SIGN LANGUAGE RECOGNITION BASED ON BODY PART RELATIONS

Marc Martínez Camarena

Promotor

Tinne Tuytelaars

Daily supervisor

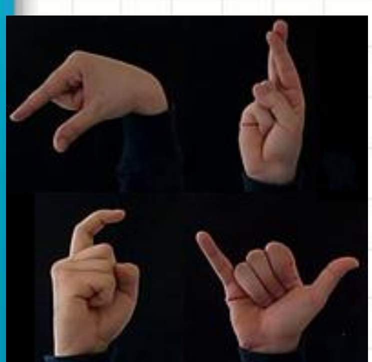
José Oramas M.

Table of contents

Project Overview

Problem statement

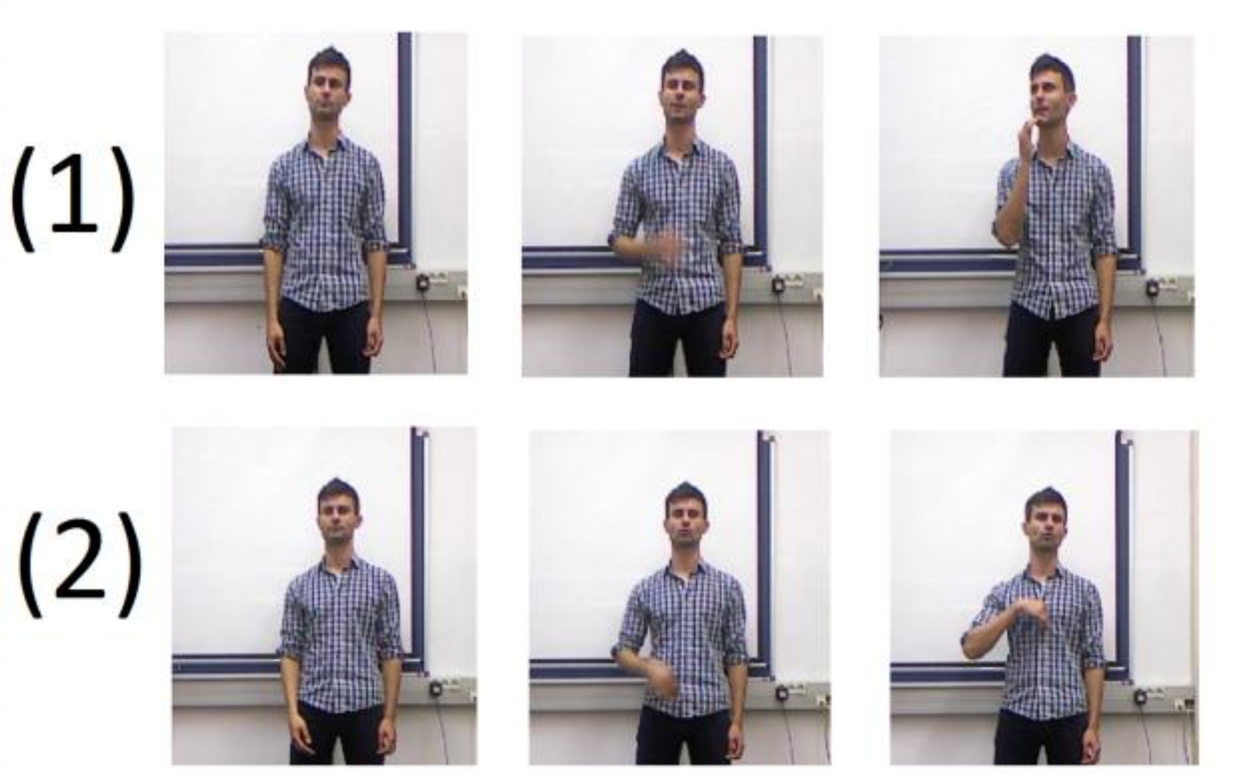
- Challenges nowadays → Sign language data → Significant for an application.
- A wide variety of sign languages.
- Each sign language: different grammar rules and different vocabulary.
- Something that have in common: **hand postures** and **hand gestures**.



Project Overview

Problem statement

- For many years, the problem of hand gesture → in following the hand trajectories.
- However, most of the sign languages, the signs are defined on a particular body area.



Project Overview

Objective

- The objective of this thesis is to explore the effect of considering relations between different parts of the body during reasoning for the task of Sign Language Recognition.

