Towards Sign Language Recognition based on Body-Parts Relations.

Marc Martínez Camarena, José Oramas M., Tinne Tuytelaars
KU Leuven
September 28th 2015

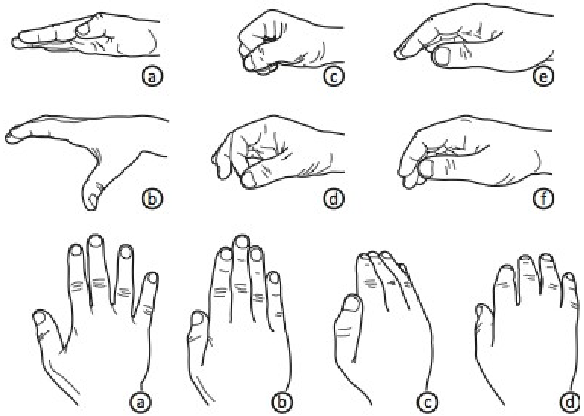
Introduction

Sign Language

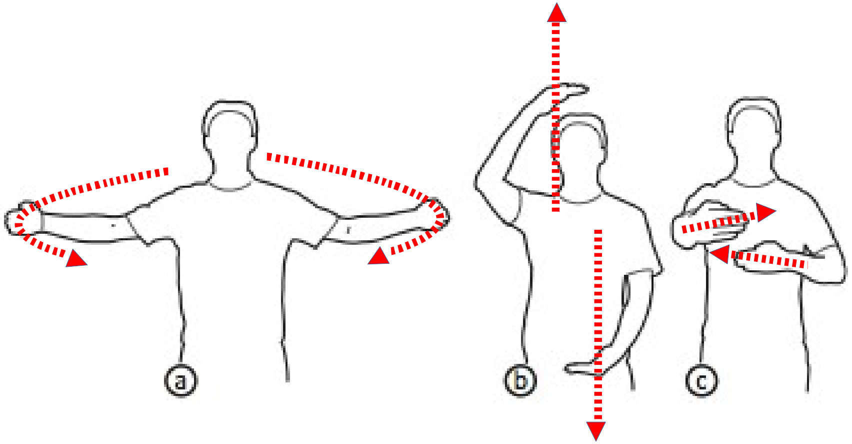
- Hearing-impaired community.
- Low teacher/student ratio.

Introduction

Some terminology



Hand Postures



Hand Gestures

*Images taken from Holz and Wilson, CHI'11.

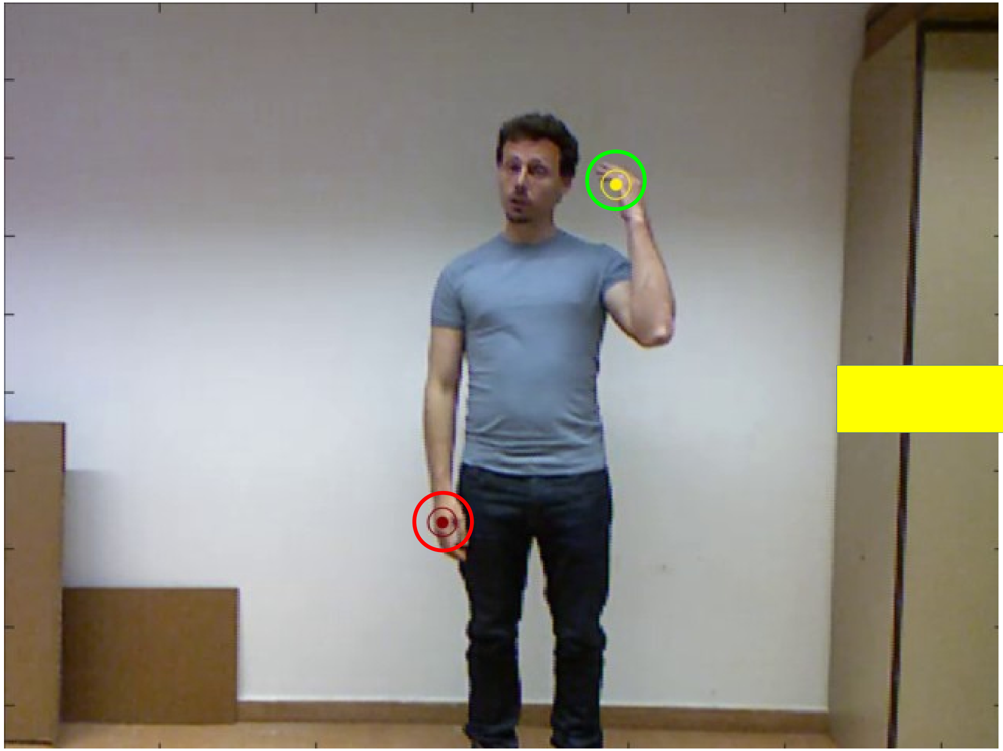
Problem Statement

Focus from:

