Realizing Speech to Gesture Conversion by Keyword Spotting

Na Zhao, Hongwu Yang

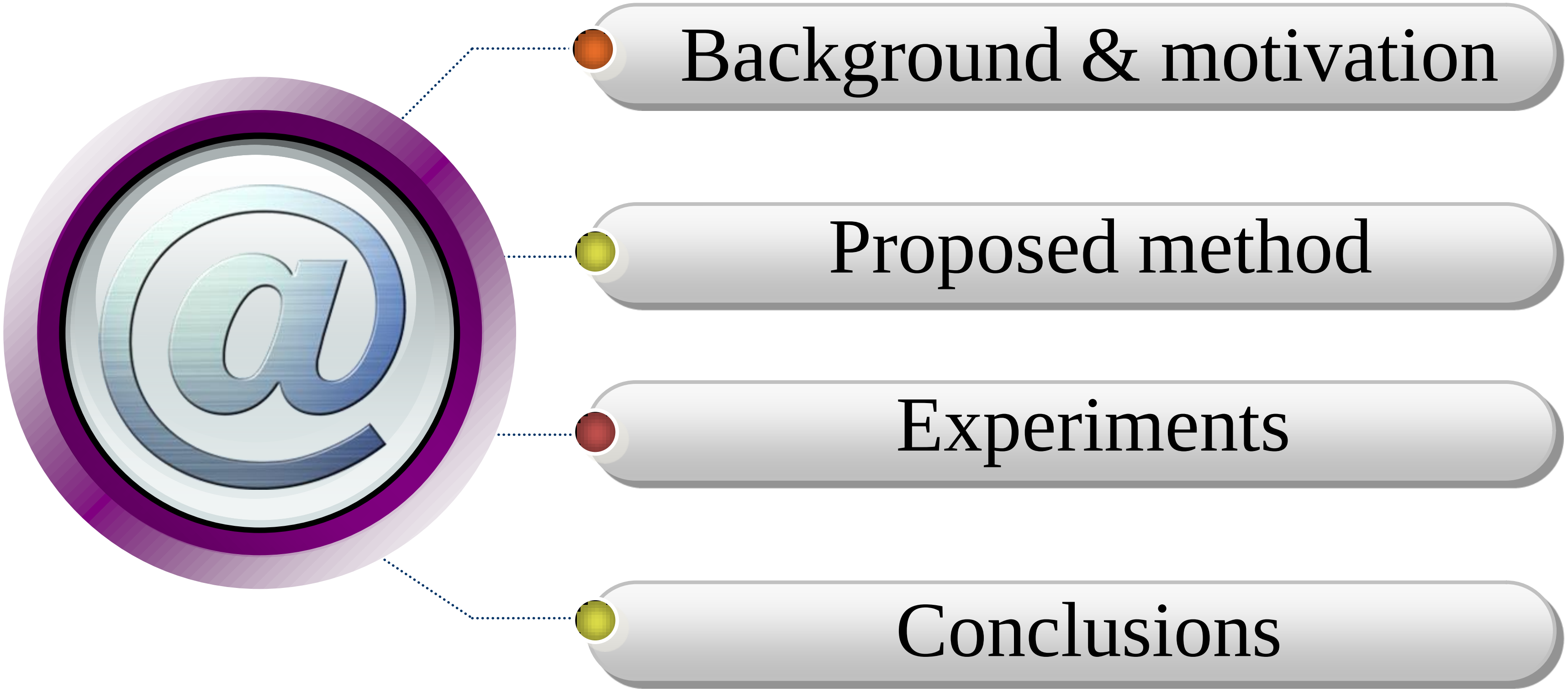
College of Physics and Electronic Engineering