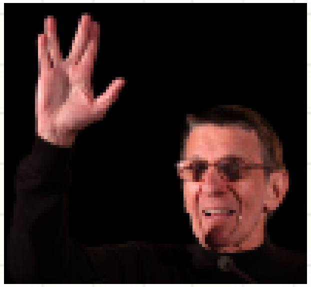
Related Work

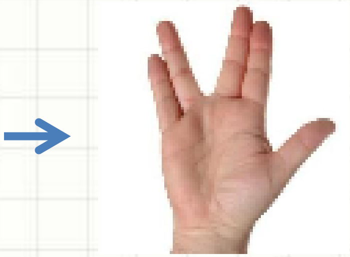
Sign Language Recognition

- Ming-Hsuan Yang et al – IEEE TPAMI,Vol.24, August 2002 [5]
- *Antonis A.Argyros – ECCV,2004* [6]
- Liu Yun and Zhang Peng – *WCSE, 2009* [7]
- Joyeeta Singha and Karen Das – *METIC, 2013* [8]

Most of them work in 2D. → Our work is focused with 3D.(Kinect)

Hand Posture and Hand Trajectory



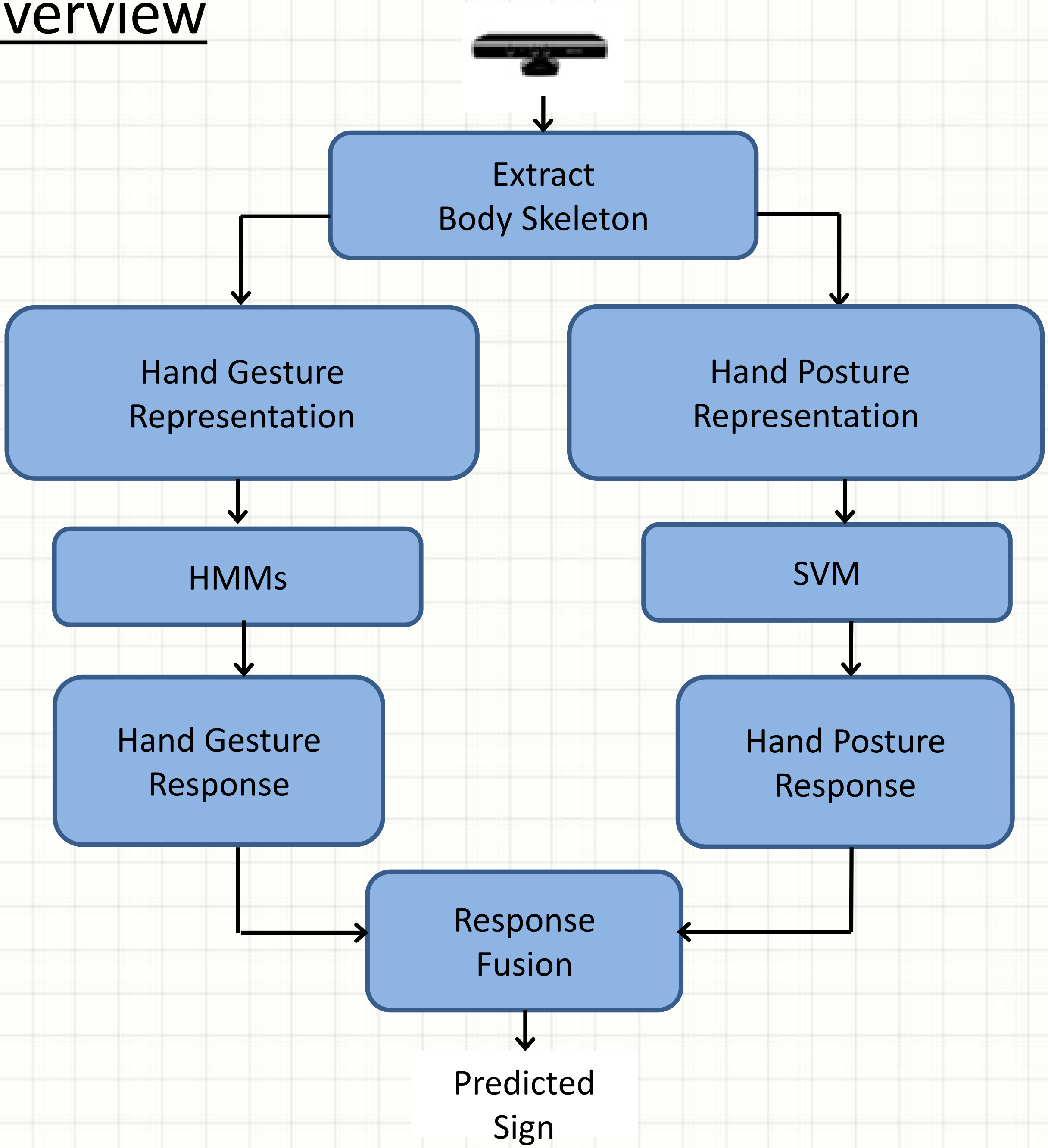


- X. Chai *et al – CAS, 2013***.**[9]
- *Iasonas Oikonomidis et al – ACCV, 2010.* [10]
- *Lalit K.Phadtare – WNYIPW, 2012***.** [11]
- Zhou Ren et al – IEEE TM,Vol.15, August 2013 [12]

They work only isolating the hand. → The different parts of the body are exploted in our work.