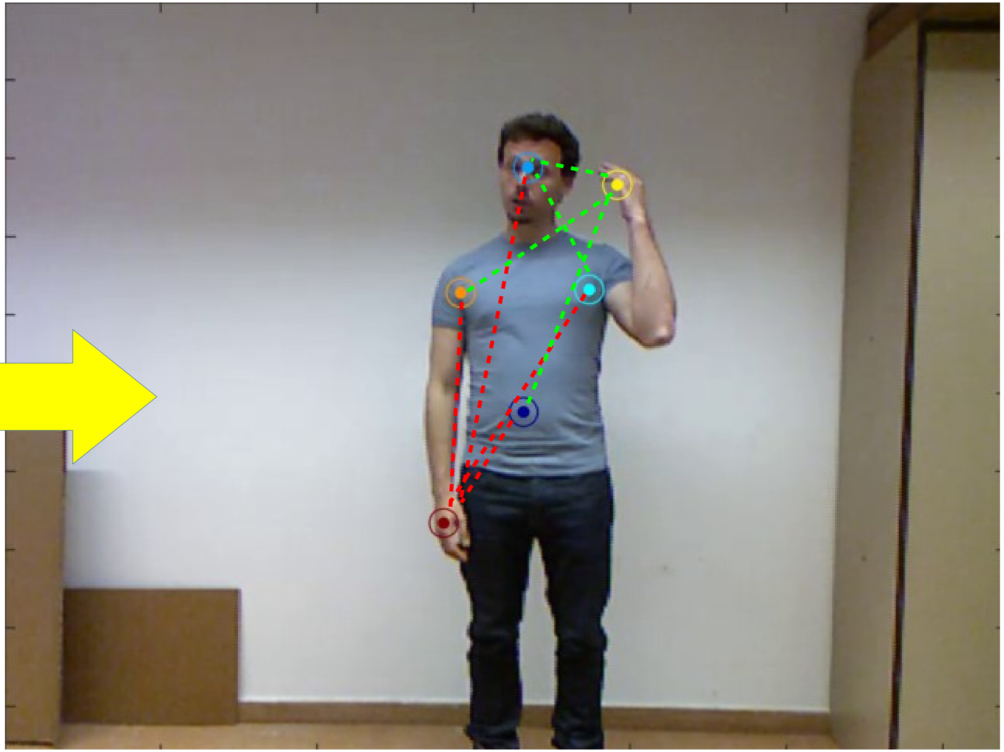
Global location of the hands

Problem Statement

Focus from:



Global location of the hands

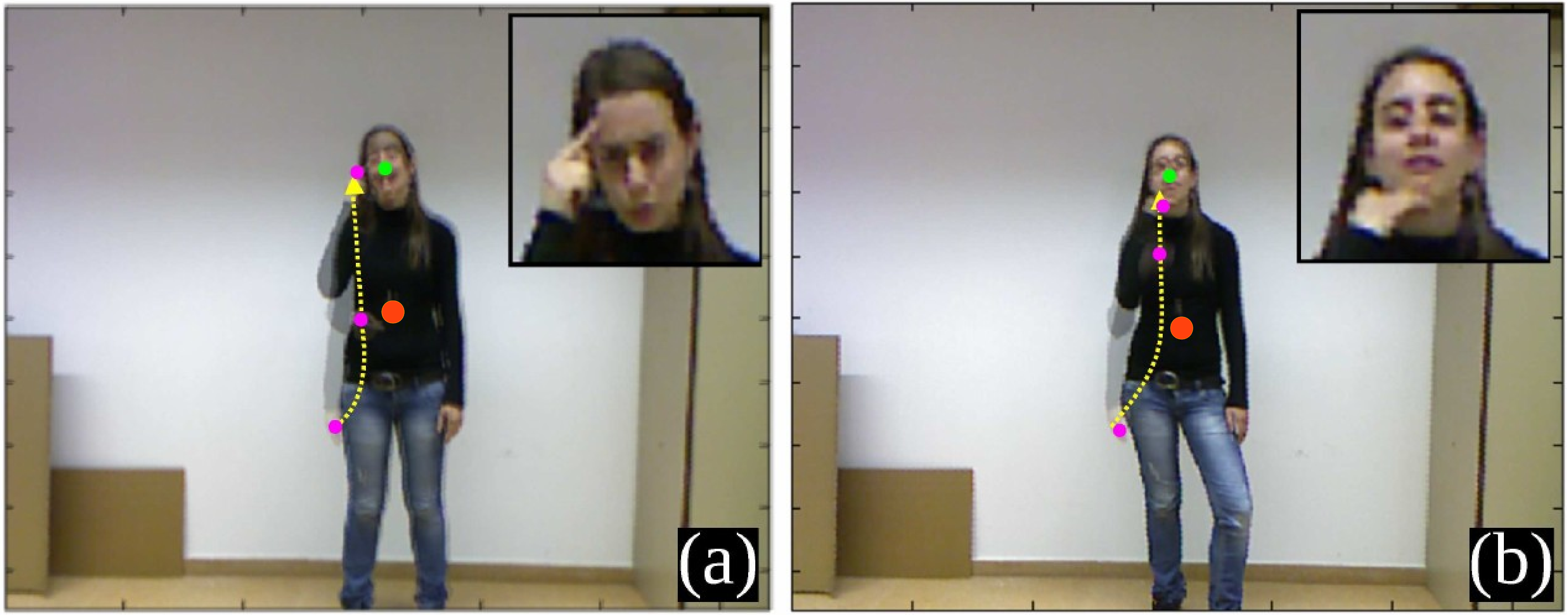


location of the hands wrt. to other parts of the body

Problem Statement

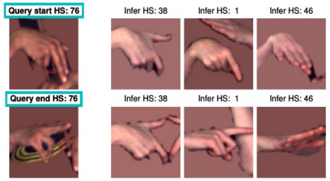
Motivation

- Signs with similar hand trajectories can be distinguished by the temporal relations between body parts.

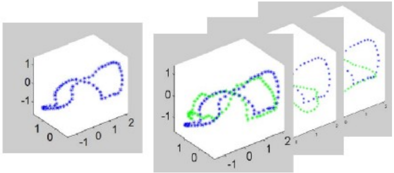


Related Work

- **Focus on Hand Postures & Trajectories**
  - Chai et al., FG'13.
  - Tanghali et al., CVPR'11.
  - Pfister et al., ECCV'14.

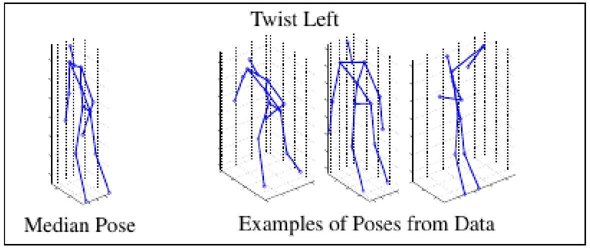


Tanghali et al., CVPR'11.



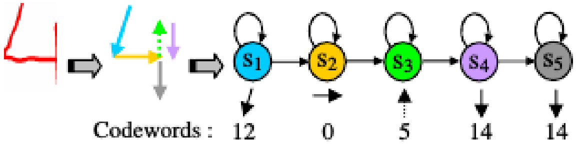
Chai et al., FG'13.