Northwest Normal University

Lanzhou

西北師範大學

Outline



Background & motivation

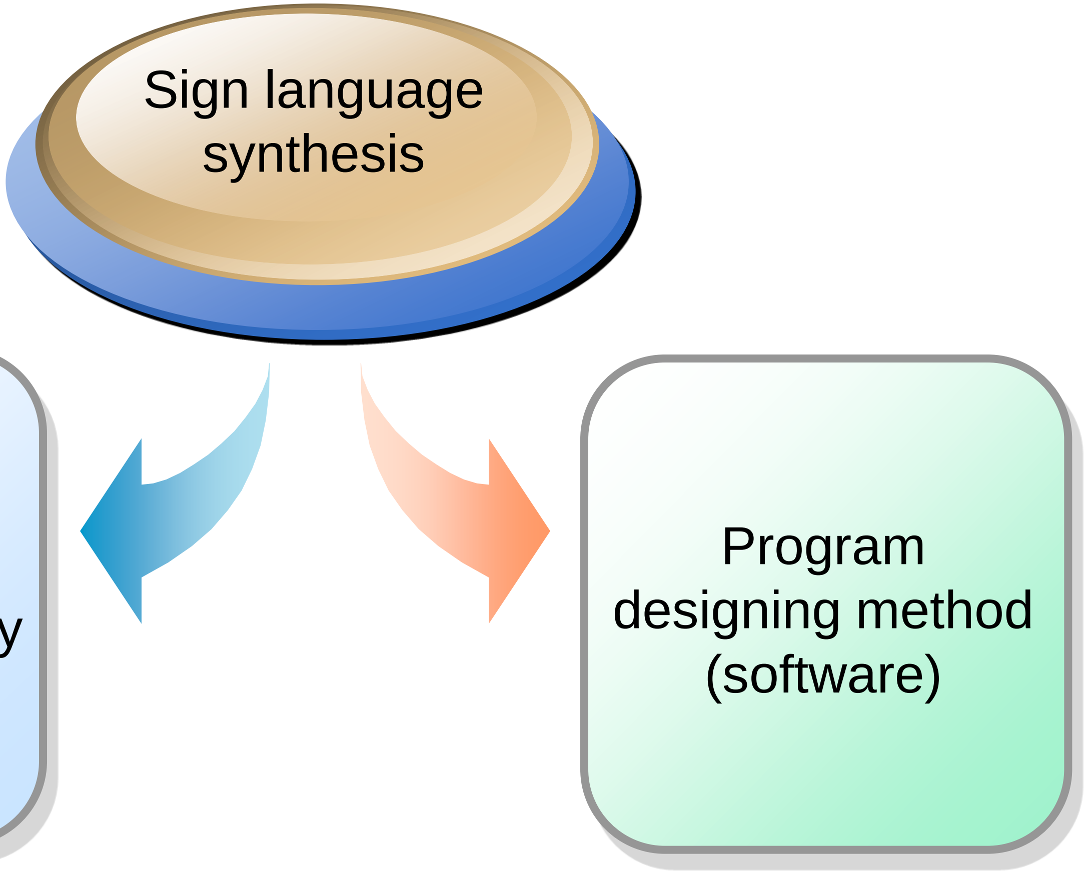
Proposed method

Experiments

Conclusions

Background & motivation

Background



Sign language synthesis

Motion tracking method (data acquisition by hardware)

Program designing method (software)

