Neural Networks 36 (2012) 157–166

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Non parametric, self organizing, scalable modeling of spatiotemporal inputs: The sign language paradigm

G. Caridakis*, K. Karpouzis, A. Drosopoulos, S. Kollias

Intelligent Systems, Content and Interaction Lab, National Technical University of Athens, Greece

ARTICLE INFO

Article history:
Received 26 July 2011
Received in revised form 3 October 2012
Accepted 4 October 2012

Keywords:
Spatiotemporal pattern recognition
Self organizing maps
Markov models
Probabilistic
Sign language recognition

ABSTRACT

Modeling and recognizing spatiotemporal, as opposed to static input, is a challenging task since it incorporates input dynamics as part of the problem. The vast majority of existing methods tackle the problem as an extension of the static counterpart, using dynamics, such as input derivatives, at feature level and adopting artificial intelligence and machine learning techniques originally designed for solving problems that do not specifically address the temporal aspect. The proposed approach deals with temporal and spatial aspects of the spatiotemporal domain in a discriminative as well as coupling manner. Self Organizing Maps (SOM) model the spatial aspect of the problem and Markov models its temporal counterpart. Incorporation of adjacency, both in training and classification, enhances the overall architecture with robustness and adaptability. The proposed scheme is validated both theoretically, through an error propagation study, and experimentally, on the recognition of individual signs, performed by different, native Greek Sign Language users. Results illustrate the architecture's superiority when compared to Hidden Markov Model techniques and variations both in terms of classification performance and computational cost.

1. Introduction

Challenges and open issues concerning the applicability and extensibility of approaches that aim at tackling spatiotemporal problems include resistance to noise and variability (w.r.t. user/repetition gesture performance) in the input channel, computational efficiency of the recognition scheme adopted, large scale dictionary registration and recognition and dictionary extension without the need of extensive retraining. Additionally, fusion of multiple modalities and usage of arbitrary, or experimentally defined, initialization parameters, such as the number of HMM states significantly influences performance, generalization and adaptability of the majority of approaches. In the proposed scheme, dedicated per modality classifiers are trained in order to model different recognition aspects and are consequently fused at decision level. This approach resembles Boosting of weak classifiers; however, the classifiers used in the proposed scheme are suitable for tackling particular aspects of the recognition task and not weak, generic classifiers. Input variability is addressed through the flexibility, provided by state transition probability dispersion during Markov chains training and by optimal path search performed during classification, both based on SOM neighborhood properties. Spatial modeling is achieved using a SOM, trained with a representative sample of hand positions during signing. Spatial modeling is performed once, annulling the need for exhaustive retraining when an unknown class is introduced to the vocabulary. Given that the initial training set is representative in terms of signing space distribution, no additional training is required, since the signing space has been well modeled. The SOM nodes neighboring relation, formed during training, consists of a crucial characteristic of the overall training and classification process. It is driving the adaptive nature of the overall approach, tackling large scale vocabulary application issues. A modified algorithm is used for calculating the Levenshtein distance, also taking into account the similarity of sequence's symbols, addressing the problem of potential variation or noise in the input channel.

Sign Language (SL) is the linguistic system used by the hearing disabled group, in order for the members of the group to communicate amongst themselves and also with hearing able people. Unlike spoken languages, sign languages are heavily based on iconicity to convey meaning. A morphosyntactic structure is employed to express linguistic relations in 3D space. Concepts are represented by signs, the basic grammatical unit of a sign language, forming a visual natural language. Sign language analysis and recognition can be viewed as a spatiotemporal problem incorporating the issues discussed previously, as well as a wide

* Corresponding author.
E-mail addresses: gcari@image.ntua.gr (G. Caridakis), kkarpou@image.ntua.gr (K. Karpouzis), ndroso@image.ntua.gr (A. Drosopoulos), stefanos@image.ntua.gr (S. Kollias).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.10.001

range of concepts. It includes pure image analysis tasks, ranging from locating and tracking the face and hands of the signer up to notions related with semantics and context, usually found in natural language processing paradigms.

Validation of the proposed approach is performed through its application to the SL analysis problem. At first we locate the head and hands of the signer in order to extract features related to both handshape and hand location. Then, hand positions are used to train Self Organizing Maps, so as to effectively represent the signing space, tackling the spatial aspect of the recognition task. First order Markov chains, which use the SOM units as states, are used to cope with the temporal aspect. Fusion of SOM and Markov chains is performed by a greedy algorithm seeking to make a locally optimal choice at each stage and converge to a global solution. Intra and inter user spatial or performance variation and random errors in the input stream are tackled by incorporating the neighborhood property of the models' states in the overall classification process, thus enhancing the overall architecture with robustness and adaptability. Separate classifiers, namely Markov models for hand position and movement and Hidden Markov Models for handshape features are fused on a decision level, in a committee-machine-like setup, further ensuring stability of the recognition process. Application of the proposed architecture has a low computational cost, making it therefore suitable for realtime applications. Experimental results, discussed in Section 4.2 and performed on two datasets, a synthetic and a Greek SL corpus, illustrate the architecture's superiority both in terms of classification performance and computational cost over popular techniques such as Hidden Markov Models and their variations (Multi-Stream, Parallel and Product HMMs). Initial validation on the synthetic dataset have been presented in Caridakis, Karpouzis, Drosopoulos, and Kollias (2010) and current work builds on this and enhances the approach with:

- transition probability spreading during training providing robustness against noise and variability in the input channel
- incorporation of a novel distance calculation algorithm based on the Levenshtein distance metric which takes into consideration similarity of symbols
- a distributed approach, tackling individual aspects of the handshape such as boundary and region, aiming to model an extremely complex pattern, such as the handshape, especially for 2D projection, and challenging, in terms of automatic recognition, finger configurations
- incorporation of multiple modalities and appropriate multimodal fusion; the latter balances each stream's contribution to the final decision according to respective unimodal classification results
- extensive experimentation with datasets featuring native signers that illustrate the architecture's superiority to current state of the art schemes, both in terms of classification performance and computational cost.