- **Exploiting Skeleton Representations**
  - Wu et al., ICMI'13.
  - Papadopoulos et al., MM'14.
  - Ellis et al., IJCV'13.
  - Hussein et al., IJCAI'13.



Ellis et al., IJCV'13.

- **Modeling Dynamics of Hand Gestures**
  - **DTW:** Rabiner & Juang, 93
  - **HMM:** Elmezain et al., ICPR'08.
    Papadopoulos et al., MM'14.



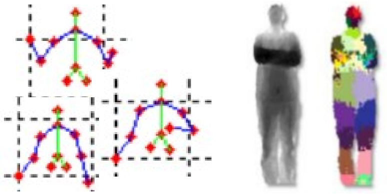
Elmezain et al., ICPR'08.

Proposed Method

In a nutshell...

Proposed Method

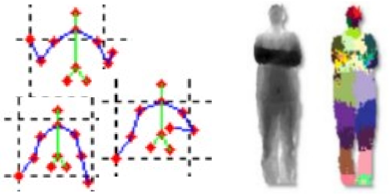
In a nutshell...



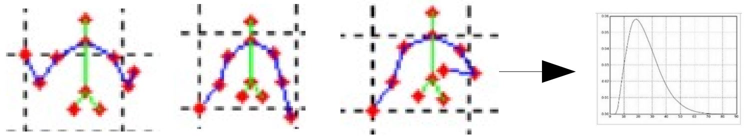
Skeleton estimation
(Shotton et al., CVPR'11)

Proposed Method

In a nutshell...



Skeleton estimation
(Shotton et al., CVPR'11)

