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*

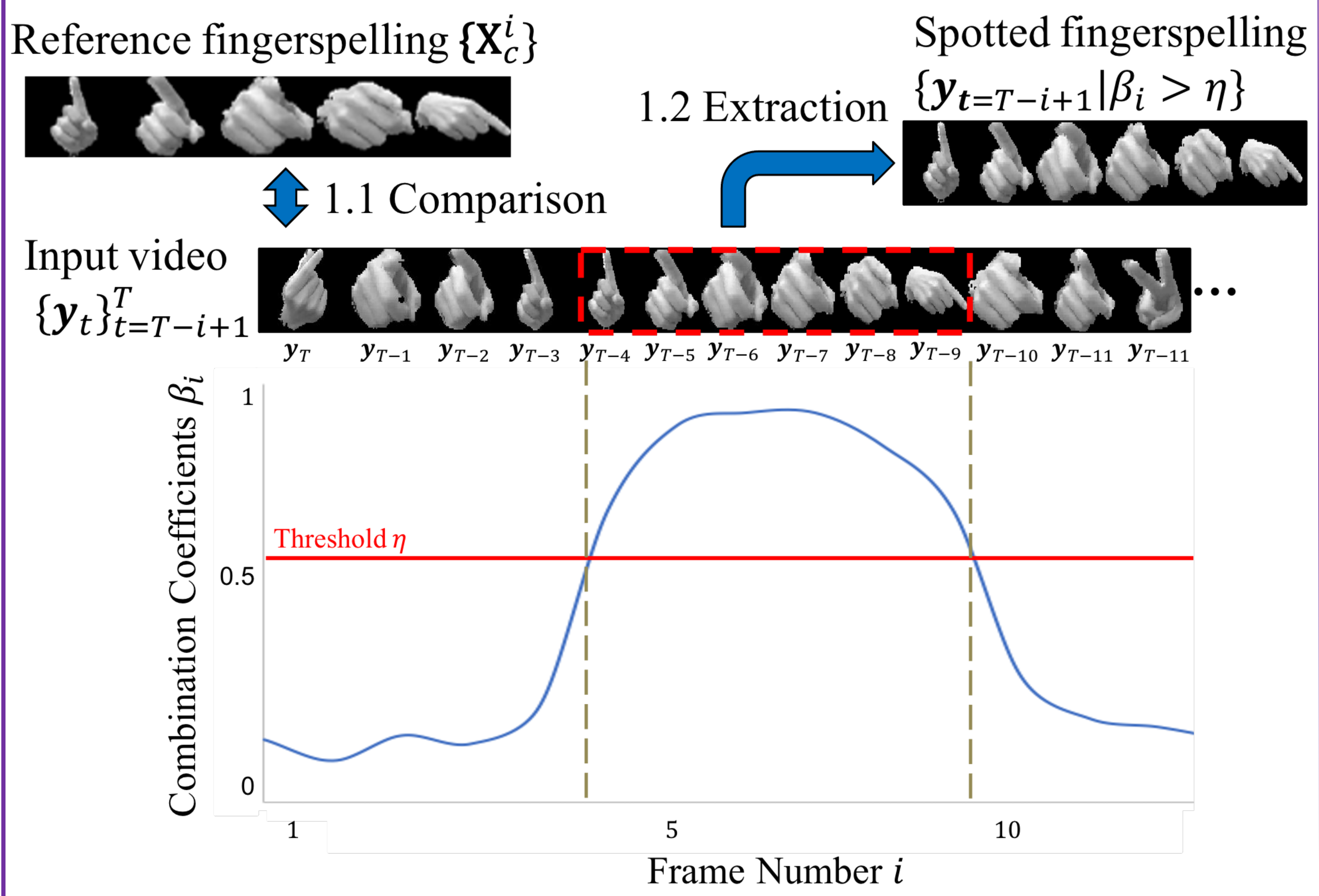
(1) Introduction

- Fingerspelling is a tool to express proper nouns and new technical terms that have not been yet defined in sign languages.
- **Main goal : Detect and recognize fingerspellings in an input video which is composed of fingerspellings and not fingerspelling signs.**
- **Basic idea :** Divide a whole process into two-steps cascade process:
  1. *Spotting* : 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.
- To efficiently incorporate two different information, we propose a fingerspelling classification framework based on three methods:
  1) *temporal regularized CCA (TRCCA)*[1] for spotting,
  2) *orthogonal mutual subspace method (OMSM)*[2] and 3) *CNN features*[3] for classification.

