Image and Vision Computing 32 (2014) 671–681

Contents lists available at ScienceDirect

Image and Vision Computing

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

Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions ☆

Jingjing Liu [a,1], Bo Liu [a,1], Shaoting Zhang [b,*], Fei Yang [a], Peng Yang [a], Dimitris N. Metaxas [a], Carol Neidle [c]

[a] Department of Computer Science, Rutgers University, Piscataway, NJ, USA
[b] Department of Computer Science, University of North Carolina at Charlotte, NC, USA
[c] Linguistics Program, Department of Romance Studies, Boston University, Boston, MA, USA

ARTICLE INFO

Article history:
Received 7 June 2013
Received in revised form 26 December 2013
Accepted 21 February 2014
Available online 12 March 2014

Keywords:
American Sign Language (ASL)
Non-manual grammatical markers
Eyebrow height
Head gestures
Facial expressions
Conditional Random Field (CRF)

ABSTRACT

Changes in eyebrow configuration, in conjunction with other facial expressions and head gestures, are used to signal essential grammatical information in signed languages. This paper proposes an automatic recognition system for non-manual grammatical markers in American Sign Language (ASL) based on a multi-scale, spatio-temporal analysis of head pose and facial expressions. The analysis takes account of gestural components of these markers, such as raised or lowered eyebrows and different types of periodic head movements. To advance the state of the art in non-manual grammatical marker recognition, we propose a novel multi-scale learning approach that exploits spatio-temporally *low-level* and *high-level* facial features. *Low-level* features are based on information about facial geometry and appearance, as well as head pose, and are obtained through accurate 3D deformable model-based face tracking. *High-level* features are based on the identification of gestural events, of varying duration, that constitute the components of linguistic non-manual markers. Specifically, we recognize events such as raised and lowered eyebrows, head nods, and head shakes. We also partition these events into temporal phases. We separate the anticipatory transitional movement (the *onset*) from the linguistically significant portion of the event, and we further separate the *core* of the event from the transitional movement that occurs as the articulators return to the neutral position towards the end of the event (the *offset*). This partitioning is essential for the temporally accurate localization of the grammatical markers, which could not be achieved at this level of precision with previous computer vision methods. In addition, we analyze and use the motion patterns of these non-manual events. Those patterns, together with the information about the type of event and its temporal phases, are defined as the high-level features. Using this multi-scale, spatio-temporal combination of low- and high-level features, we employ learning methods for accurate recognition of non-manual grammatical markers in ASL sentences.

1. Introduction

Signed languages are full-fledged natural languages, comparable in structure and complexity to spoken languages but manifested in the visual–gestural modality. Computer-based recognition of sign language from video, which has been the object of various research efforts over the last two decades (e.g., [42,49,47]), is particularly challenging in that it requires attention to detection and interpretation of linguistic information conveyed through both the manual and the non-manual channels. Although the equivalents of words are articulated primarily through movements of the hands and arms, important linguistic information of various kinds is also expressed non-manually: through head movements and facial expressions. In particular, essential grammatical information about such things as negation, clausal type, question status, and topics is conveyed by clusters of specific facial gestures and head movements [3–5,9,22,23,34]. For instance, the marking of yes/no questions typically includes raised eyebrows. However, raised eyebrows are a component of many other grammatical markers, as well (e.g., topics, conditional clauses, "relative clauses"); these expressions are distinguished by other differences in facial expressions, including eye gaze and eye aperture, nose configuration, and head positions and movements.

Despite their linguistic importance, it is only relatively recently that sign language recognition has begun to focus on detection of the non-manual components of signed languages, in some cases as an aid to the recognition by computer of manual signs [2,39], and in other cases, with a view toward interpretation of the grammatical information carried by non-manual markers [26,30,37]. Computer-based recognition of non-manual aspects of signed languages has made use of methods for facial expression recognition [44,13] and head pose estimation [32], which have been active research topics in computer vision for

☆ This paper has been recommended for acceptance by Qiang Ji.
* Corresponding author. Tel.: +1 732 4450636; fax: +1 732 4450537.
E-mail address: szhang16@cs.uncc.edu (S. Zhang).
[1] Indicates equal contribution.

