PRODUCT-HMMS FOR AUTOMATIC SIGN LANGUAGE RECOGNITION

Stavros Theodorakis, Athanassios Katsamanis and Petros Maragos

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece

ABSTRACT

We address multistream sign language recognition and focus on efficient multistream integration schemes. Alternative approaches are investigated and the application of Product-HMMs (PHMM) is proposed. The PHMM is a variant of the general multistream HMM that also allows for partial asynchrony between the streams. Experiments in classification and isolated sign recognition for the Greek Sign Language using different fusion methods, show that the PHMMs perform the best. Fusing movement and shape information with the PHMMs has increased sign classification performance by 1,2% in comparison to the Parallel HMM fusion model. Isolated sign recognition rate increased by 8,3% over movement only models and by 1,5% over movement-shape models using multistream HMMs.

Index Terms— sign language recognition, Product HMM, integration, asynchrony, HMM+

1. INTRODUCTION

Sign languages, i.e., languages that essentially convey information via visual patterns, commonly serve as an alternative or complementary mode of human communication [1]. Visual patterns, as opposed to the audio ones used in the oral languages, are formed by hand shapes and hand or general body motion, lip movements and facial expressions. Their expressiveness facilitates human interaction and exchange of information not only in the existence of hearing-impaired people but also in situations where speech is impractical, e.g., in loud workspaces. However, efficient communication by these means is only feasible between specially trained interacting parties. In this context, automatic sign-to-text and text-to-sign translation can be viewed as the intermediate technological modules that can partially lift this restriction. In our work, we address the problem of automatic sign language recognition (sign-to-text) and we investigate schemes to efficiently handle the multistream/multimodal character of this task.

The field of sign language recognition is certainly in the focus of quite intense research lately [2]. It is considered to be a multilevel problem and it poses significant challenges regarding data collection, visual processing, i.e., hand localization, tracking and feature extraction, information stream modeling for recognition and general language modeling. Bowden and his colleagues, [3], to cope with limited training data, proposed a two stage classification procedure and achieved 97.67% classification rate for a lexicon of 43 words using only single instance training. At the initial stage, a high level description of hand shape and motion was extracted while at the second classification stage the temporal transitions of individual signs were also taken into consideration. For this purpose, a classifier bank of Markov chains was applied in combination with Independent Component Analysis.

For modeling, many previous approaches build on experience gained from the corresponding area of speech recognition. Bauer et al., [4], based on a 97-sign vocabulary of the German Sign Language used Hidden Markov Models (HMMs) creating one model per sign. The signer wore colored cotton gloves and the extracted features included the position of both hands, distances between all fingers and distances between the hands. They also implemented a language model and finally achieved 93.2% recognition accuracy.

To effectively handle the multiple information streams apparent in sign languages, e.g., left and right hand or head movement, Vogler and Metaxas [5] moved one step further and applied the so-called Parallel HMMs. Their basic assumption is that the involved streams essentially evolve independently. Though it is accepted that this is due to an engineering tradeoff and not valid in reality it can lead to significant improvements compared to the single stream HMM approach. Their experiments were on continuous American sign language recognition based on a 22-sign vocabulary. They broke down signs into their constituent phonemes using the basic ideas of the Movement-Hold model [6]. They also broke up the features into movement and shape channels. The former comprises features related to the location and movement of the hands while the latter describes the handshape. To handle these two channels they used Parallel HMMs (PaHMMs) and achieved 96.15% word accuracy increasing recognition rate by 1.6% over movement-only models.

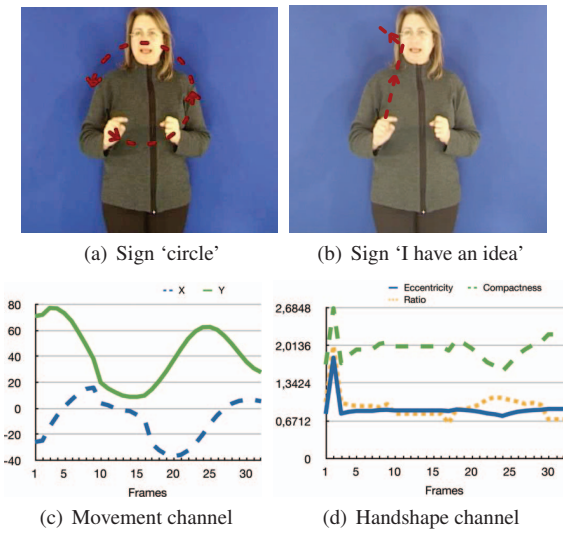
In our current work, we focus on the multistream character of sign languages. We investigate alternative multistream integration approaches for sign language recognition in an effort to account for possible interstream interactions. We regard hand movement and hand shape as separate streams and, motivated by analogous work in audio-visual speech recognition, we apply the so-called Product HMM to model each sign. Our classification and recognition experiments for the Greek Sign Language (GSL) and a 93-sign vocabulary demonstrate that it can be quite beneficial to consider the separate information streams as partially interacting and not completely independent.