Motivation

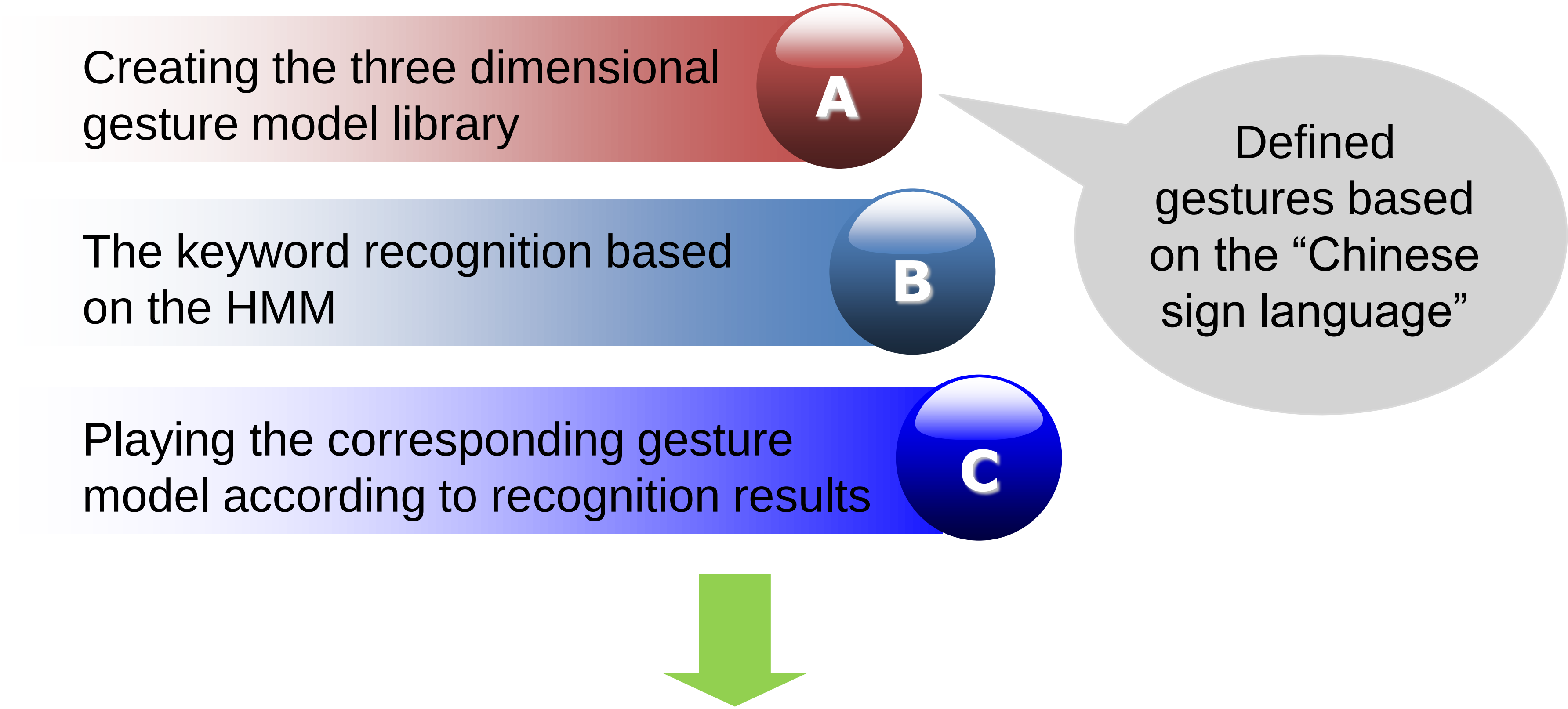
the Hidden Markov Model based keyword spotting

Lacking of study on speech to sign language conversion

To promote and apply difficultly in real life.

To satisfy the need of communication between normal and speech-impaired people, the paper realizes a speech-to-gesture conversion system.

Proposed method



A Creating the three dimensional gesture model library

B The keyword recognition based on the HMM

C Playing the corresponding gesture model according to recognition results

Defined gestures based on the “Chinese sign language”

The keyword spotting system is combined with the sign language synthesis to realize a speech to gesture conversion system.