Japanese fingerspelling examples

(2) Proposed framework

Step1: Spotting process



- Fingerspelling images in an input video are detected by considering the temporal information using TRCCA.

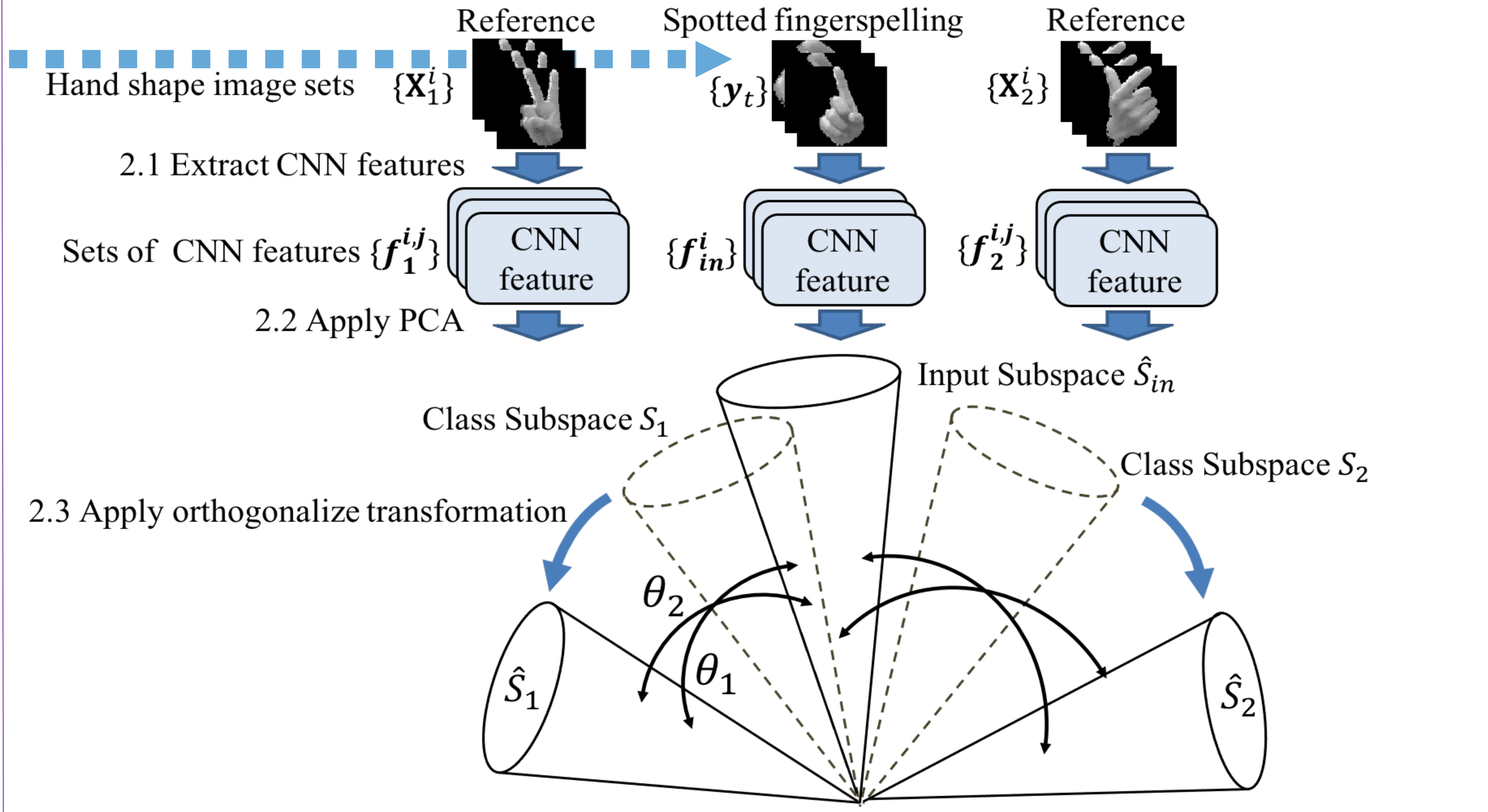
  TRCCA: Calculate similarity between two sequences considering temporal information.