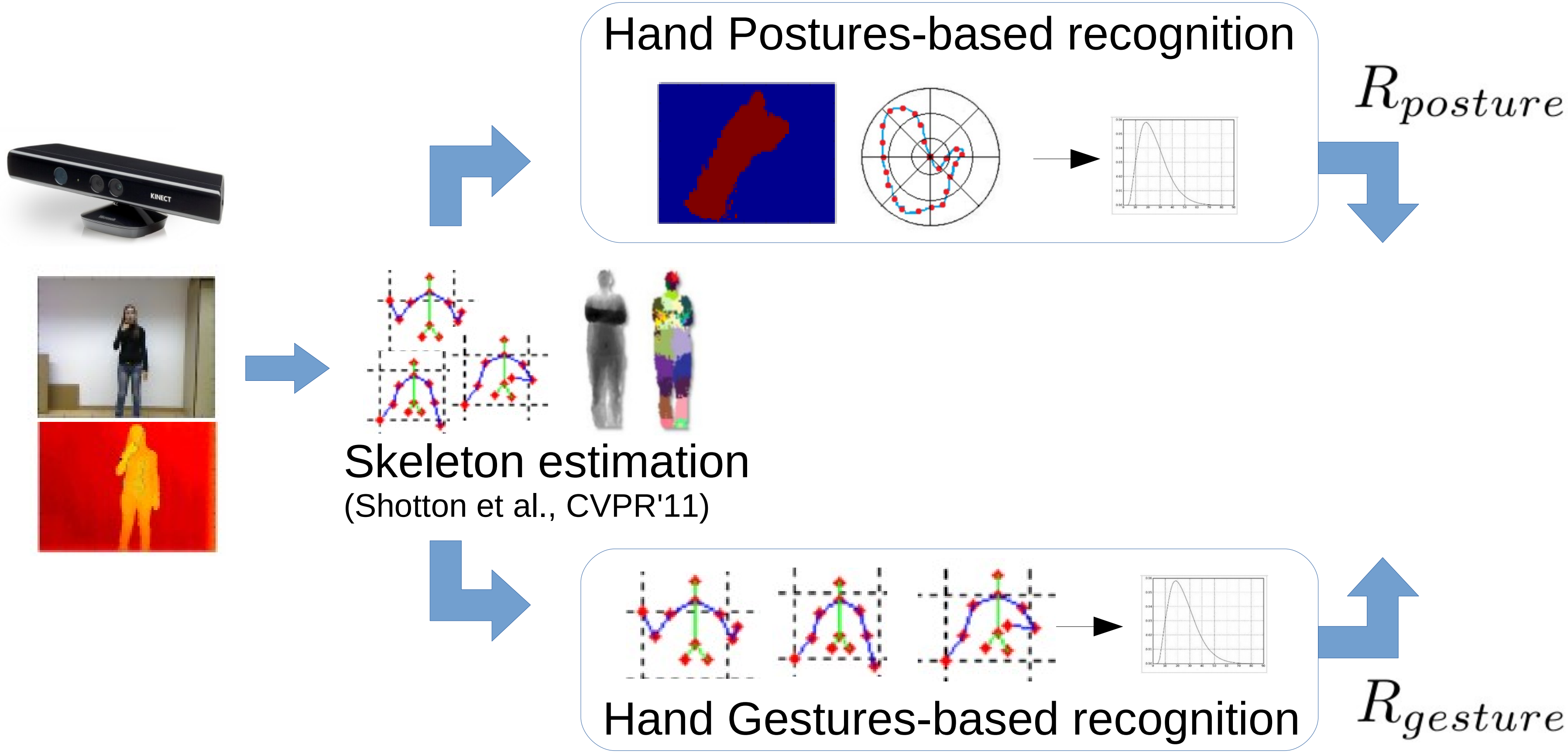


Hand Gestures-based recognition

$R_{gesture}$

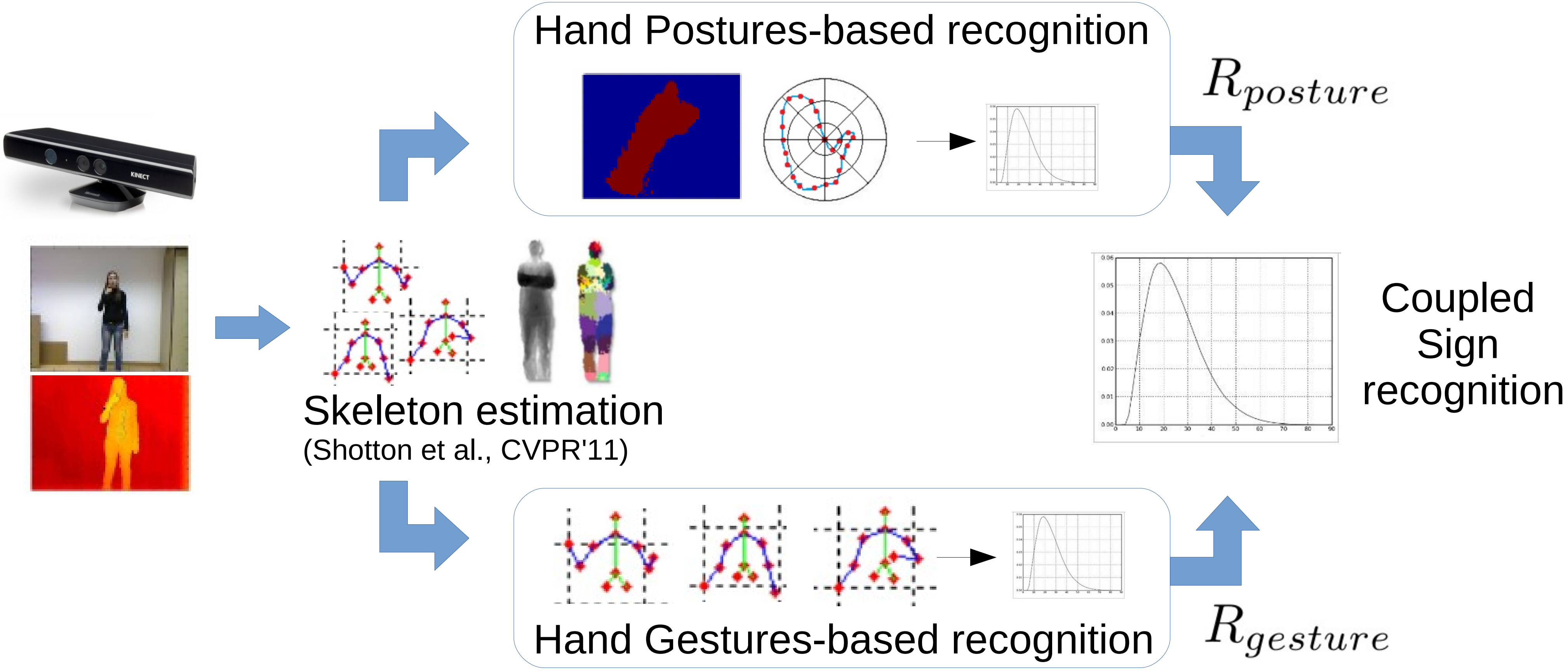
Proposed Method

In a nutshell...



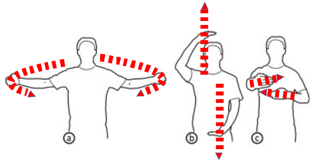
Proposed Method

In a nutshell...



Proposed Method

Hand Gestures-based Sign Recognition



- Given a set of body joints

$$J = \{j_1, j_2, ..., j_{11}\}$$

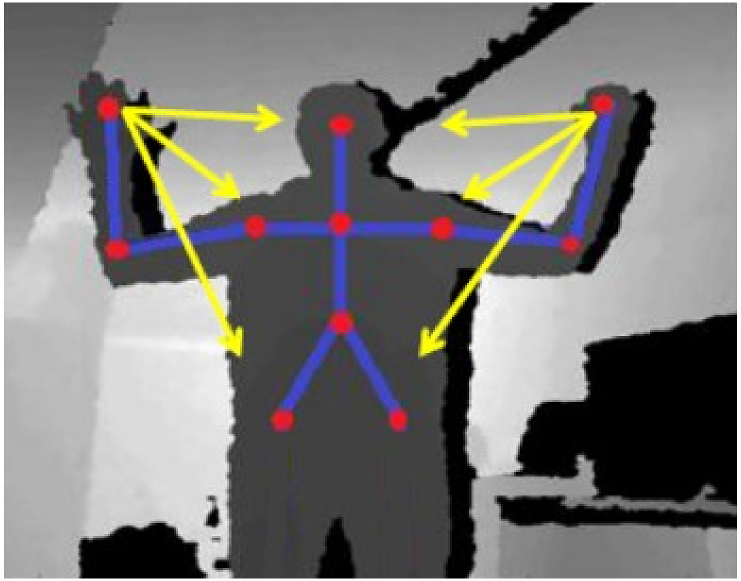
- We define the descriptor *(RBPD)* as:

$$RBPD = [\delta_1, \delta_2, ..., \delta_m]$$

where:

$$\delta_i = (j_i - j_h)$$

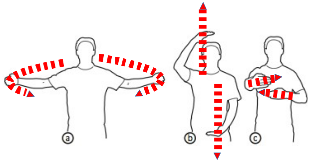
j_h : hand joint.



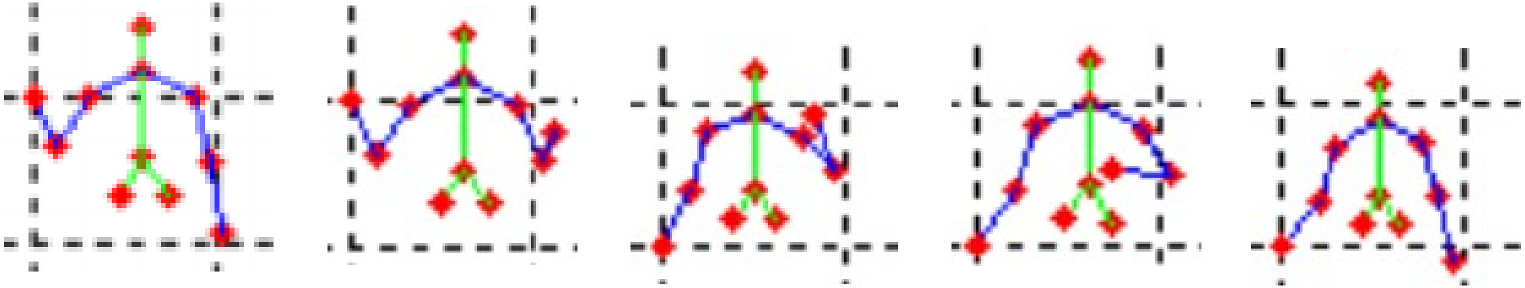
* This procedure is performed for each frame of the video

