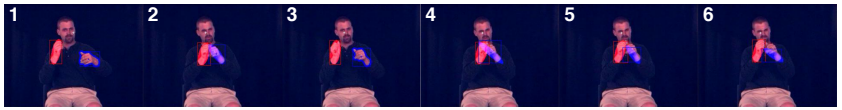
MODEL-LEVEL DATA-DRIVEN SUB-UNITS FOR SIGNS IN VIDEOS OF CONTINUOUS SIGN LANGUAGE

Stavros Theodorakis, Vassilis Pitsikalis and Petros Maragos

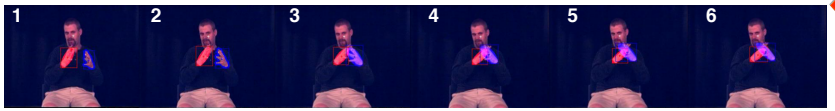
School of ECE, National Technical University of Athens 15773, Greece

2010 ICASSP Dallas

/WITH/



/FOOTBALL/



Sub-sampled frame sequence of /WITH/, /FOOTBALL/ realizations.

1. Sign Language – Motivation

❑Visual patterns are formed by *hand shapes, manual* or general body *motion* and *facial expressions*.

❑Word in spoken language → Sign.

❑**Position, Movement**, Hand Shape, Orientation, Facial.

❑**Phonological Sub-Units**: no well-defined unit equivalent to the phoneme in speech.

❑We focus on **automatic data-driven modeling of sub-units** *without* any linguistic – phonological information.

❑**Goal:** continuous Sign Language Recognition.

2. Outline-Contributions

❑ **Visual Front-End + Feature Extraction:** Hands' centroid movement-position measurements: 2D location, 2D derivatives, velocity, dynamics .

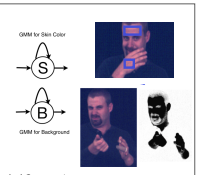
❑ **Data-Driven Sub-Unit Construction**

- ❑State-level sign segmentation.
- ❑Dynamic *(vs. Static)* Specific Modeling → Structure.
- ❑Model-level Sub-Unit Construction: HMM clustering.

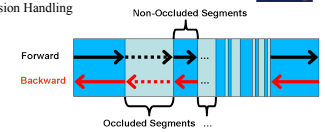
3. Visual Front-End

❑Hand and Head Detection

- ❑Probabilistic skin color model,
- ❑Color force in a Geodesic Active Contours model [10,11].



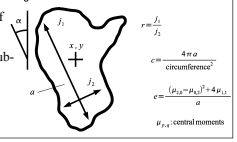
❑Occlusion Handling



❑Features: Movement-Position of the hands.

❑Simple Features → Effect on Sub-Unit modeling ?

❑Movement-Position: main cues that describe a sign [1, 2].



$r = \frac{j_1}{j_2}$

$c = \frac{4\pi a}{\text{circumference}^2}$

$e = \frac{(\mu_{2,0}-\mu_{0,2})^2 + 4\mu_{1,1}}{a}$

$\mu_{p,q}$: central moments

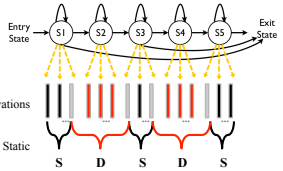
4. Segmentation-Dynamic/Static

❑Lack of annotation within sign units !

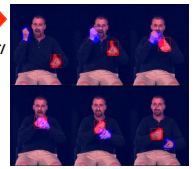
❑**Sign Segmentation**: data driven, model-based, state-level

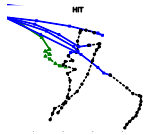
- ❑**Segmentation Points** + **Dynamic vs. Static** classification

❑**Pronunciation Variation**: Pronunciation Clustering (PC) | Sign



/HIT/





Sub-sampled frame sequence of /HIT/ realization: Upper: 1-3, Bottom: 4-6.

Right hand centroid trajectories and segmentation (color); multiple /HIT/ realizations

5. Sub-Unit Modeling

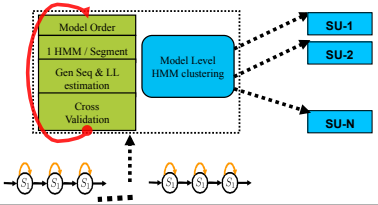
❑Segmentation + Dynamic/Static Classification
→ prosperous initialization to model intra-sign segments.

❑Dynamic *(vs. Static)* modeling

❑Sub-Unit modeling at the **model level**.

❑**Normalization** wrt. Initial Position

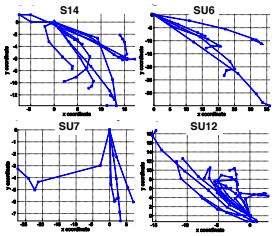
❑Hierarchically cluster ***whole dynamic models*** (HMMs) [8] based on a similarity measure among models.



6. Qualitative Results

Sample Sub-Unit Clusters

❑Capture geometrical properties: Direction.



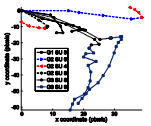
Right hand centroid trajectories
Cluster Index → Sub-Unit Index

Sub-Unit Sharing

❑ *G1* /FOOTBALL/ : SU8

❑ *G2* /WITH/ : SU6 +SU4, SU8 + SU4

❑ *G3* /HIT/ : SU8+SU9



Right hand centroid trajectories and Sub-Units (color).