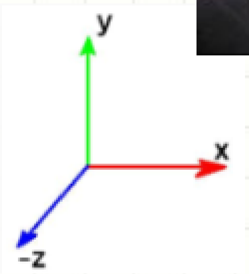
Table of contents

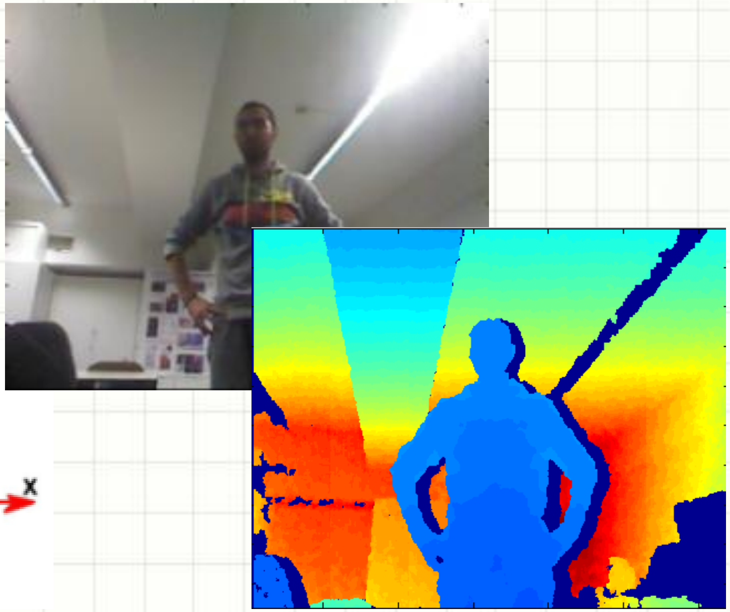
System Overview



Data Adquisition

- Microsoft Kinect Camera
  - Rgb and depth images.
  - Low-cost depth camera.
  - 3D points in the world coordinate space.
  - Provides information about objects range of 2 meters.





- Skeleton Body Representation Algorithm [1]
  - Extract body pixels by thresholding depth.
  - Random Forest to classify the body parts.
  - Mean-shift clustering algorithm to find joint positions.

Tracks the body in **REAL TIME** → **15 KEY JOINTS**

Body parts and joint positions

- Shotton et al. CVPR ,2011. [1]

Sign Recognition based on Hand Gesture Features

Relative Body Part Descriptor (RBPD)

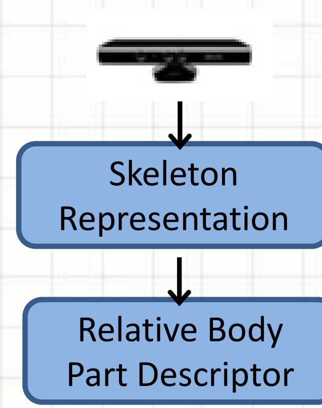
- 11 Keys Points in 3D.
- Two new world CS → Each hand.
- The Dimension of the RPBD : 66 dim

  (Two arrays of 33 dim**).**
- Z-score Normalization

RPBD$_{1-1}$	...	...	RPBD$_{1-66}$

Hand Gesture representation

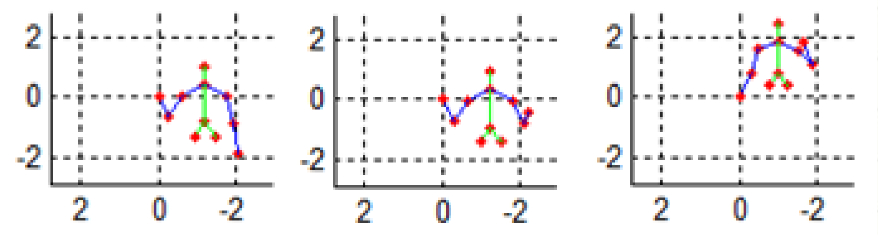
- RBPD for all the frames of a given sign.
- The dimension : 66 x N. →N depends of each sign.