Proposed Method

Hand Gestures-based Sign Recognition

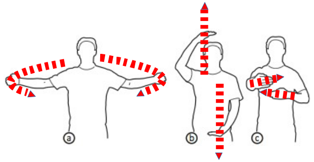


- Data re-encoding via K-Means

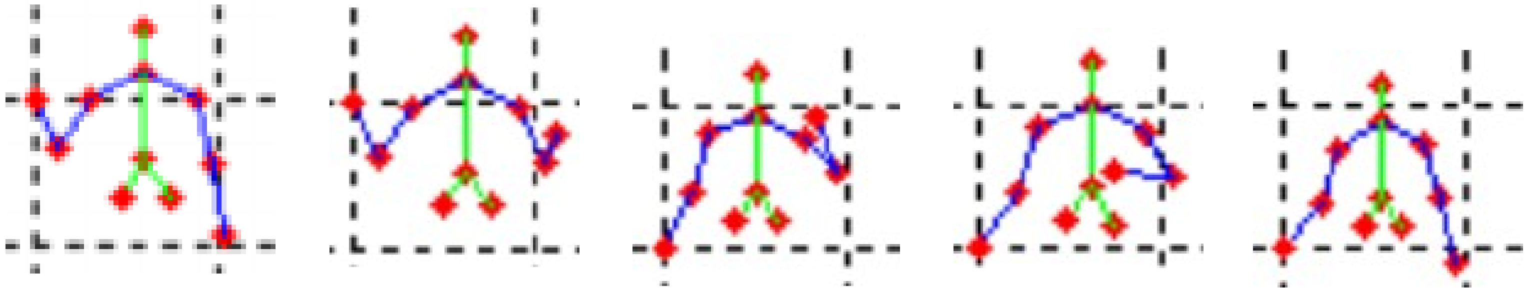


Proposed Method

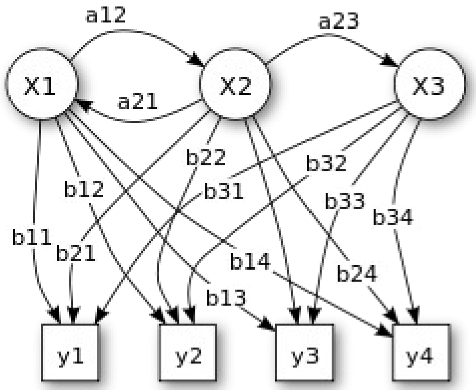
Hand Gestures-based Sign Recognition



- Data re-encoding via K-Means



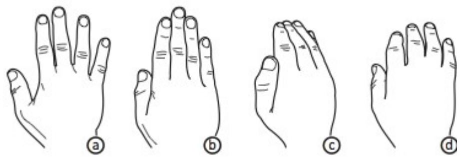
- Gesture modeling via Hidden Markov Models (HMMs)



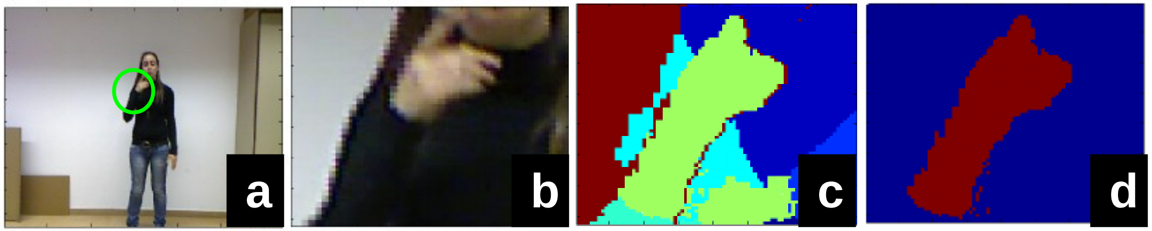
- An HMM is trained for each sign class.
- Training observations → gesture(sequence of centers).
- # observation symbols == # clusters (K).

Proposed Method

Hand Postures-based Sign Recognition

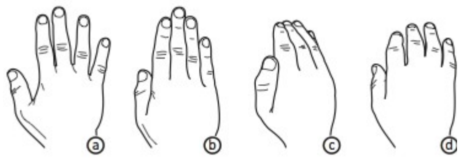


- Hand segmentation

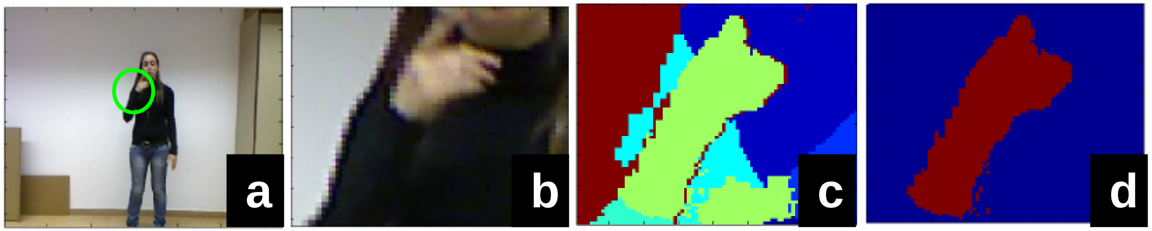


Proposed Method

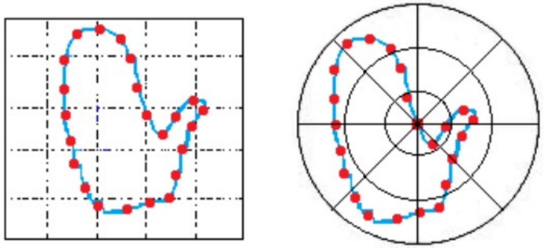
Hand Postures-based Sign Recognition



- Hand segmentation



- Posture description



- Shape context descriptors
  (Belongie et al, TPAMI'02)
- Bag of words (BoW) encoding
  (Salton & McGill, 1983)

