Single-Frame Hand Gesture Recognition Using Color and Depth Kernel Descriptors

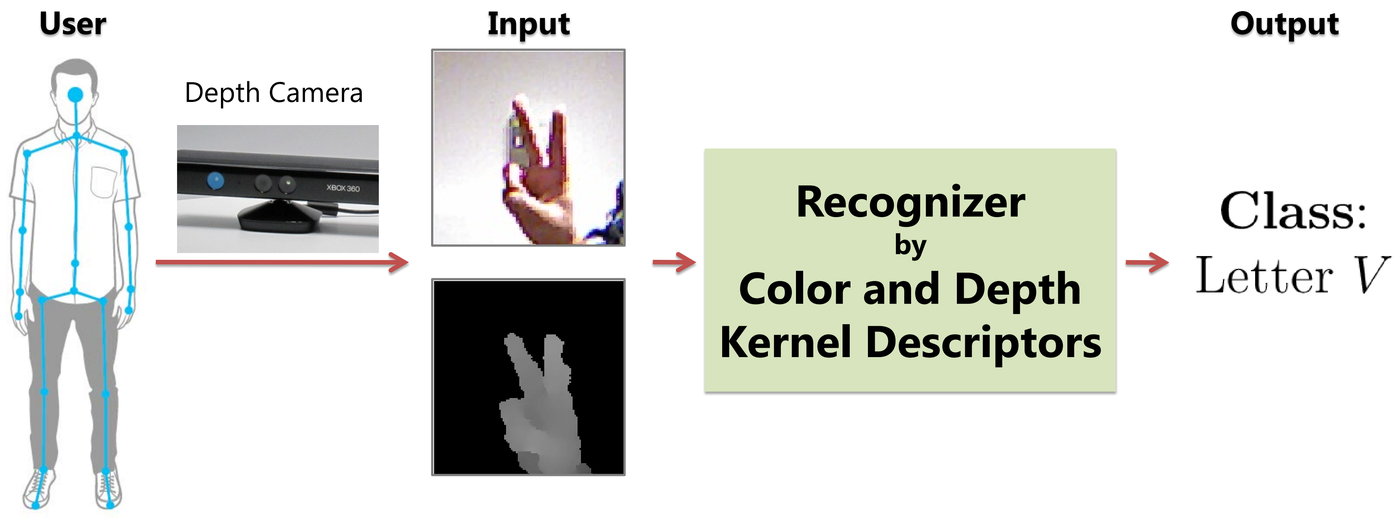
Xiaolong Zhu and *Kwan-Yee K. Wong*
Department of Computer Science, The University of Hong Kong. China

ICPR 2012 The 21st International Conference on Pattern Recognition

Introduction

- We present a flexible method for single frame hand gesture recognition by fusing information from color and depth images.
- Depth makes it possible to obtain a reliable bounding box of the hand regardless of light and distance changes.
- Our method extracts common patch-level features, and fuses them by means of kernel descriptors.

System Overview



Efficient Match Kernels for SVM

The **match kernel** of two images $\mathbf{I}_i$ and $\mathbf{I}_j$ for kernel SVM can be written as

$$K(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{|\mathbf{X}_i||\mathbf{X}_j|} \sum_{\mathbf{a}\in\mathbf{X}_i} \sum_{\mathbf{b}\in\mathbf{X}_j} k(\mathbf{a}, \mathbf{b}),$$