http://dx.doi.org/10.1016/j.imavis.2014.02.009
0262-8856/

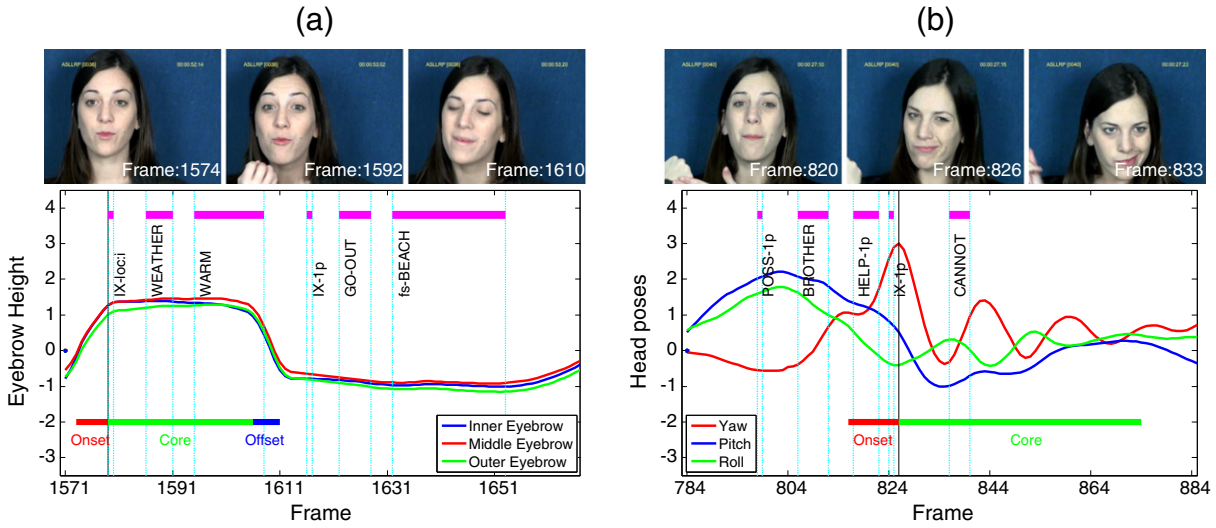

Fig. 1. Eyebrow raise and head shake: (a) If the weather is warm, I'll go to the beach; (b) As for my brother helping me, he can't.

decades. Most approaches to interpretation of non-manual information have involved construction of mapping functions between low-level features and grammatical markers. Related work will be reviewed in Section 2. However, the accurate interpretation of non-manual events (e.g., eyebrow gestures and periodic head movements) that carry linguistic information requires attention to the spatio-temporal patterning of these events over syntactic domains that may vary significantly in size and duration.

In this paper we propose an automatic recognition system for non-manual grammatical markers based on multi-scale and spatio-temporal analysis of head poses and facial expressions. Although there has been previous research focused on low-level features alone [26,30,36], or on high-level features that can be derived from those [37], our approach relies on combining both types of features. To extract the relevant features, we use a 3D deformable face tracker based on an adaptive ensemble of ASMs [17]. This 3D tracker deals well with the large head movements and occlusions of the face by the hands that occur during signing. Another advantage of the 3D face model approach is that it eliminates the need for facial pose alignment necessary in 2D approaches, which is often a source of significant errors in facial pose estimation and expression recognition. Another feature of our 3D face tracker is that it is generic and not person-specific. Our face tracker can adaptively fit each person's face without the need for person-specific training. The low-level features related to the 3D head pose and facial expressions (including eye aperture and eyebrow configuration) are based on geometry and appearance features from the 3D face tracker. The low-level features are used, in part, to recognize non-manual events such as raised/lowered eyebrows and periodic head movements (nods and shakes). The recognition of these events allows us to extract important high-level features, including the type of non-manual event, its temporal phases, and specific aspects of its motion patterns over phrasal domains. Finally, we combine the low- and high-level features based on learning methods [1] to recognize the five types of non-manual grammatical markers under consideration in this paper.

We pay particular attention to the computer-based recognition of raised and lowered eyebrows, and of periodic head nods and shakes, since they function as components of many non-manual grammatical markers in signed languages generally, and in American Sign Language (ASL) in particular. The focus on eyebrow gestures and periodic head movements is thus linguistically motivated. We further partition these events into temporal phases, to facilitate the accurate localization of the relevant grammatical markers. For example, in eyebrow events, our research to date has revealed that the linguistically significant portion of a raised or lowered eyebrow gesture begins after an anticipatory, transitional phase (which we refer to as the *onset*), during which the eyebrows raise (or lower) to the maximal extent. The *core* of the gesture begins at that point, linguistically aligned with the start of the relevant sign or group of signs associated with the grammatical marker. The eyebrows often start returning to neutral (a phase that we refer to as the *offset*) a few frames before the end of the final sign in the phrase being marked by the relevant non-manual expression. This is illustrated in Fig. 1(a). For periodic head gestures, the head tends to begin with a transitional, anticipatory movement (*onset*) that involves rotating the head to the maximal angle, so that the first head nod or shake can cover the maximal angular range, and the start of the linguistically significant portion of these periodic head movements is usually initiated from this maximally rotated position. Periodic head movements tend to involve successive head rotations of diminishing amplitude, eventually damping out (with no identifiable *offset*). This pattern is illustrated in Fig. 1(b). Hence, it is critical to separate out the preparatory phase in order to focus on the linguistically meaningful part.

In order to recognize these high-level features, we introduce a hierarchical Conditional Random Field (CRF) [20] framework to recognize raised/lowered eyebrows, head nods, and head shakes from video sequences, and to further partition these non-manual events into temporal phases, i.e., onset, core and offset (if there is one). After recognizing the non-manual events, we analyze their motion patterns, which, together with the temporal phase partitioning, are regarded as high-level features. Experiments are designed to evaluate the recognition results for eyebrow gestures and periodic head movements, as well as the further effects of such identifications on successful recognition of non-manual grammatical markers. More specifically, we investigate the improvement in performance for recognition of these markers achieved as a result of the combined use of low-level and high-level features (i.e., eyebrow and head events and their spatio-temporal patterning). Our method outperforms the baseline approach that uses only low-level features. The methodology in this paper is a significant extension of our previous approach in [25]. The innovations here include: 1) utilization of drastically improved facial tracking that combines a 3D deformable model with an ASM approach; this new method is capable of reporting reliably when tracking is not successful; 2) improvement in the computation of high-level features as sequence models, which in turn enhances the incorporation of linguistic knowledge into the recognition of grammatical markers; and 3) validation of the results through a substantial number of additional experiments, demonstrating success in the identification and temporal localization of events and their phases, and in the higher-level detection and identification of grammatical markers in ASL.