System framework

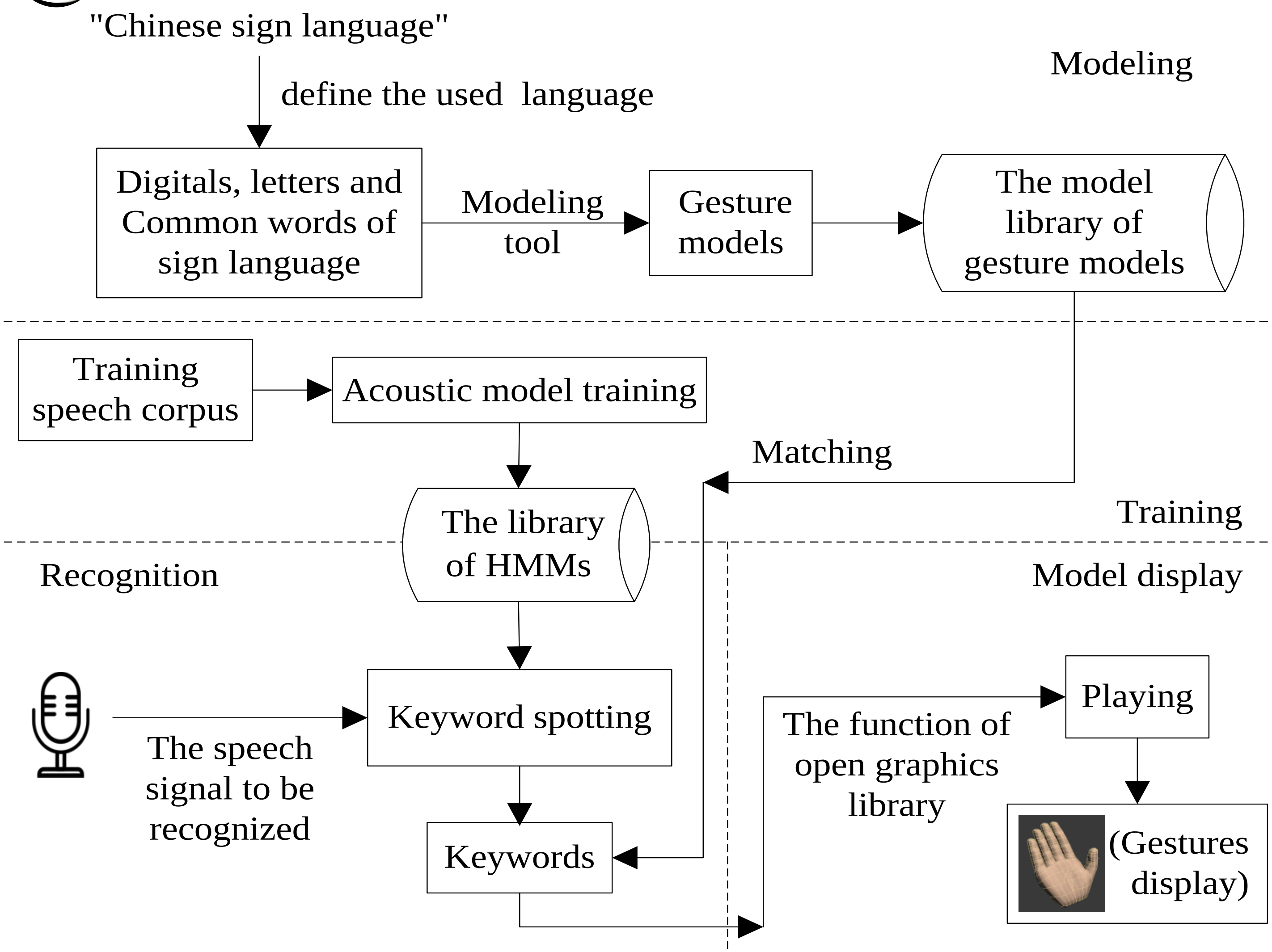


Fig.1 A speech-to-gesture conversion system

Creating gesture model library

1

- Analyzing the physical structure of hand model

2

- Calculating the similarity of hand-shape
- Calculating the free bending angle of knuckle