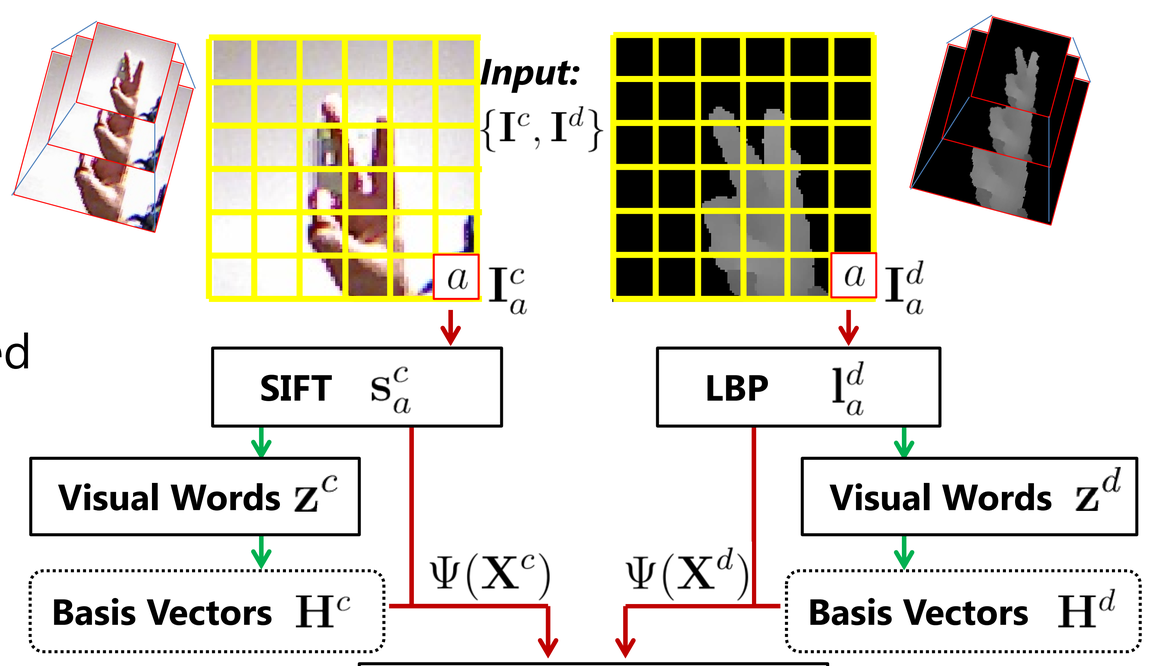
where $k(\mathbf{a}, \mathbf{b}) = \phi(\mathbf{a})^\top \phi(\mathbf{b})$.

The infinite-dimensional kernel vector $\phi(\mathbf{a})$ is approximated by a D-dimensional vector $\psi(\mathbf{a}) = \mathbf{H}\mathbf{v}_\mathbf{a}$.

We construct $\mathbf{H}$ by extracting the visual words of the patch, hence

$$\begin{aligned} k(\mathbf{a}, \mathbf{b}) &\doteq \psi(\mathbf{a})^\top \psi(\mathbf{b}) \\ &= (\mathbf{H}\mathbf{v}_\mathbf{a}^\star)^\top \mathbf{H}\mathbf{v}_\mathbf{b}^\star \\ &= (\mathbf{H}^\top \phi(\mathbf{a}))^\top \cdot (\mathbf{H}^\top\mathbf{H})^{-1} \cdot (\mathbf{H}^\top \phi(\mathbf{b})) \\ &= \mathbf{k}_\mathbf{Z}(\mathbf{a})^\top \cdot \mathbf{K}_\mathbf{ZZ}^{-1} \cdot \mathbf{k}_\mathbf{Z}(\mathbf{b}) \end{aligned}$$

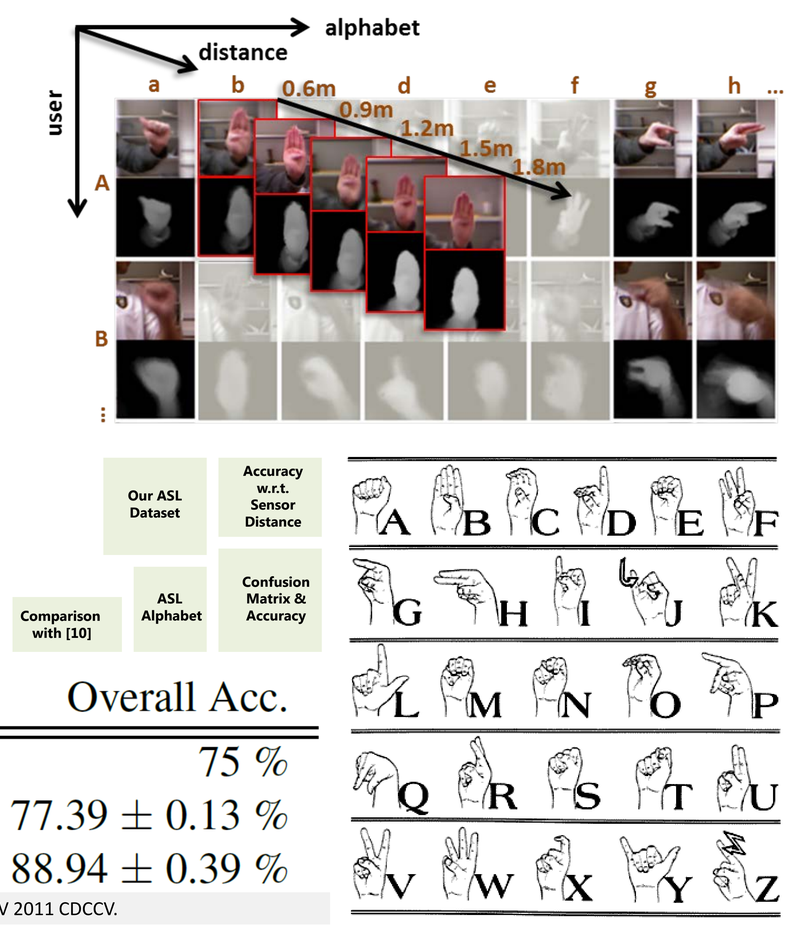
Color and depth descriptors are **concatenated** into an image-level kernel feature vector.

