Computer Vision and Image Understanding 141 (2015) 108–125

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers

Oscar Koller*, Jens Forster, Hermann Ney

Human Language Technology and Pattern Recognition, RWTH Aachen University, Germany

ARTICLE INFO

Article history:
Received 30 October 2014
Accepted 23 September 2015

Keywords:
Sign language recognition
Statistical modelling
Tracking
Visual modelling
Signer dependency
Signer adaptation

ABSTRACT

This work presents a statistical recognition approach performing large vocabulary continuous sign language recognition across different signers. Automatic sign language recognition is currently evolving from artificial lab-generated data to 'real-life' data. To the best of our knowledge, this is the first time system design on a large data set with true focus on real-life applicability is thoroughly presented. Our contributions are in five areas, namely tracking, features, signer dependency, visual modelling and language modelling. We experimentally show the importance of tracking for sign language recognition with respect to the hands and facial landmarks. We further contribute by explicitly enumerating the impact of multimodal sign language features describing hand shape, hand position and movement, inter-hand-relation and detailed facial parameters, as well as temporal derivatives. In terms of visual modelling we evaluate non-gesture-models, length modelling and universal transition models. Signer-dependency is tackled with CMLLR adaptation and we further improve the recognition by employing class language models. We evaluate on two publicly available large vocabulary databases representing lab-data (SIGNUM database: 25 signers, 455 sign vocabulary, 19k sentences) and unconstrained 'real-life' sign language (RWTH-PHOENIX-Weather database: 9 signers, 1081 sign vocabulary, 7k sentences) and achieve up to 10.0%/16.4% and respectively up to 34.3%/53.0% word error rate for single signer/multi-signer setups. Finally, this work aims at providing a starting point to newcomers into the field.

1. Introduction

Sign language (SLs), the natural languages of the Deaf, are known to be as grammatically complete and rich as their spoken language counterparts. Science discovered SLs a few decades ago and research promises new insights into many human language related fields from language acquisition to automatic processing.

SLs are not international and convey meaning by more than just the moving hands. They make use of both 'manual features' (hand shape, position, orientation and movement) and linguistically termed 'non-manual' features consisting of the face (eye gaze, mouthing/mouth gestures and facial expression) and the upper body posture (head nods/shakes and shoulder orientation). All of these language components are used in parallel to complement each other, but depending on the context of an utterance, a specific component may or may not be required to interpret the sign. Sometimes, an individual component plays an integral role within the sign, sometimes modifies the meaning, and sometimes provides spatial or temporal context. Furthermore, the different information channels do not share a fixed temporal alignment, but are rather loosely coupled.

Computer vision methods exist to extract features for these different channels. However, SL constitutes an extremely challenging test bed as it incorporates huge variations inherent to natural languages. High signing speed, motion blur, different lighting and view-point-dependent appearance have to be tackled. Furthermore, ambiguity is inherent to SLs, as each movement, each change in eye gaze or each appearance of the tongue may or may not have a grammatical or semantic function depending on the context. Thus, learning features and training classifiers that can be applied to SL recognition must cope with a natural variation seldom present in other tasks. At the same time, it constitutes a very well-defined environment for assessing gesture recognition techniques by providing rules and boundaries for naturalness and intelligibility.

Historically, research on (ALSR) had mainly access to small data sets, limited number of signers and a limited recognition vocabulary. Recently, a very exciting era has started. SL research is moving out of the lab into 'real-life' scenarios.

In this paper, we present extensive results and thorough analysis on, to our knowledge, the currently biggest publicly available corpus of continuous SL (RWTH-PHOENIX-Weather). It covers only 'real-life'

* Corresponding author.
E-mail addresses: koller@cs.rwth-aachen.de (O. Koller), forster@cs.rwth-aachen.de (J. Forster), ney@cs.rwth-aachen.de (H. Ney).