- The detailed procedure

1.1 The input image sequence $\{y_t\}_{t=T-i+1}^{T}$ is compared with the reference fingerspelling $\{X_c^i\}$ by TRCCA. If the input sequence has high similarity with the reference fingerspelling, the input sequence is classified as a fingerspelling.

1.2 Fingerspelling images $\{y_{t=T-i+1} | \beta_i > \eta\}$ 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, then classify it by subspace similarity.

- The detailed procedure

2.1 CNN features $\{f_{in}^i\}$ and $\{f_c^{i,j}\}$ are extracted from $\{y_{t=T-i+1} | \beta_i > \eta\}$ and $\{X_c^i\}$.

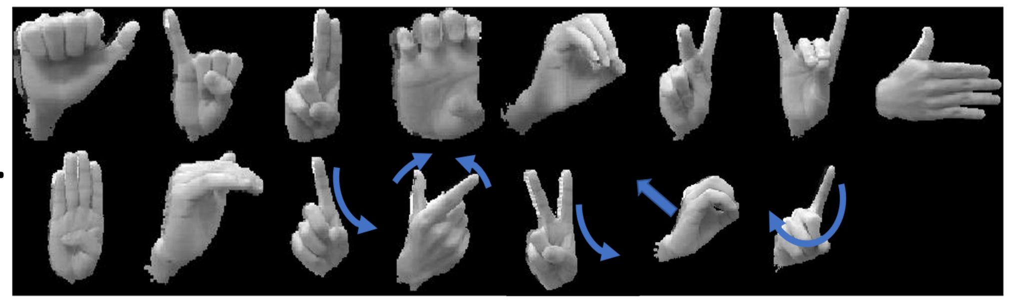
2.2 Each class subspace $\{S_c\}$ and an input subspace S_{in} are generated by applying PCA to the sets of CNN features.

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

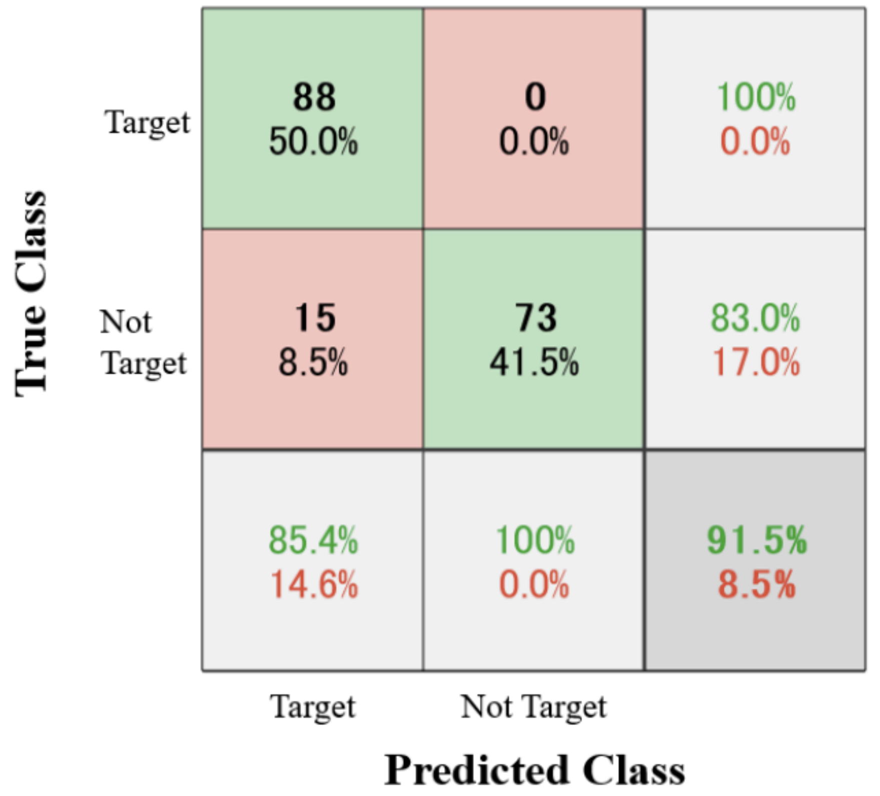
2.4 The spotted fingerspelling is classified based on similarities between the input subspace S_{in} and reference subspaces $\{\hat{S}_c\}$.