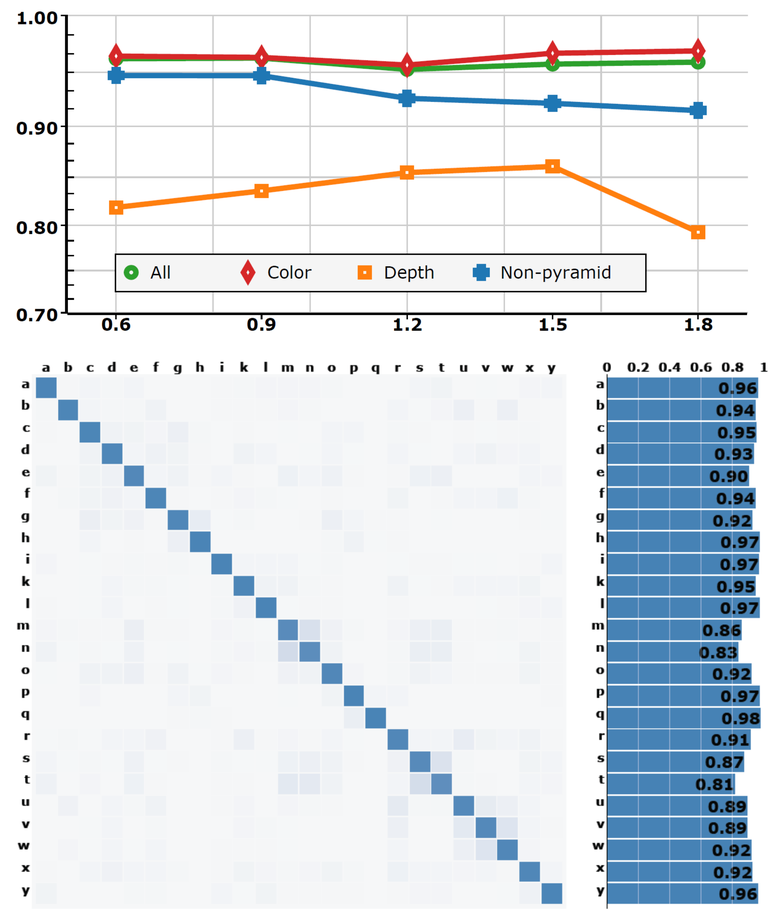


Experimental Result

We have evaluated our method on two datasets: ASL FingerSpelling Dataset [10] and our own dataset.

- *FingerSpelling dataset:* 500 color images for each of 5 users are obtained for each sign.
- *Our dataset:* It consists of 24 static signs of 5 users at 5 different distances.





Method	#training samples	Overall Acc.
Pugeault [10]	1250	75 %
Our Approach (NP)	40	77.39 ± 0.13 %
Our Approach (IP)	40	88.94 ± 0.39 %

[10] N. Pugeault and R. Bowden. *Spelling It Out : Real-Time ASL FingerSpelling Recognition*. ICCV 2011 CDCCV.

Take-home Message

1. Hand can be easily tracked by depth camera, i.e., we can easily obtain a reliable bounding box of the hand.
2. Patch-based approaches can be used here, so we can achieve high accuracy with efficient algorithm.
3. The method works in a normal indoor setting (0.6m~2m).

Contact Us

Xiaolong ZHU:
xlzhu@cs.hku.hk
Dr. Kwan-Yee K. Wong:
kykwong@cs.hku.hk