7. Current Work

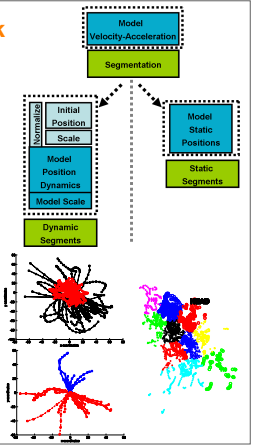
❑ We cluster *not* the independent frames [5],

❑ *neither* the frames' sequences as segments [6, 7] at the feature level.

❑ **Model Level Dynamics** Incorporation.

❑ **Modeling Structure** Each modeling level/ appropriate feature and modeling.

❑ **Feature Normalization** focus on the actual phenomena, factors.



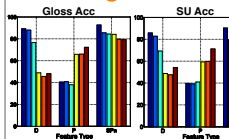
8. Experimental Data Setup

❑Continuous American Sign Language [9]: 843 utterances, 406 words, 4 signers, Uniform background. Sign level transcriptions; English Glosses; annotated start/end points.

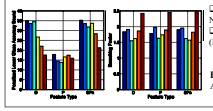
❑BU400 HQ, 6 videos, 648×484 frames, 60fps. Most frequent Glosses; cross-validate; train/test 60-40%

REALLY	56	SAY	31	BUT	25	ONE	24	ON	18	HERE	17
SEE	16	COP	13	LOOK	12	TO	12	FRIEND	12	MOTHER	11
MANY	10	MAN	10	GET	9	GO	7	BETTER	7	BECAUSE	7
CIGARETTE	6	SOME	6	EXAMPLE	6	WEATHER	5	END	5	FORMERLY	5
FINISH	5	DEGREE	5	_80	5	SHOWER	5	KIND	5	WORK	5

9. Recognition Results



Who	Segm.	Ft.	Str.	Clas
[6]	3S	P	-	DTW
Ours	5SE	P	-	DTW
Ours	PC+5SE	P	-	DTW
Ours	PC+5SE	P	D	HMM
Ours~	2SEr	V	D+	HMM
Ours~	2SEr	V	D	HMM



❑**x-axis** (left to right) Feature: Direction (D), Position (P), Normalized Position (SPn)

❑**Table**: Dynamic (D); Velocity (V); x States (xS); Escape (E); Ergodic (Er); Pronunciation Clustering (PC).

Bottom Figures: Penalized Lower Bound Gloss Accuracy, Lexicon Branching Factor

References

[1] S. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," Trans. Pattern Anal. Mach. Intellig., vol. 27, no. 6, pp. 873–891, 2005.
[2] K. Emmorey, Language, cognition, and the brain: insights from sign language research, Erlbaum, 2002.
[3] C. Vogler and D. Metaxas, "Handshapes and movements: Multiple-channel american sign language recognition," in Gesture Workshop, 2003, pp. 247–258.
[4] S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," Sign Language Studies, vol. 64, pp. 195 – 277, 1989.
[5] B. Bauer and K-F. Kraiss, "Towards an automatic sign language recognition system using subunits," in Int'l Gesture Workshop, 2001, vol. 2298, pp. 64–75.
[6] G. Fang, X. Gao, W. Gao, and Y. Chen, "A novel approach to automatically extracting basic units from chinese sign language," in Proc. ICPR, USA, 2004, vol. 4, pp. 454–457.
[7] J. Han, G. Awad, and A. Sutherland, "Modelling and segmenting subunits for sign language recognition based on hand motion analysis," Pat. Rec. Lett., vol. 30, no. 6, pp. 623–633, 2009.
[8] P. Smyth, "Clustering sequences with hidden markov models," in In Advances in Neural Information Processing Systems, 1997, vol. 9, pp. 648–654.
[9] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and Ney H., "Benchmark databases for video-based automatic sign language recognition," in Proc. International Conference on Language Resources and Evaluation (LREC), May 2008.
[10] G. Papandreou and P. Maragos, "Multigrid geometric active contour models," IEEE Trans. on Image Process., vol. 16, no. 1, pp. 229–240, Jan. 2007.
[11] O. Diamanti and P. Maragos, "Geodesic active regions for segmentation and tracking of human gestures in sign language videos," in Proc. ICIP, 2008.
[12] V. Digalakis, P. Monaco, and H. Murveit, "Genones: generalized mixture tying in continuous hidden markovmodel-based speech recognizers," IEEE Trans. on Speech and Audio Processing, vol. 4, no. 4, pp. 281–289, Jul 1996.
[13] B.-H. Juang and L. R. Rabiner, "A probabilistic distance for hidden markov models," AT & T Technical Journal, 1985.

Acknowledgments

This research work was supported by the EU under the research program Dictasign with grant FP7-ICT-3-231135. We also wish to thank Boston University and C. Neidle for providing the BU400 video database.

For further information

Please contact. [sth, vpitsik, maragos]@cs.ntua.gr More information can be found at http://cvsp.cs.ntua.gr and http://www.dictasign.eu