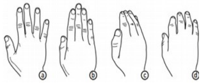
- Classification via one-vs-all SVM classification
  (Crammer & Singer et al, JMLR'01)

Proposed Method

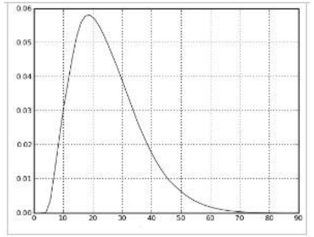
Coupled Sign Recognition

- Late fusion of gestures/postures responses.

$$R_{posture}$$

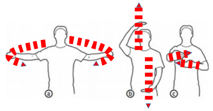


$$R = [R_{posture}, R_{gesture}]$$



$$\hat{c}_i = \arg\max_{(c_k)}(\omega_k \cdot R_i)$$

$$R_{gesture}$$



- Classification via one-vs-all SVM classification

(Crammer & Singer et al, JMLR'01)

Evaluation

Experimental Settings

- **Chalearn (2013) gesture dataset** (Escalera et al. ICMI'13 WS)
  - **20** sign classes | 27 subjects.

- **MSRC-12 dataset** (Fothergil et al. CHI'12)
  - (No RGBD data)
  - **12** gesture classes | 30 subjects.

Evaluation

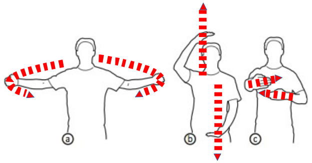
Experimental Settings

Cross-Validation for:

- Estimating the number of words for BoW.
- Estimating the number of states for each HMM.
- Training the multiclass SVM classifiers used for late fusion.

Evaluation

Hand Gesture-based Sign Recognition (I)

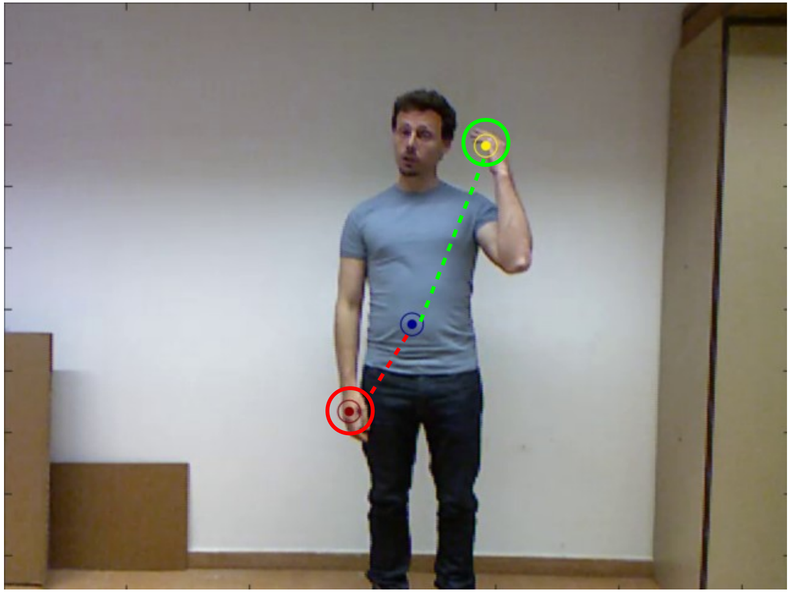


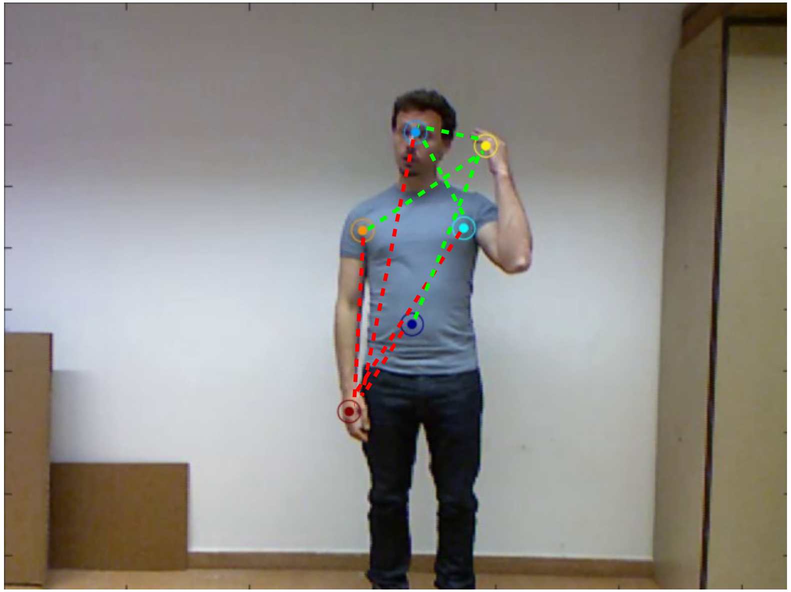
Methods:

Purely spatial

HD : Hand locations.

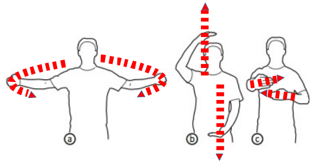
RBPD : Relative location of the hands wrt. to parts of the body.