3

- Examples of gestures model library

➢ Analyzing the physical structure of hand model

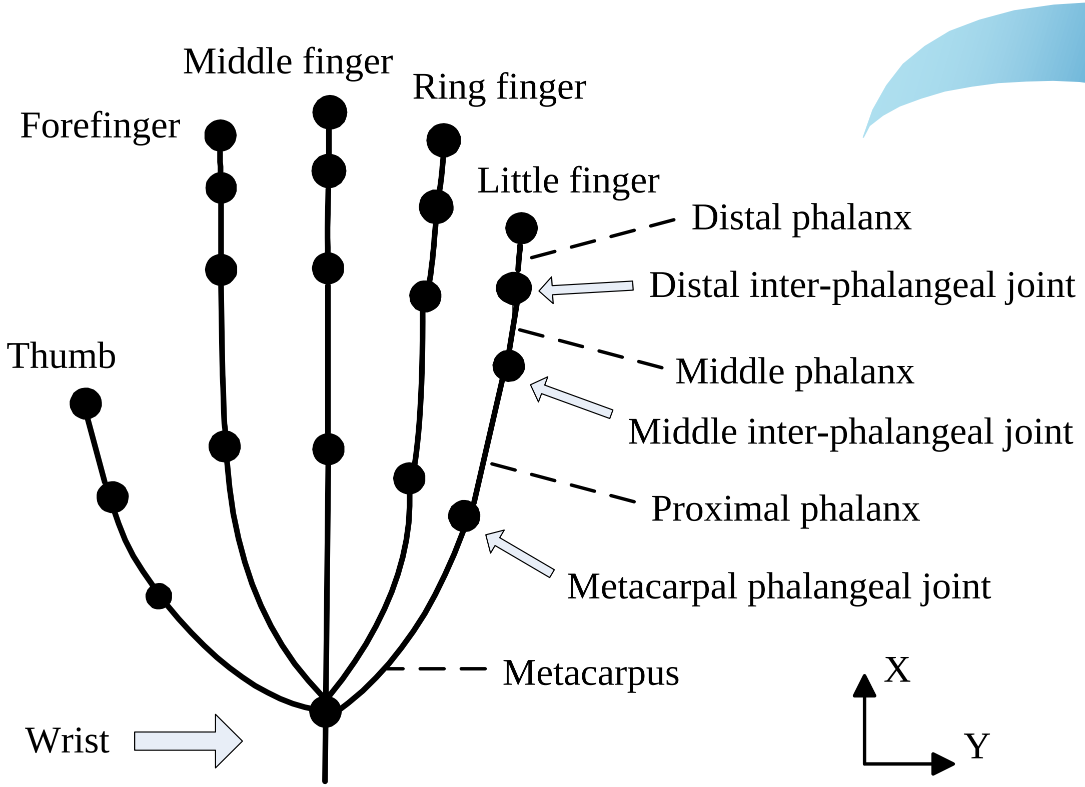


Fig.2 Extracted point and line model as well as the names of each joint

- ✓ Composed of four known segments except thumb
- ✓ The active plane of knuckle is perpendicu-lar to the center plane of palm

Calculating the similarity of hand-shape

The similarity between adjacent gesture is compared with the weighted Euclidean distance.

$$S_{AB} = \sqrt{\sum_{k=0}^{9}\left[\left(\theta_{AK}-\theta_{BK}\right)^{2} \bullet W_{k}\right]}$$

Calculating the free bending angle of knuckle

$$\begin{cases} L=L_{M}+L_{P}+L_{D} & (1) \\ H=L_{M}*\sin\theta_{1}+L_{P}*\sin(\theta_{1}+\theta_{2}) \\ \quad +L_{D}*\sin(\theta_{1}+\theta_{2}+\theta_{3}) & (2) \end{cases}$$

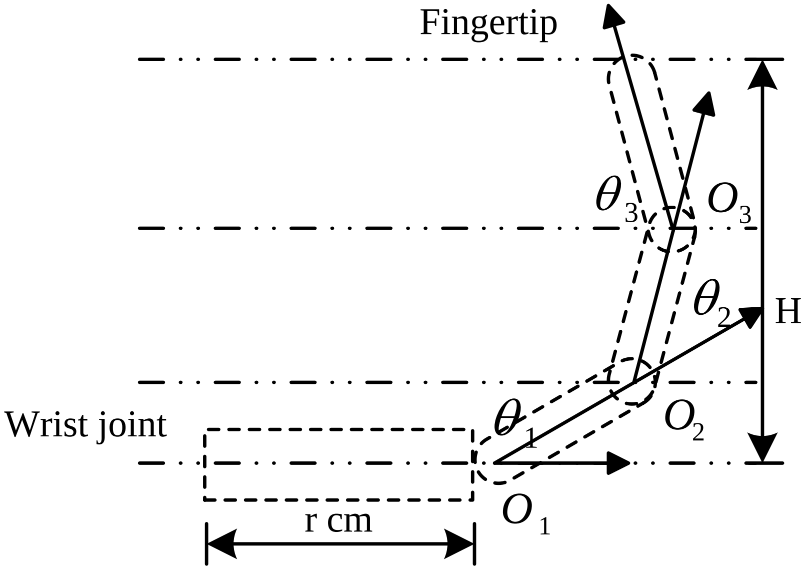
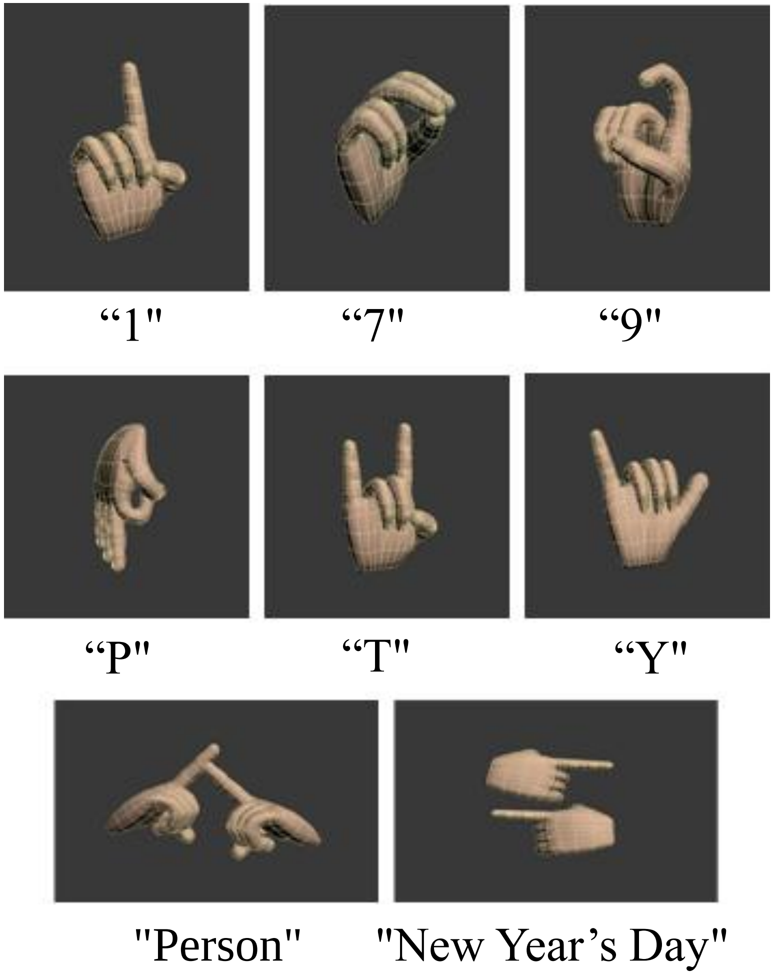