Frame 1	RPBD$_{1-1}$	...	...	RPBD$_{1-66}$
...	...	...	...	...
...	...	...	...	...
Frame N	RPBD$_{N-1}$	...	...	RPBD$_{N-66}$

Sign Recognition based on Hand Gesture Features

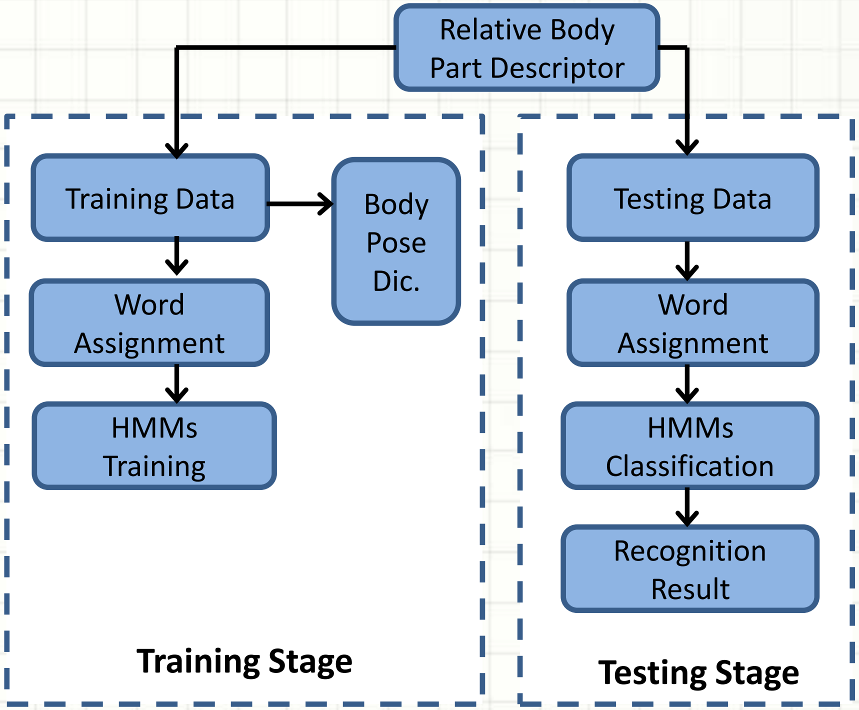
Body Pose Dictionary



- A data set of RBPD is collected.
- K words using K- means [1].
- Each hand gesture representation → Sequence of words.

Hidden Markov Models (HMMs)

- A physical sign → Markov chain with their different states.
- Each sign class → one HMM.
- Training → Adjust a model.
- Testing → Choose the training model with high response.

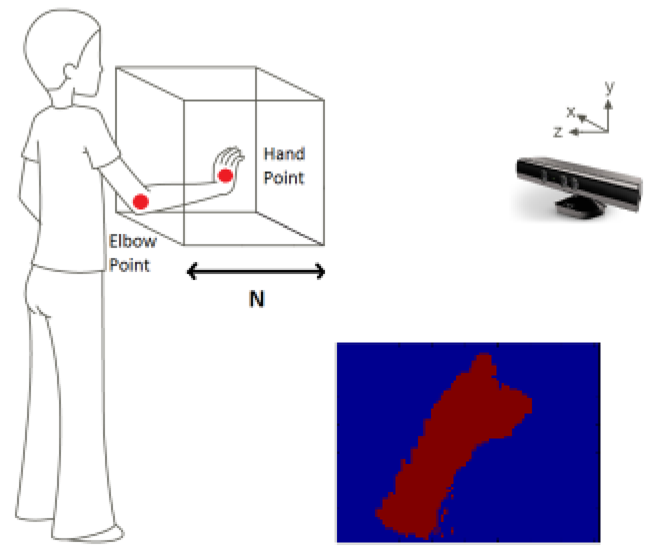


- J.B. MacQueen . PBS, 1967 [1]

Sign Recognition based on Hand Posture Features

1. Hand Posture Representation

- The 3D data around the hand is collected creating a cube.
- Nearest Neighbor with the 15 3D body points.
- Uniform resized image and binarization.



2. Shape Context Descriptor [1]

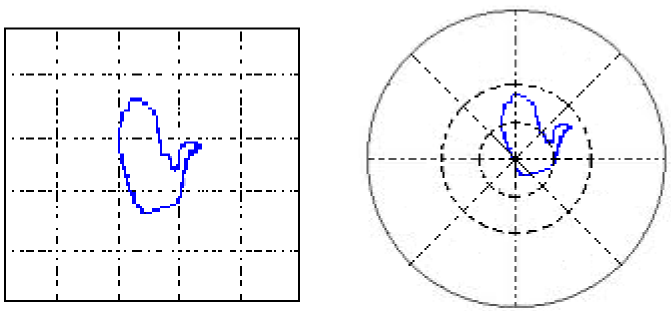
- Sampled edge equally-spaced points.
- Log-logar coordinate system.
- Histogram accumulates the amount of points.
- Each hand posture → Set of S. Context Desc.