Evaluation

Hand Gesture-based Sign Recognition (II)

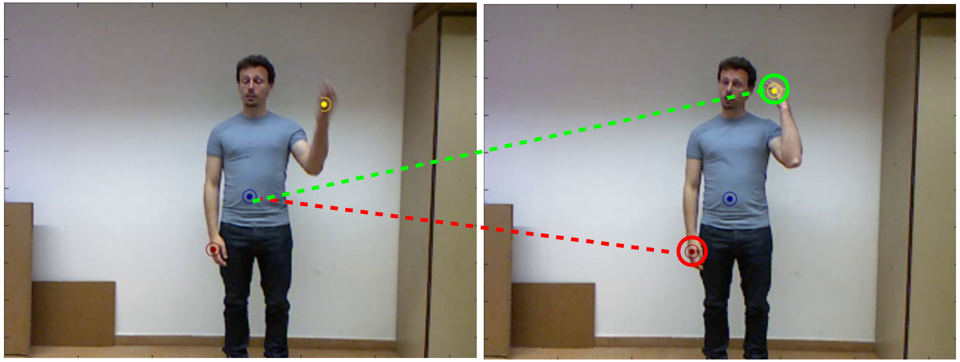


Methods:

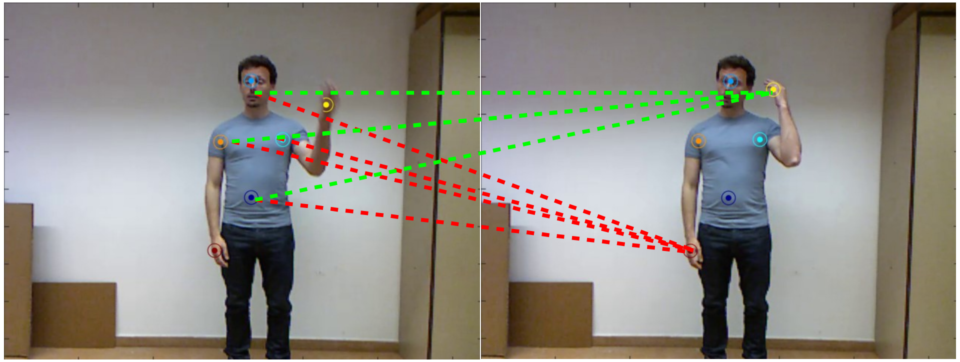
Locally-temporal extension

HD-T : Hand locations wrt. torso location in the previous frame.
RBPD-T : Relative location of the hands wrt. to parts of the body in the previous frame.

HD-T **RBPD-T**



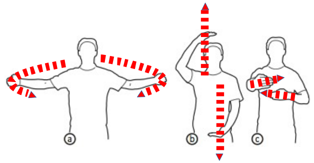
t-1 *t*



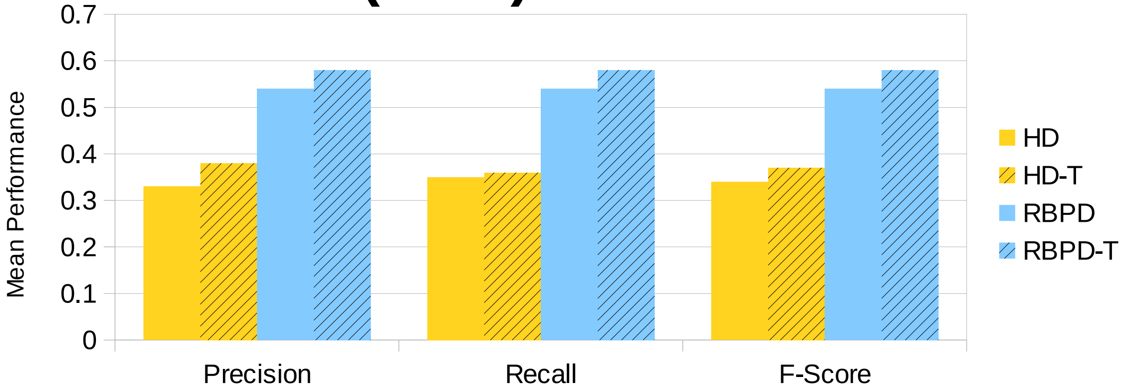
t-1 *t*

Evaluation

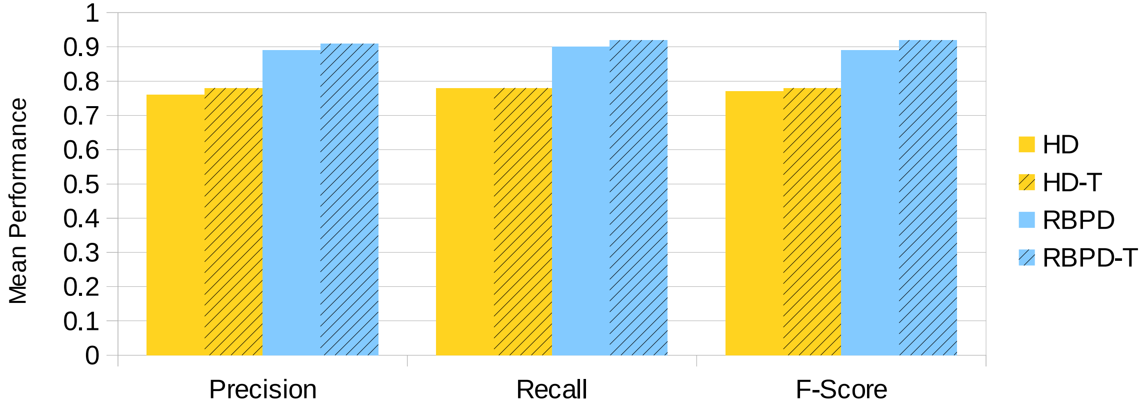
Hand Gesture-based Sign Recognition (III)



Chalearn (2013) dataset



MSRC-12 dataset



Discussion:

- HD performance is significantly lower. (especially for sign language setting)

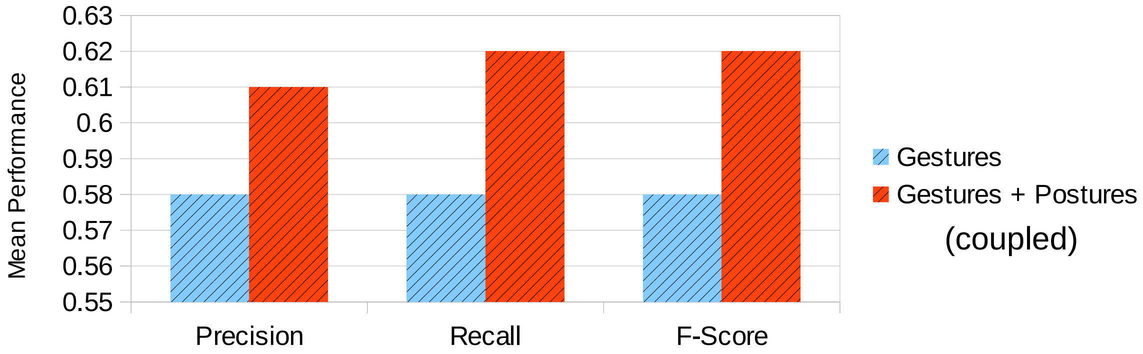
- The temporal extensions (-T) benefit both methods (HD & RBPD) .