Fig.3 Calculation model of finger knuckles

Examples of gestures model library



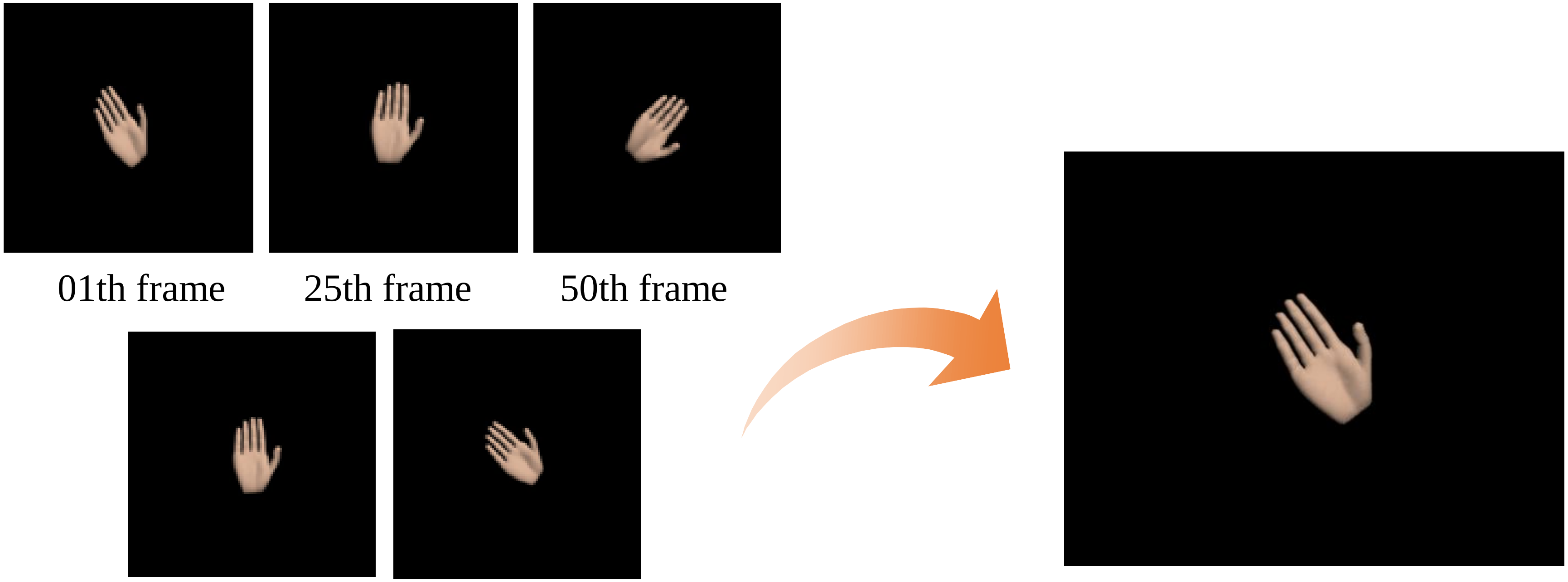
“1" “7" “9"