3. Bag-of-Words Descriptor [2]

- Dictionary of K words.
- Each hand posture → Histogram of words.
- Hand postures along a sign by accumulating the words.
- Procedure for both hands → Concatenation.

4. Support Vector Machine (SVM)

- Training: One-vs-all multiclass SVM classifier.
- Testing: Desc. fed into the SVM to predict



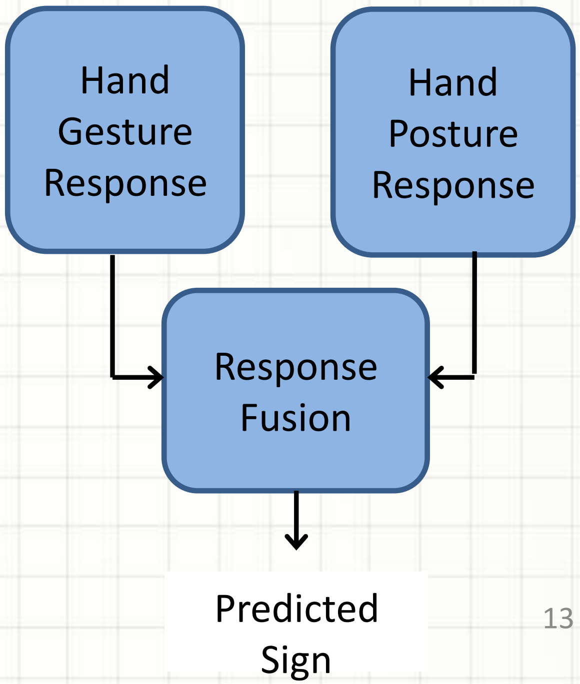
- S. Belongie and J. Malik. ICAIL, 2000 [1]
- T. Li et al. IEEE, TCSVT, 2011 [2]

Responses Combination

- Each introduced sign in the system → Two responses. (Likelihood vectors)
  - Response based on hand gestures features.
  - Response based on hand posture features.

- Each response vector indicates the likelihood of the introduced sign along the different sign classes of the system.

Joint Descriptor → Concatenation of these two likelihood vectors. → The dimension 2 x L (Nºsigns)



Fusion Response:

- Training: Multiclass classifier.
- Testing: Descriptor fed into the multiclass classifier to predict the given sign.

Evaluation Protocol

- Italian cultural signs data set provides [1]:

Rgb images. Depth images. Mask person. Skeleton (Body points).	20 different Italian cultural signs 27 different users			
		Training Data	Validation Data	Testing Data
	Italian Signs	4000	3960	3324

- Evaluation
  - Taking the depth images and the skeleton provided by the data set.
  - Sign classification rather than sign detection.
  - The performance results in terms of accuracy (Acc).

- Italian Cultural Sign Data Set . ChaLearn Multimodal Challenge Data , 2013[1]

Results – Hand Gesture Features

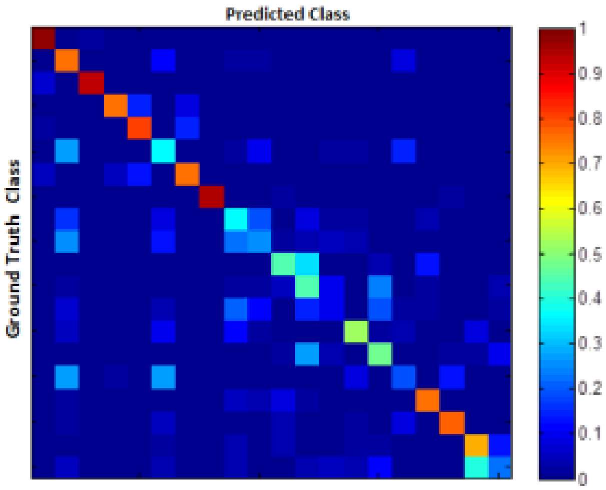
- Implementation descriptors:
  - RBPD → The descriptor proposed in methodology.
  - RBPD - 2 → As the RBPD. However, the hand of the previous frame.
  - TORSO → Reference the torso joint.
  - HD → Hand trajectory approach.

	Validation Data	Final Data
RBPD	51%	55%
RPBD-2	54%	57%
TORSO	42%	-
HD	33%	-

- Best mean accuracy of **57%** in the test set.

Discussion:

- Descriptors which take into account relations between body parts perform better.
- Difference between TORSO, RBPD and RBPD-2.
  - The hands the main element.
- Difference between RBPD and RBPD-2 → Minimum.
  - RPBD → Spatial.
  - RPBD-2 → Spatial and temporal.



