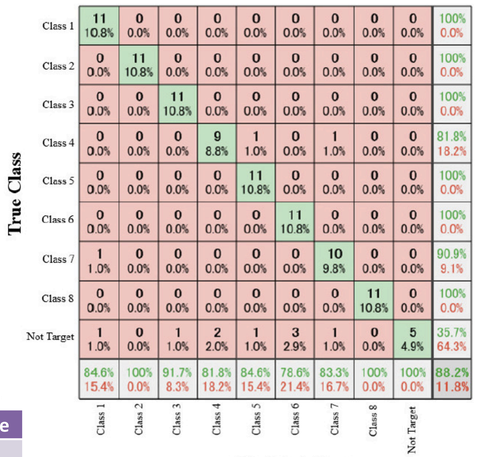
Sample images of Japanese fingerspelling dataset.



Confusion Matrix: Results of the spotting process.

Accuracies and recognition times of different methods.

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	86.9%	187.3 ms
TRCCA - CNN feat - OMSM(Proposed)	**88.2%**	91.2 ms



Confusion Matrix:
Results of the classification process

(4) Conclusion

- We proposed fingerspelling recognition framework based on a complementary combination of TRCCA and OMSM with CNN features.
- We confirmed that our two-steps process significantly outperforms conventional one-step methods in terms of classification accuracy and recognition time.

(5) References

[1]S. Tanaka, A. Okazaki, N. Kato, H. Hino and K. Fukui, Spotting fingerspelled 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.

[2]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.

[3] N. Sogi, T. Nakayama, and K. Fukui, A method based on convex cone model for image-set classification with cnn features, in 2018 *International Joint Conference on Neural Networks*, 2018, pp. 1-8.