“P" “T" “Y"

"Person" "New Year’s Day"

Fig.4 Examples of some static gesture models

Examples of gestures model library


01th frame

25th frame

50th frame

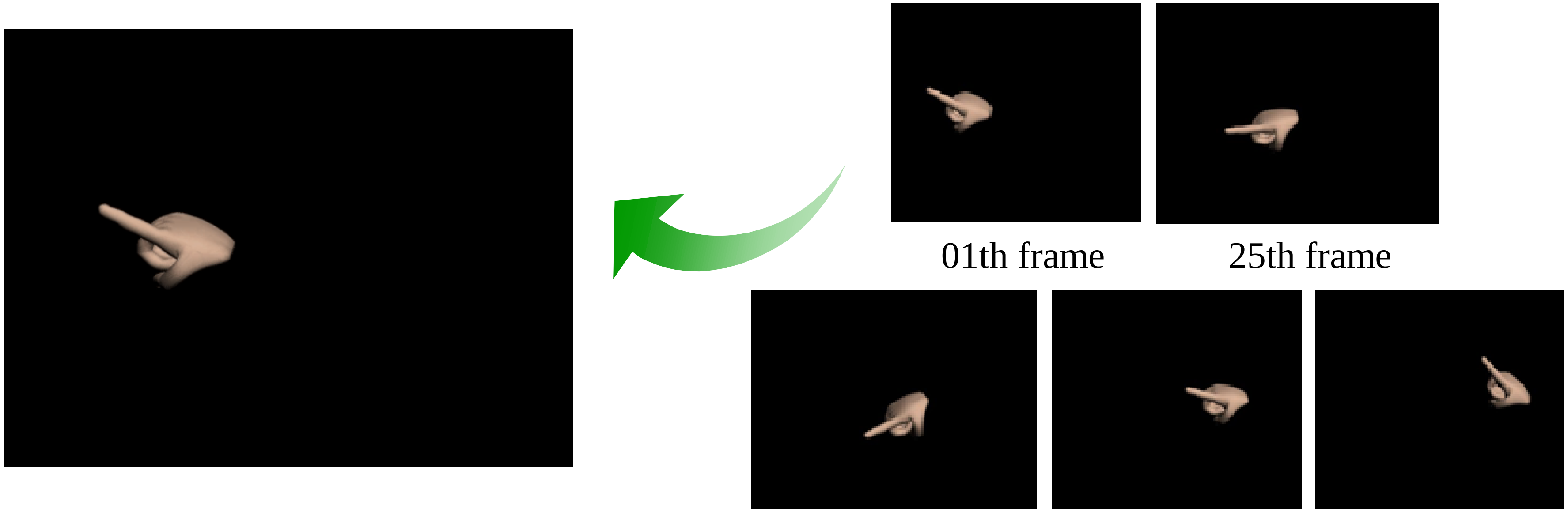
75th frame

100th frame

(a) The key frames of "no"

Fig.5 Examples of the key frames of some dynamic gesture models

Examples of gestures model library



01th frame 25th frame

50th frame 75th frame 100th frame

(b) The key frames of "approve"

Fig.5 Examples of the key frames of some dynamic gesture models

Experiments

Keyword spotting

A total of 592 sentences are recorded under the office environment by the eight speakers including four women and four men.