Confusion Matrix RBPD -2

Results – Hand Posture Features

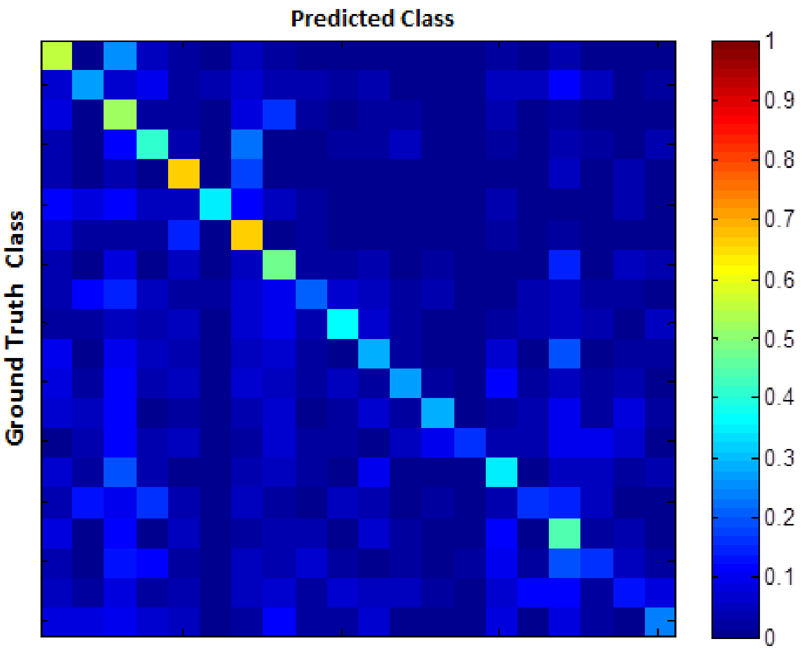
- Mean accuracy of 35% in the test set.

Discussion:

- Low average due to:
  - Images low resolution → Captured 2 or 3m about the camera.
  - Hands come in contact with the body.
  - Signs with similar hand posture sequences → Only different in a particular hand posture.

Comparison:

- Against other works [1][2].
  - Methods not well-suited for this dataset.
  - These methods obtain good results:
    - Images of high resolution.
    - The hands separated from the user.



Confusion Matrix Hand Posture Features

- C. Keskin et al. ICCV, 2011 [1]
- J. Knopp et al. ECCV, 2010 [2]

Results – Responses Combination

- Two different multiclass classifiers:
  - SVM → Mean accuracy of 62%.
  - ODKDE → Mean accuracy of 56%.
- **Discussion:**

- SVM better than ODKDE.

- SVM improves the system.

  - Improvement of 5pp

- ODKDE deteriorates the system.

  - Deterioration of 1pp

- **Comparison:**

- We compare against Jiaxiang Wu et al. [1] 1st in the ChaLearn Gesture 2013.

- Improvement of 3pp over their method.

Sign Class	Using Posture Features	Using Gesture Features	Responses Combination (SVM)	Wu et al. [1]
1	55%	97%	97%	85%
2	27%	76%	70%	70%
3	52%	94%	95%	87%
4	42%	75%	88%	80%
5	67%	80%	86%	79%
6	35%	37%	33%	36%
7	66%	76%	83%	86%
8	47%	94%	93%	95%
9	21%	37%	39%	37%
10	36%	25%	34%	85%
11	30%	44%	49%	22%
12	27%	45%	42%	43%
13	29%	10%	23%	35%
14	17%	52%	63%	33%
15	36%	47%	51%	26%
16	17%	19%	30%	47%
17	45%	75%	78%	56%
18	17%	77%	81%	65%
19	13%	70%	77%	79%
20	24%	23%	25%	39%
Mean Accuracy	35%	57%	62%	59%

- Wu et al. ICMI, 2013 [1]

Conclusion

- System representing each sign by a combination of hand posture descriptors and hand gesture descriptors.
- Hand gesture descriptors taking into account the different parts of the body perform better than hand global trajectory methods.
- Robust hand posture descriptor for images of low resolution.
- The combination of the responses of hand posture descriptors and hand gesture descriptors helps to improve the system.
  - The best configuration 62% mean accuracy.
    - Improvement of 5 pp → Hand gesture features.

Future work

- Add new features to make this system work with a real sign language. → Facial expressions, grammar rules…
- Compare with more descriptors and methods to emphasize the benefits of using the relations between body parts.
- In many signs, there is only a particular hand posture that has valuable information.
  - SVMs with latent variables.
- Method for sign detection/location.