The remainder of the paper is organized as follows: Section 2 discusses aspects, challenges and previous work related to automatic Sign Language recognition, by critically reviewing each approach, bringing forth the focus and the strong points of each article. Section 3 introduces the proposed architecture and is roughly divided into the learning process (Section 3.2) and the classification process (Section 3.6). The proposed approach is validated: (a) theoretically in Section 3.6.1, by studying the propagation of error when a random error is introduced in the input stream and (b) experimentally in Section 4, by applying the learning and classification scheme on the Greek Sign Language Corpus (Efthimiou & Fotinea, 2007) featuring three native signers performing representative lemmata of the Greek Sign Language. Finally, the article is summarized in Section 5 where future directions of the presented research work are also discussed.

2. Machine learning and SL recognition

An abundance of automatic sign language recognition techniques can be found in the literature, differentiating in terms of input streams, extracted features, vocabularies, signer dependence, isolated or continuous recognition (Ong & Ranganath, 2005; von Agris, Zieren, Canzler, Bauer, & Kraiss, 2008). The input stream can be either based on the use of motion capture (direct-measure device) data gloves (Wang, Leu, & Oz, 2006; Zhang, Yao, Jiang, Zhao, & Sun, 2005) or consist of visual signals. Datagloves are quite expensive and intrusive, however they constitute a robust and accurate way of capturing 3D hand location and finger flexion in real time. Motion capture is used in Vogler (2002) while time-of-flight camera is employed in Fujimura and Liu (2006) and visual and device inputs are fused in Brashear, Starner, Lukowicz, and Junker (2005). In all approaches features are extracted from the gestured input stream mainly based on the position of the dominant right hand. Usually, when motion capture is employed, the 3D position is included in the features set, but for vision based approaches, only the 2D projection of the hand position can be extracted and 3D can only be calculated in conjunction with stereo vision. The position of the hand is relative to some reference point, for example, the head of the user or his/her back in case of data capturing by placing an additional sensor on the back of the signer. Another important issue on which automatic sign language recognition is based on, is the vocabulary size of the experimental corpora. In most cases, the experimental dataset is composed of a quite restricted number of signs ($\approx$50); only in a small number of cases this is extended (Derpanis, Wildes, & Tsotsos, 2004; Fujimura & Liu, 2006; Gao, Fang, Zhao, & Chen, 2004a; Zhang et al., 2005). Additionally, signer dependence and sign variation are decisive aspects when trying to implement architectures into real world. Signer independence, and ways to tackle both intersigner variation in the performance of the signs and grammatical idioms of signer groups, are vital for achieving good recognition rates in an arbitrary setting by an unregistered user.

An important issue in the wide research area of automatic sign language recognition is the vocabulary size of the experimental corpora used to verify the robustness and generalization capabilities of the proposed systems, as can be seen in Table 1. Most articles construct their experimental dataset using a quite restricted number of signs, varying from 10 to 65, while others extend their vocabulary, to what would be a more representative sample of the respective sign language, but still use between 164 and 274 signs. The articles that approximate a universal recognizer are those who reach impressive vocabulary sizes that enumerate up to more than 5000 signs. Publications belonging to the last group focus mainly on large vocabulary recognition and issues related to a complete, real time system that could support automatic sign language recognition. The number of repetitions performed for each vocabulary entry is also important, as is the training/testing sample ratio. Typically each sign is repeated 5–10 times e.g. Bowden, Windridge, Kadir, Zisserman, and Brady (2004), Cooper and Bowden (2007a, 2007b), Kadir, Bowden, Ong, and Zisserman (2004), Ma, Gao, and Wang (2000), Wang and Gao (2000), Wang, Gao, and Shan (2002) and Zieren and Kraiss (2005), while there are cases where more (Hernandez-Rebollar, Kyriakopoulos, & Lindeman, 2004; Imagawa et al., 2000; Lee & Tsai, 2007) or fewer (Assan & Grobel, 1998; Fang, Gao, & Ma, 2001; Infantino, Rizzo, & Gaglio, 2007; Su, 2000; Wang, Chen, Wang, & Gao, 2006) repetitions are performed for each lemma in the restricted, experimental vocabulary.

Residing in the heart of each recognition system, the classification architecture is considered the most important one. Although a plethora of works propose a single off the shelf classifier, there is also a significant number of approaches that utilize a combination of classification schemes. Such schemes include HMM and variants, Neural Networks, Boosting Techniques, Linear Models, Tree structures, Clustering and State Sequence Comparison.