http://dx.doi.org/10.1016/j.cviu.2015.09.013
1077-3142/

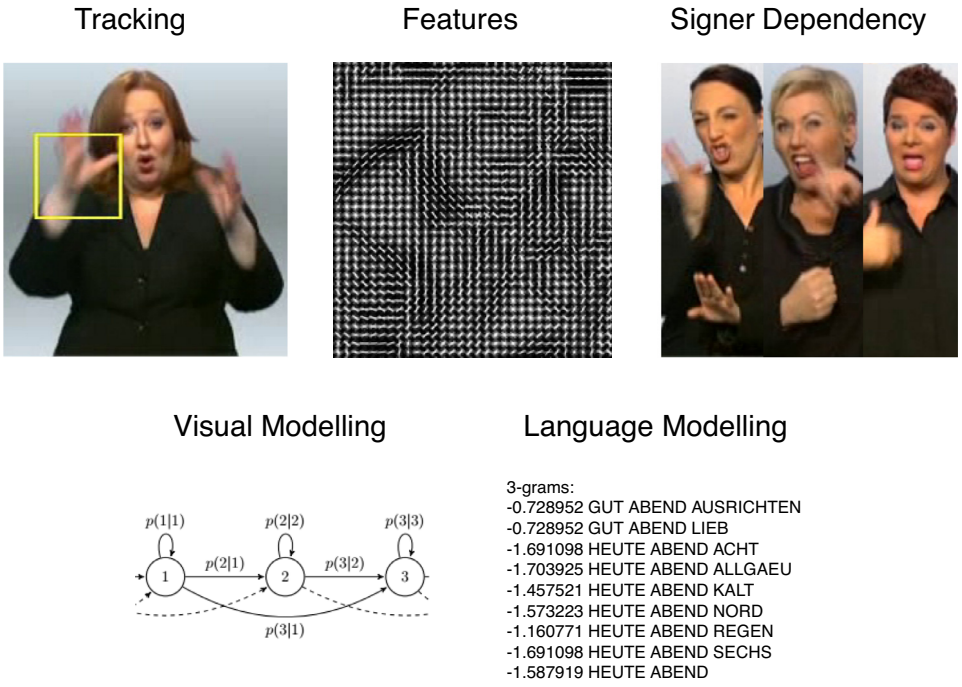

Fig. 1. Areas tackled by this paper.

signing recorded on public TV broadcast that has been manually labelled by native speakers. To the best of our knowledge, this is the first time, system design on a large data set with true focus on real-life applicability is thoroughly presented. Our contributions are in five areas, namely tracking, features, signer dependency, visual modelling and language modelling.

We experimentally show the importance of tracking for SL recognition, with respect to the hands and facial landmarks. We further contribute by explicitly enumerating the impact of multimodal SL features describing hand shape, hand position and movement, inter-hand-relation and detailed facial parameters, as well as temporal derivatives. Among these, the combination of hand gesture features and face features is novel, as well as the definition of the high-level face features.

In terms of visual modelling we evaluate non-gesture-models, length modelling and universal transition models. Signer-dependency is tackled using (CMLLR) adaptation. Further, class language models (LMs), CMLLR adaptation, as well as non-gesture-models are the new aspects to ASLR.

In Section 2, we introduce the state-of-the-art in the context of SL recognition and its related sub-fields. In the following two sections, we first present the employed data sets used for evaluating this work (Section 3) and then, in Section 4, the overall recognition system is explained in detail.

The subsequent sections tackle each of the five areas depicted in Fig. 1, giving first the technical details and then the experimental evidences. This is meant to open up the field to newcomers, who can estimate the impact of the most important design decisions. In Section 5, the employed tracking techniques are discussed and their impact with respect to the hands and facial landmarks is given. Section 6 presents the employed features covering most important modalities for SL and shows the impact on overall recognition results. Methods improving the visual modelling are presented in Section 7. Our approach to tackling multiple signers is presented in Section 8. The experimental sections end with our contribution to language modelling in Section 9. Finally, the paper closes with a conclusion and discussion of future work in Sections 10 and 11.

2. Related work

This section describes related work in ASLR and its related disciplines. The field evolved from recognising isolated signs of very limited number, articulated by only a single signer toward more complex settings with continuous natural signing of multiple signers. Thereby, the scientific community advances three tracks simultaneously:

1. The methods to extract relevant information become more sophisticated and precise, moving from expensive glove- and accelerometer-based setups to non-intrusive computer vision techniques.
2. The modelling of SL evolves to accommodate both linguistic and data-driven findings, aiming to fully reflect the complexity of the visual language.
3. The available data sets become more challenging, bigger and closer to real-life signing.

Although more recently ASLR is starting to tackle 'real-life' continuous signing data, the majority of work in the community still focuses on the recognition of isolated signs mostly in artificial settings.

2.1. Sign language recognition

Tamura and Kawasaki [64] were the first to start exploring the world of ASLR. They built a system to recognise isolated signs of Japanese SL by modelling the shape, movement and location of the hand using a simple colour segmentation. A lot of the early ASLR systems then employed glove-based motion tracking systems to overcome difficulties with vision-based feature extraction and tracking. This allowed to increase the recognition vocabularies while still achieving high accuracy on simpler tasks. In this way, Kadous [34] distinguished 95 (AUSLAN) signs with accuracies of around 80% using decision trees as classifier. Two years later Liang and Ouhyoung [42] proved to recognise a lexicon of 250 different signs of Taiwanese SL with a similar error rate. However, due to high cost of motion capture systems and thus low real-world applicability, coloured gloves and computer vision techniques started emerging [13].