THANK YOU FOR YOUR ATTENTION

References I

- [1] - Ye Gu, Ha Do, Yongsheng Ou and Weihua Sheng.
  Human Gesture Recognition through a Kinect Sensor.
  ICRB China - December 2012
- [2] – Leandro Miranda, Thales Vieira and Dimas Martinez.
  Real-time gesture recognition from depth data trough key poses learning and decision forests.
  SIBGRAPI Brazil - August 2012
- [3] – Hugo Jair Escalante and Isabelle Guyon.
  Principal motion: PCA-based reconstruction of motion histogramas.
  INAOE Mexico - May 2012
- [4] - Italian Cultural Signs data set:
  http://sunai.uoc.edu/chalearn/
- [5] – Ming-Hsuan Yang,Narendra Ahuja and Mark Tabb, Member.
  Extraction of 2D Motion Trajectories and Its Application to Hand Gesture Recognition.
  IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 8, AUGUST 2002.
- [6] – Antonis A.Argyros, Manolis I.A.Lourakis.
  Real-Time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera.
  Computer Vision – ECCV 2004.
- [7] – Liu Yun and Zhang Peng.
  An Automatic Hand Gesture Recognition System Based on Viola-Jones Method and SVMs.
  IEEE - WCSE 2009

References II

- [8] – *Joyeeta Shingha and Karen Das.*
  Hand Gesture Recognition Based on Karhunen-Loeve Transform.
  Mobile & Embedded Technology International Conference – 2013.
- [9] – *X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, and X. Chen.*
  Sign Language Recognition and Translation with Kinect.
  CAS. 2013.
- [10] – *X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, and X. Chen.*
  Sign Language Recognition and Translation with Kinect.
  Key.Lab of Intelligent Information Processing of Chinese Academy of Sciences. Institute of Computing Technology, CAS, 2013..
- [11] – *Lalit K. Phadtare, Raja S. Kushalnagar and Nathan D. Cahill.*
  Detecting Hand-Palm Orientation and Hand Shapes for Sign Language gesture recognition suing 3D images. *(WNYIPW), 2012 Western New York.*
- [12] – *Zhou Ren, Junsong Yuan, Jingjing Meng and Zhengyou Zhang.*
  Robust Part-Based Hand Gesture Recognition Using Kinect Sensor
  IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 5, AUGUST 2013
- [13] – *Xia Liu and Kikuo Fujimura.*
  Hand Gesture Recognition using Depth Data.
  6th IEEE Internacional Conf. on Auto. Face and Gesture Recognition – 2004.
- [14] - *X. Zabulisy, H. Baltzakisy and A. Argyroszy.*
  View-based Interpretation of Real-time Optical Flow for Gesture Recognition.
  Chapter 34, in "The Universal Access Handbook" Jun 2009.

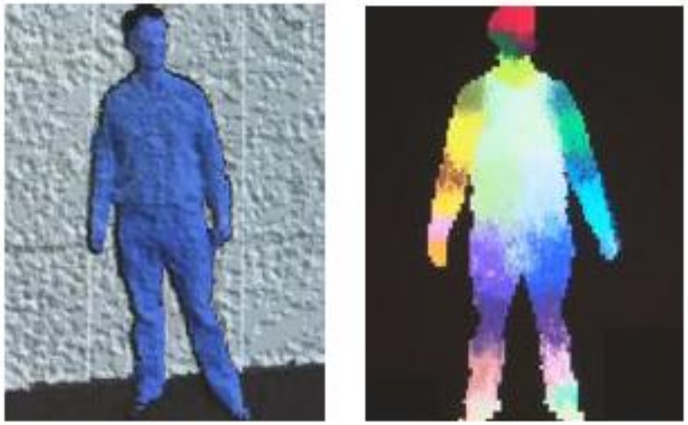
References III

- [15] - *J. B. Kruskal and M. Liberman.*
  The symmetric time-warping problem:from continuous to discrete.
  Addison-Wesley, Reading, Massachusetts, 1983.
- [16] - *L. Rabiner and B. Juang.*
  Fundamentals of speech recognition.
  Prentice Hall, 1993.
- [17] - *H. Y. Chung and Hee-Deok.*
  Conditional random field-based gesture recognition with depth information.
  Optical Engineering, Volume 52, id. 017201, 2013.
- [18] - *C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxa.*
  Conditional Models for Contextual Human Motion Recognition.
  TTI-C, University of Toronto, Rutgers University, 2010.
- [19] - *J. Yamato, J. Ohya, and K. Ishii.*
  Recognizing human action in time-sequential image using hidden Markov model.
  NTT Human Interface Labs., Yokosuka. In proceeding of: Computer Vision and Pattern Recognition. Proceedings CVPR.IEEE, 1992.
- [20] - *M. Elmezain, A. Al-Hamadi, and B. Michaelis.*
  Hand Gesture Recognition Based on Combined Features Extraction.
  World Academy of Science, Engineering and Technology.Vol,3, 2009.
- [21] - *J. Wu, J. Cheng, C. Zhao, and H. Lu.*
  Fusing Multi-modal Features for Gesture Recognition.
  ICMI Proceedings of the 15th ACM on International conference on multimodal interaction, pages 453–460, 2013.