(3) Experiments

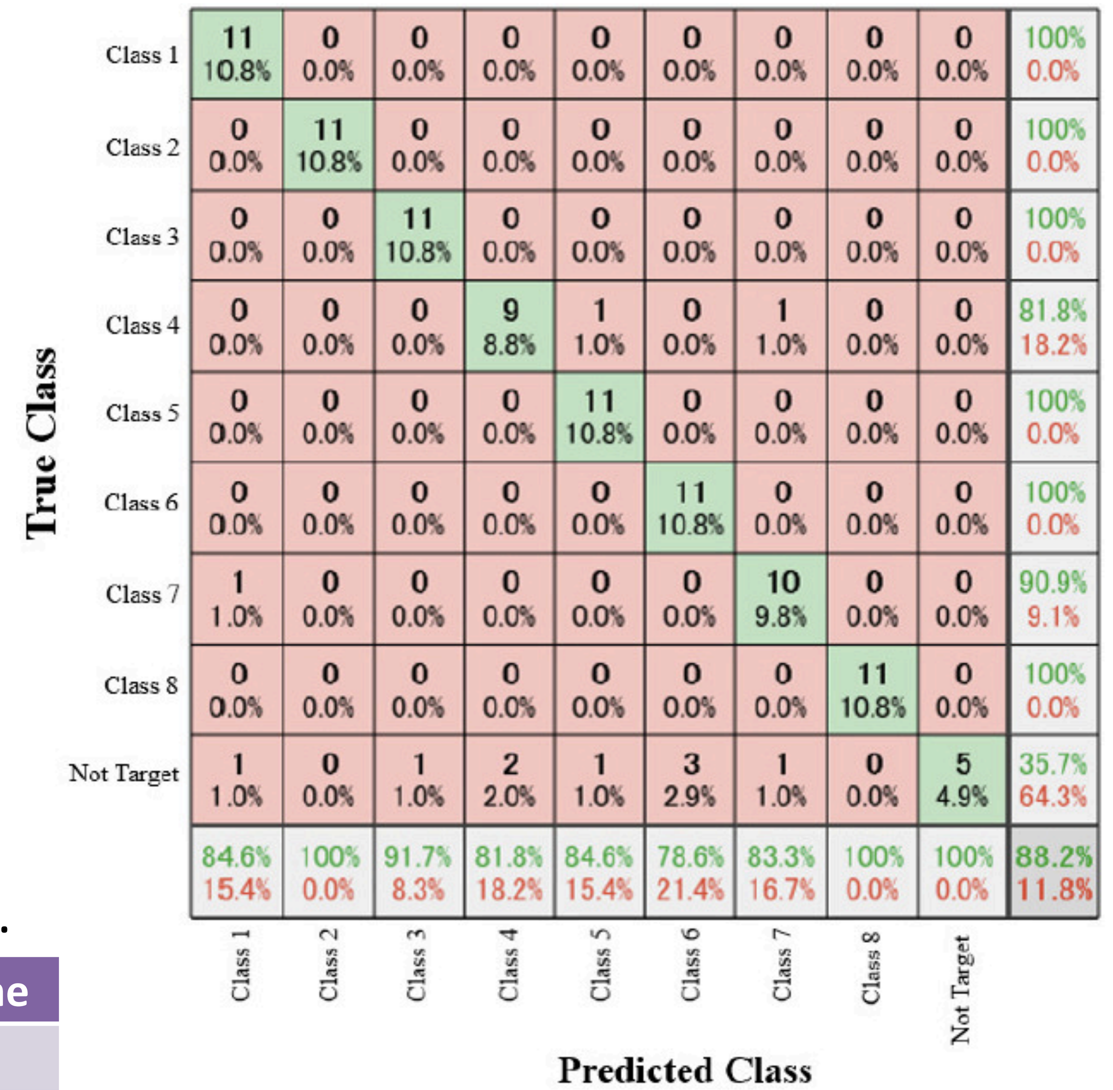
- Dataset:
  - We recorded 15 fingerspelling classes by a depth camera.
  - We synthesized an input video, which continuously inputs fingerspelling and not fingerspelling sequences alternately.
- Evaluation index:
  - Spotting performance, Classification accuracy, Recognition time.



Sample images of our fingerspelling dataset.



Confusion Matrix: Results of the spotting process.



Confusion Matrix: Results of the classification 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

(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.