Information Sciences 428 (2018) 30–48

Contents lists available at ScienceDirect

Information Sciences

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

Independent Bayesian classifier combination based sign language recognition using facial expression

Pradeep Kumar[a,*], Partha Pratim Roy[a], Debi Prosad Dogra[b]

[a] Department of Computer Science & Engineering, Indian Institute of Technology, Roorkee, India
[b] School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, India

ARTICLE INFO

Article history:
Received 7 September 2016
Revised 3 October 2017
Accepted 23 October 2017
Available online 27 October 2017

Keywords:
Sign language recognition
Depth sensors
Hidden Markov model (HMM)
Bayesian combination

ABSTRACT

Automatic Sign Language Recognition (SLR) systems are usually designed by means of recognizing hand and finger gestures. However, facial expressions play an important role to represent the emotional states during sign language communication, has not yet been analyzed to its fullest potential in SLR systems. A SLR system is incomplete without the signer's facial expressions corresponding to the sign gesture. In this paper, we present a novel multimodal framework for SLR system by incorporating facial expression with sign gesture using two different sensors, namely Leap motion and Kinect. Sign gestures are recorded using Leap motion and simultaneously a Kinect is used to capture the facial data of the signer. We have collected a dataset of 51 dynamic sign word gestures. The recognition is performed using Hidden Markov Model (HMM). Next, we have applied Independent Bayesian Classification Combination (IBCC) approach to combine the decision of different modalities for improving recognition performance. Our analysis shows promising results with recognition rates of 96.05% and 94.27% for single and double hand gestures, respectively. The proposed multimodal framework achieves 1.84% and 2.60% gains as compared to uni-modal framework on single and double hand gestures, respectively.

1. Introduction

With the development of low cost depth sensors such as Leap motion and Kinect, there is an increased interest to develop gesture or expression based techniques in the areas of Sign Language Recognition (SLR), gaming, security, etc. This has opened-up new way of research in Human-Computer-Interaction (HCI). Sign Language is considered to be the only way of communication between deaf and hearing impaired peoples [5,15]. Sign Language is generally composed of two types of signs, i.e., Manual signs and Non-manual signs. Manual signs are composed of sign gestures which are performed using hand and finger movements, whereas non-manual signs are represented by various facial expressions, head tilting, lip pattern, mouthing, and other similar signals [16,21]. These are then added with the hand or manual signs to create a useful meaning. Thus, the meaning of a hand sign is incomplete without facial expressions. Non-manual signs plays an important role in SLR systems because they carry grammatical and prosodic information [9,43].

Sign gestures that does not contain any facial expression, are easily distinguishable by manual features while in other gestures it remains ambiguous until the availability of some facial expression. For example, in Fig. 1, where the hand gestures for the words '*who*' and '*what*' (two handed) look similar, however, it depicts different facial expressions for both sign

* Corresponding author.

https://doi.org/10.1016/j.ins.2017.10.046
0020-0255/

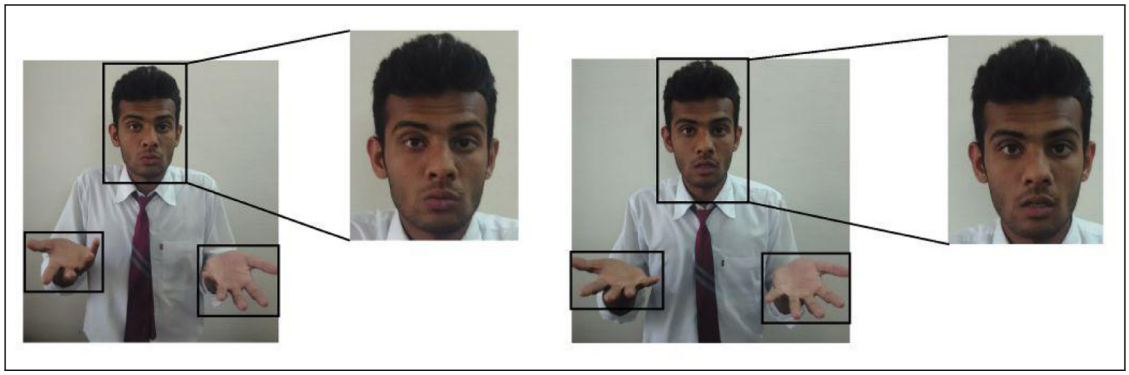

Fig. 1. Sign gesture representation for two sign word '*who*' and '*what*' which have same hand gesture but are discriminated by facial expressions.