References IV

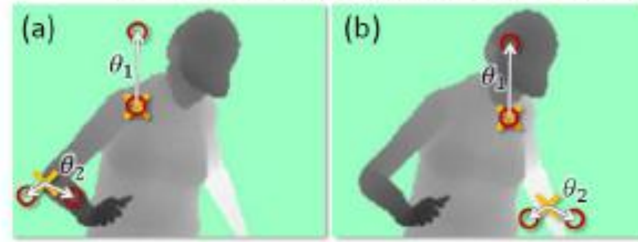
- [18] - *S. Belongie and J. Malik.*
  Matching with Shape Contexts. IContentbased
  Access of Image and Video Libraries. Proceedings. IEEE Workshop on, pages 20–26, 2000.
- [19] - *T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua.*
  Contextual Bag-of-Words for Visual Categorization.
  IEEE Transaction on Circuits and Systems for Video Technology, Vol. 21, No. 4, 2011.
- [20] *S. S. Keerthi, S. Sundararajan, K.-W. Chan, C.-J. Hsieh, and C.-J. Lin.*
  A sequential dual method for large scale multi-class linear SVMs.
  In KDD, 2008.
- [21] - *M. Kristan and A. Leonardis.*
  Online discriminative kernel density estimation.
  International Conference on Pattern Recognition, 2010.

Extract Body Pixels by Thresholding Depth



Features

- Difference of depth at two pixel
  - Offset is scaled by depth at reference pixel

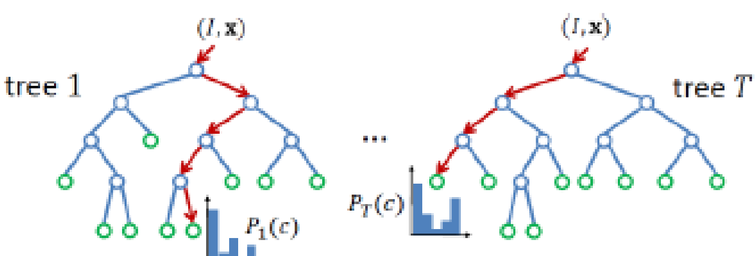


$$f_\theta(I, \mathbf{x}) = d_I\left(\mathbf{x} + \frac{\mathbf{u}}{d_I(\mathbf{x})}\right) - d_I\left(\mathbf{x} + \frac{\mathbf{v}}{d_I(\mathbf{x})}\right)$$

$d_I(\mathbf{x})$ is depth image, $\theta = (\mathbf{u}, \mathbf{v})$ is offset to second pixel

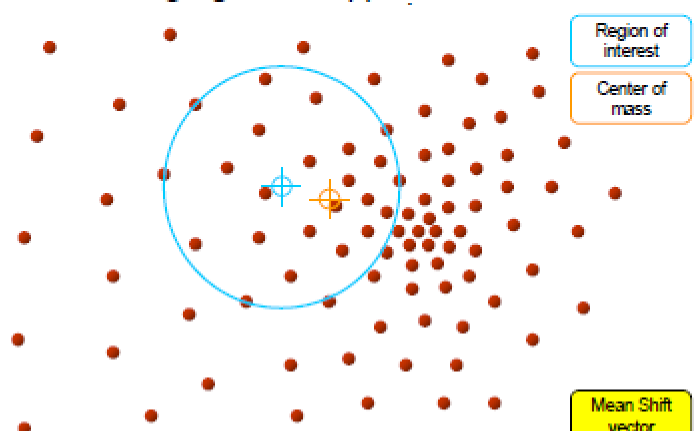
Part Classification with Random Forests

- **Randomized decision forest:** collection of independently-trained binary **decision trees**
- Each tree is a classifier that predicts the likelihood of a pixel **x** belonging to body part class *c*
  - Non-leaf node corresponds to a thresholded feature
  - Leaf node corresponds to a conjunction of several features
  - At leaf node store learned distribution $P(c|I, \mathbf{x})$



Joint Position Estimation

- Joints are estimated using the **mean-shift clustering** algorithm applied to the labeled



Objective : Find the densest region
Distribution of identical billiard balls