Enhancing a Sign Language Translation System with Vision-Based Features

P. Dreuw, D. Stein, and H. Ney
Human Language Technology and Pattern Recognition

Introduction

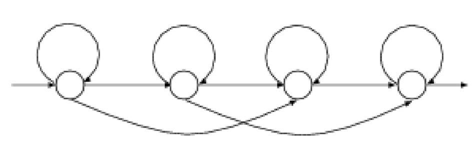
- automatic sign language recognition (ASLR):
  - ▹ vision-based approach, no special hardware, inhomogeneous background, occlusions, ...
  - ▹ hand and face positions are important features for ASLR
  - ▹ tracking problems:
    - ∘ hands are signing in front of the face
    - ∘ hands are moving very fast and abrupt
  - ▹ we avoid preliminary decisions and propose to use the same techniques that are successfully applied in automatic speech recognition (ASR)
- automatic sign language translation (ASLT):
  - ▹ problem: handling of spatial reference points in the signing space
  - ▹ locations are encoded at static points in signing space as spatial references for motion events
  - ▹ proper representation and translation

Tracking System

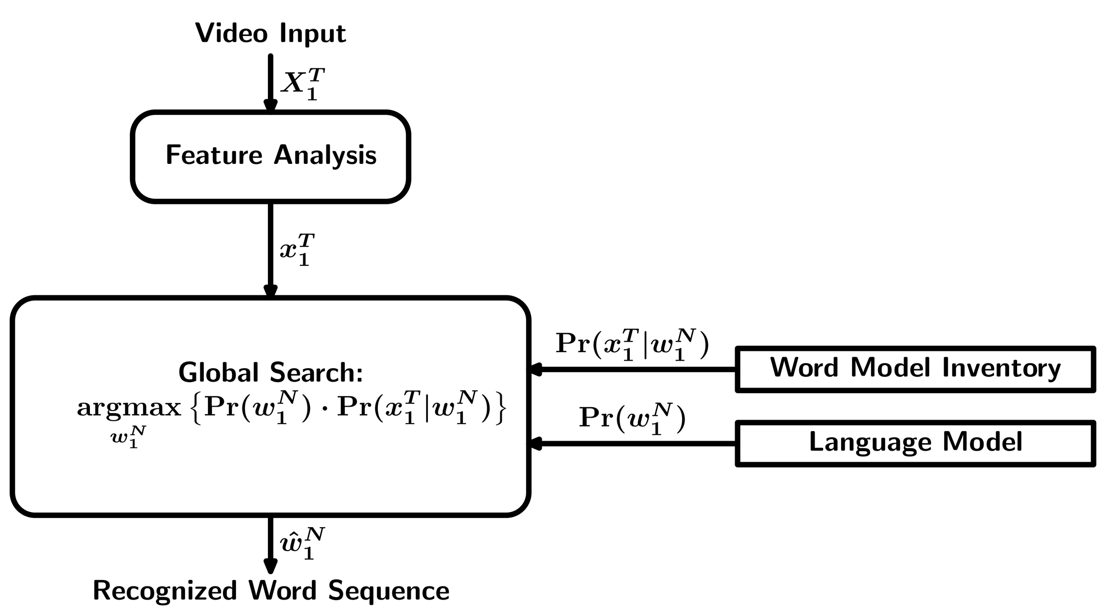
- tracking is necessary to extract sign language features: hand position relative to the head, facial expressions, ...
- environment problems: tracking features are usually not correct and can only be used to produce candidate regions
- idea: prevent taking possibly wrong local decisions
- how: tracking is done at the end of a sequence by tracking back the decisions to reconstruct the best path
- the best path is the path with the highest score wrt. a given scoring function
- 2 steps:
  1. score calculation: calculate a global score $S(t, x, y)$ and a back-pointer table B for the best tracking until time step t which ends in position (x, y)
  2. traceback: reconstruct the best path $t \rightarrow (x, y)$ using S and B
- framework can be used for head and hand tracking just by using different scoring functions

Sign Language Recognition

- a sign/gesture is a sequence of images
- important features
  - ▹ hand-shapes, facial expressions, lip-patterns
  - ▹ orientation and movement of the hands, arms or body
- HMMs are used to compensate time and amplitude variations of the signers



- goal: find the model which best expresses the observation sequence
- to classify an observation sequence X_1^T, we use the Bayesian decision rule:



Sign Language Translation

- state-of-the-art phrase-based statistical machine translation system
  - ▹ for a recognized sequence f_1^J we maximize a translation probability for target sentences e_1^I
  - ▹ log-linear combination model:

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^M \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)}$$

  - ▹ set of different features h_m, scaling factors λ_m
  - ▹ trained with downhill simplex algorithm
- tracking positions of the sentences were clustered and their mean calculated
- for deictic signs, the nearest cluster according to the Euclidean distance was added as additional word information for the translation model

Experimental Results

- RWTH-Boston-Hands database:
  - ▹ 1000 annotated frames, 2.3% tracking error rate
  - ▹ tracking of head and dominant-hand for ASLR

- RWTH-Boston-104 database:
  - ▹ 161 training sentences, 40 test sentences
  - ▹ preliminary translation experiments with dominant-hand tracking features

	Recognition Example
source	JOHN IX GIVE MAN IX NEW COAT
result	JOHN __ GIVE ___ IX NEW COAT

	Translation Example
without tracking	John gives that man a coat
with tracking	John gives the man over there a coat.

Translation Features	WER[%]	PER[%]
without tracking	44.2	42.5
with tracking	42.5	40.3

Conclusion

- if no pruning is used in tracking, the optimal path is guaranteed to be found
- the proposed tracking algorithm enables to track a target disregarding information gaps (e.g. due to occlusions)
- advantages under noisy circumstances
- dominant-hand tracking position information improves recognition error rates
- incorporation of the tracking data for the deixis words helps the translation system to discriminate between deixis as
  - ▹ distinctive article,
  - ▹ locative reference or
  - ▹ discourse entity reference

Outlook

- model for all entities
- handling spatial verb flexion, time information
- speech-to-speech

Lehrstuhl für Informatik 6 – Computer Science Department – RWTH Aachen University – Aachen, Germany
{dreuw,stein,ney}@cs.rwth-aachen.de
http://www-i6.informatik.rwth-aachen.de