words, where for the sign word '*who*', the signer's mouth has rounded shape and the head movement goes upward direction while for sign word '*what*', the signer's mouth is open and has raised eyebrows. Such facial information helps in discriminating different gestures and has great influence on the recognition results. Aran et al. [4] have proposed a method to integrate both manual and non-manual features simultaneously using a two-step process. First step is to classify a gesture based on manual features, whereas, in case of ambiguity in decision, a second stage classifier considers the non-manual features to resolve the ambiguity. However, their model could not be extended for the complete SLR problems.

It has been observed that the signer's facial expressions change frequently to deliver the exact meaning and sense to the performed gesture. It is possible that one facial expression may correspond to many sign gestures. Similarly, the same sign gestures may correspond to multiple facial expressions. Moreover, both hand gesture and facial expression act as complementary to each others, because the signer might deliver the gesture information through face expressions or using small hand gesture and vice-versa. Therefore, a multimodal framework is needed to combine both modalities (i.e. manual and non-manual signs) into a single recognition framework. Such a multimodal system can achieve better accuracy than existing SLR systems that fully rely on manual signs. However, facial expressions are usually ignored because of high complexities involved in the interpretation of various features and their understanding. Hence most of the existing studies are either based on recognition of manual signs (discarding the non-manual) using 2D vision based (i.e. camera) or using 3D depth data approaches with the help of depth cameras such as Leap motion and Kinect.

Zafrulla et al. [50] have developed a gesture recognition and verification system for deaf children based on 3D depth data using Kinect. Similarly, non-manual signs are also getting attention by various researchers for recognition purposes. However, majority of the existing work attempt to recognize only non-manual signs independently by recognizing the head movement [8,35], facial expressions and lip reading [28], thus, discarding the manual sign information. However, a complete SLR system requires the analysis of both manual and non-manual features, simultaneously.

While developing a SLR system with multiple modalities, it becomes difficult to extract a good feature set from the observed raw data because of different scaling and temporal variance between the modalities. Hence we obtain a unsatisfactorily biased classifier which can classify only certain classes samples because of the poor feature set [41]. In such cases, it is necessary to aggregate the decision from different modalities because in sign language both facial expression and hand gestures carries complementary information and helps in better recognition. However, there exist multiple schemes for combining the decision of multiple classifiers such as Majority voting, Bagging and Boosting. These techniques are popularly used in many real world pattern recognition tasks including gesture recognition, security, biometrics, object detection, etc. In [38,47] the authors proposed fusion algorithms, namely Multi-view Intact Space Learning (MISL) and Class Consistent Multi-Modal (CCMM) that integrate homogeneous features by combining multiple views and input modalities to improve the performance of recognition problems. Hong et al. [19] have proposed dynamic captioning methodology in videos to assist hearing impaired people using heterogeneous features such as face detection and recognition, visual saliency analysis, and text-speech alignment. For better understanding of the video contents their approach highlights the scripts word-by-word by aligning them with the speech signal and also illustrate the variation of voice volume. Likewise, the authors in [20] have proposed the use of audio features for indexing and retrieval of video contents. However, these approaches either rely on the provided feature set or the confidence score and consequently, provide a limited improvement in recognition performance.

In this study, we present a feature independent classifier combination that is based on the results of different classifiers for different modalities using Independent Bayesian Classification Combination (IBCC). The methodology takes decisions of uni-modality, i.e., hand gesture and facial expression as input. Then it uses IBCC algorithm to increase the overall performance of the system. The scheme has proven to be advantageous over existing decision fusion methodologies because it is directly based on classifier's decision instead of confidence, probability and likelihood scores, etc. Simpson et al. [39] have derived a variational inference algorithm for IBCC and applied it to detect cluster of users depicting the same behavior based on the decision of multiple classifiers.