- Purely reasoning about hand gestures may not solve sign language recognition.

Evaluation

Coupled Sign Recognition (I)

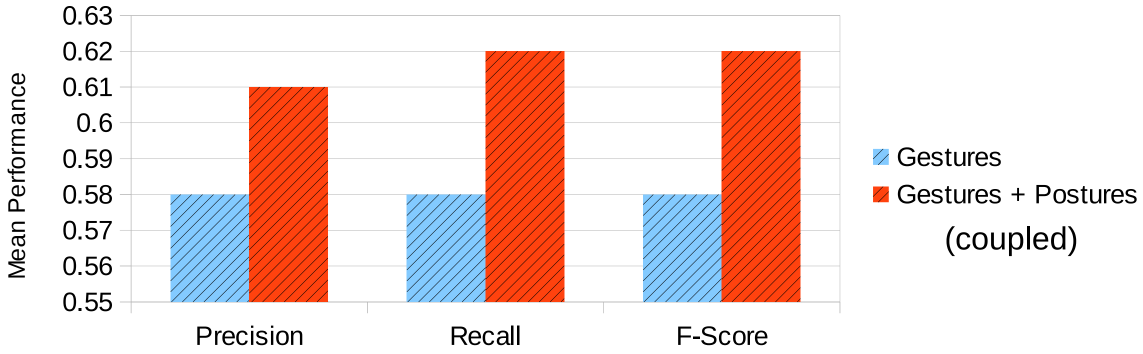
Chalearn (2013) dataset (focus on RBPD-T)



Evaluation

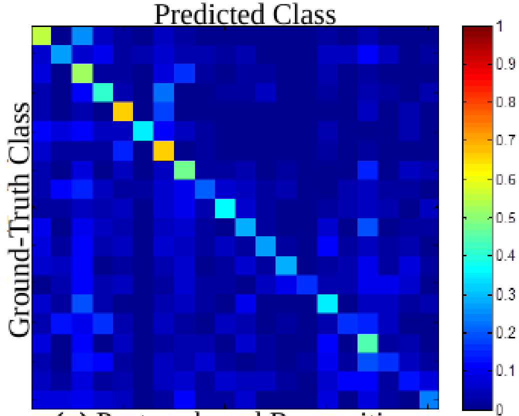
Coupled Sign Recognition (I)

Chalearn (2013) dataset (focus on RBPD-T)

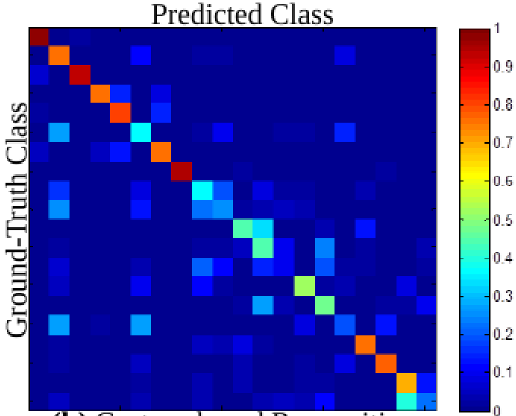


Discussion:

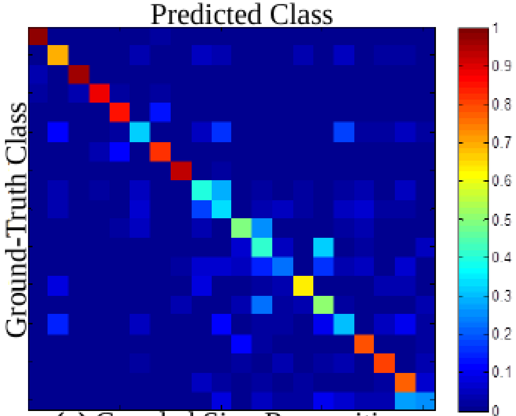
- Ambiguity between sign classes is clarified when combining postures and gestures.



(a) Posture-based Recognition



(b) Gesture-based Recognition



(c) Coupled Sign Recognition

Evaluation

Coupled Sign Recognition (II)

Chalearn (2013) dataset

	Precision	Recall	F-Score
Wu et al., ICMI'13	0.60	0.59	0.60
Yao et al., CVPR'14	-	-	0.56
Pfister et al., ECCV'14	0.61	**0.62**	**0.62**
Ours *(RBPD-T based)*	0.61	**0.62**	**0.62**

* Mean performance values are indicated.

Discussion:

- Improved performance over Wu et al. ICMI'13. (winners of Chalearn 2013 gesture challenge)

- Competitive performance with Pfister et al., ECCV'14. (which focuses on hand postures)

Conclusions

- Considering spatial relations between parts of the body provides richer descriptor for sign language recognition.

- Considering local temporal changes improves sign recognition performance.

Future Work

- **Integrate better methods to model hand postures.**
- **Focus on sign detection/localization.**
- **Integrate other features of sign language.**
  (e.g. language models or facial gestures)
- **Integrate recent methods for modeling video dynamics.**
  (e.g. VideoDarwin (Fernando et al., CVPR'15))

Questions?

Towards Sign Language Recognition based on Body-Parts Relations.

Marc Martínez Camarena, <u>José Oramas M.</u>, Tinne Tuytelaars
KU Leuven
September 28th 2015