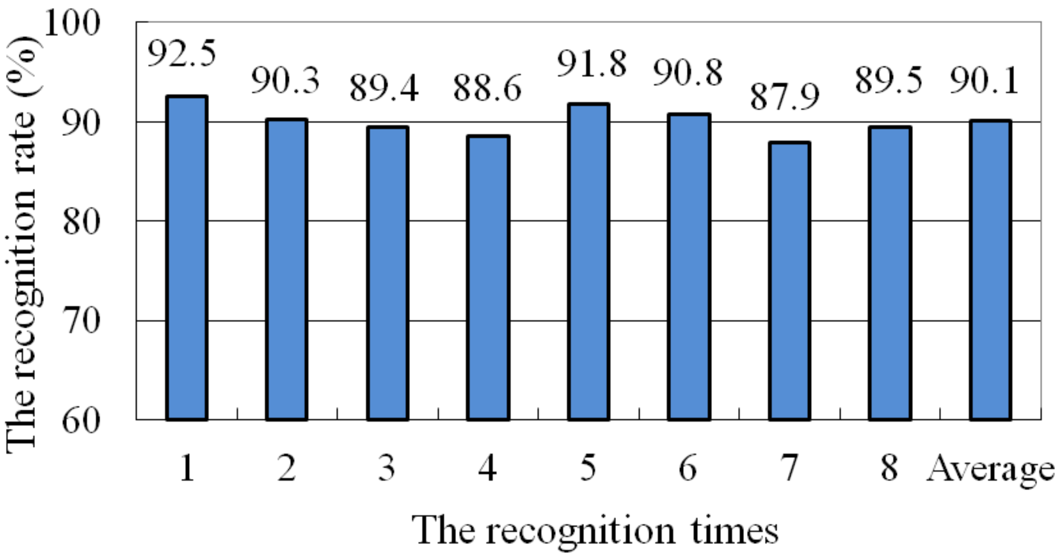


Fig.6 The results of keyword spotting

The accuracy evaluation of converted gestures

To evaluate whether the converted gestures can accurately express the meaning of key words.

Table 1.The standard of MOS evaluation

Score	The evaluating standards
0-1	Bad, the very bad match
1-2	Poor, barely matching
2-3	Medium, accepting the match
3-4	Good, willing to accept the match
4-5	Excellent, the very natural match

Table 2.The results of MOS evaluation

The average value of MOS	The standard deviation
4.4	0.3

Conclusions

System

Realizing speech-to-gesture conversion

Further

✓ Enlarging the gesture model library

✓ Improve the keyword recognition rate

Results

Recognition rate 90.1%

The average value of MOS 4.4

The standard deviation 0.3

References

[1] S Wang, L Wang, D Kong, “Prosodic Chinese Sign Language Synthesis Driven by Speech,” *Proceedings of the 5th International Conference on Computational and Information Sciences (ICCIS)*, 2013, pp. 1951–1954.

[2] D Song, X Wang, X Xu, “Chinese Sign Language Synthesis System on Mobile Device,” *Procedia Engineering*, vol. 29, pp. 986–992, 2012.

[3] M.E. Sargin, O Aran, A Karpov, et al, “Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis,” *Proceedings of the IEEE International Conference on Multimedia and Expo*. pp. 893–896, 2006.

[4] A Fischer, V Frinken, H Bunke, et al, “Improving HMM-Based Keyword Spotting with Character Language Models,” *Proceedings of the 12th IEEE International Conference on Document Analysis and Recognition*, 2013, pp. 506–510.

[5] L Jiang, Y Liu, D Yang, et al, “Analysis of Human Hand Posture Reconstruction under Constraint and Non-constraint Wrist Position,” *Proceedings of the International Conference on Intelligent Robotics and Applications (ICIRA)*, 2015, pp. 269–281.

[6] M.K. Bhuyan, C Narra, D.S. Chandra, “Hand gesture animation by key frame extraction,” *Proceedings of the IEEE International Conference on Image Information Processing*, 2011, pp. 1–6.

[7] N Sun, T Ayabe, K Okumura, “An Animation Engine with the Cubic Spline Interpolation,” *Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP)*, 2008, pp. 1185–1192.

[8] W Shen, W Zeng, “Research of VR modeling technology based on VRML and 3DS MAX,” *Proceedings of the IEEE International Conference on Computer Science and Network Technology (ICCSNT)*, 2011, pp. 487–490.

[9] N Shimada, K Kimura, Y Shirai, et al, “Hand posture estimation by combining 2-D appearance-based and 3-D model-based approaches,” *Proceedings of the 15th IEEE International Conference on Pattern Recognition*, 2000, pp. 705–708.

[10] R.K. Dubey, A Kumar, “Comparison of subjective and objective speech quality assessment for different degradation/noise conditions,” *Proceedings of the IEEE International Conference on Signal Processing and Communication (ICSPC)*, 2015, pp. 261–266.

Thank You !

Q&A