2. MULTISTREAM FUSION FOR SIGN LANGUAGE RECOGNITION

The goal of automatic sign language recognition, may be viewed as the recovery of a sign sequence S from the sequence of observations O. Given the multicue nature of the visual patterns forming the sign sequence it is essential that these observations represent information conveyed by all involved information channels. Hand shapes and movement, the body pose, face expression and head movement should all be taken into consideration in the general case [2]. Neglecting for example the hand movement would not allow disambiguation of the GSL signs for ‘circle’ and ‘I have an idea’ since the

This work has been supported by the European research program DICTA-SIGN (grant FP7-ICT-3-231135) and partially by grant ΠΕΝΕΔ-2003ΕΔ866 [cofinanced by E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%)]

same handshape is used for both, i.e., fist with the index finger projected shown in Figs. 1(a),1(b). The movement of the dominant hand is quite distinct for the two signs though. Observations from these different channels are supposed to form separate streams which may then be exploited in appropriate fusion schemes, either synchronous or asynchronous.



(a) Sign 'circle' (b) Sign 'I have an idea'

(c) Movement channel (d) Handshape channel

Fig. 1. Examples of the Greek Sign Language signs 'circle' and 'I have an idea'. In the bottom row, observations for the sign 'circle' are shown: on the left, the coordinates (x,y) of the dominant hand from the movement channel and, on the right, the eccentricity, compactness and ratio from the handshape channel.

2.1. Synchronous Fusion

2.1.1. *Feature Fusion*

The so-called feature fusion scheme, also known as early integration, is the simplest case and is based on the assumption that the involved streams are synchronous and there is no need for any kind of stream-dependent treatment. Given time-synchronous movement and handshape feature vectors $\mathbf{o}_{m,t}$ and $\mathbf{o}_{h,t}$, respectively, feature fusion considers

$$\mathbf{o}_t = [\mathbf{o}_{m,t}, \mathbf{o}_{h,t}] \in R^{\,l}\,,\; l = l_m + l_h \tag{1}$$

as the joint observation of interest, modeled by a single-stream HMM as:

$$P[\,\mathbf{o}_t \mid c\,] = \sum_{j=1}^{J_c} w_{cj}\; N_l(\mathbf{o}_t\;;\;\mathbf{m}_{cj}\;,\;\mathbf{s}_{cj}) \tag{2}$$

where $c \in C$ denote the HMM context dependent states, J_c denotes the number of mixtures, w_{cj} are the mixture weights and $N_l(\mathbf{o}\;;\;\mathbf{m}\;,\;\mathbf{s}\;)$ is the l-variate normal distribution with mean $\mathbf{m}$ and a diagonal covariance matrix $\mathbf{s}$.

2.1.2. *State-Synchronous Multi-Stream HMM*

In state-synchronous multi-stream HMMs though still the streams share common underlying state dynamics (Fig. 2 for the case of movement and handshape streams), each stream may be separately weighted. The observation likelihood of the multi-stream HMM is the product of the observation likelihood of each single-stream raised to an appropriate stream weight [7]:

$$P[\,\mathbf{o}_t \mid c\,] = \prod_{s\in\{M,H\}} [\sum_{j=1}^{J_{sc}} w_{scj}\; N_{l_s}(\mathbf{o}_{st}\;;\;\mathbf{m}_{scj}\;,\;\mathbf{s}_{scj})\,]^{\,\lambda_{sct}} \tag{3}$$

where λ_{sct} are the stream weights, that are positive and are related to the reliability of the information each stream carries or to other prior discriminative criteria. For example, the involvement of the stream carrying information for the movement of the non-dominant hand, e.g., the left one for the signs shown in Fig.1, should be considered of secondary importance and thus weighted less.

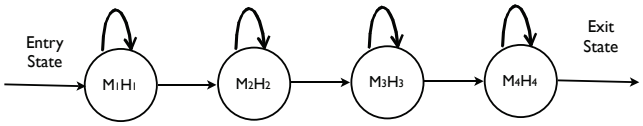


Fig. 2. Example of a multi-stream HMM with 2 streams and 4 states. Each state is shared by both the movement (M) and handshape streams (H).

2.2. Asynchronous Fusion Approaches

2.2.1. *Parallel HMM*

The feature fusion and the synchronous multi-stream model we discussed enforce state synchrony between information streams. This is quite restrictive and it has been shown that can limit recognition performance [5].

Parallel HMMs (PaHMMs) are on the opposite extent in terms of interstream synchrony related restrictions. They are an extension to HMMs and they have been applied for sign language recognition [5] based on the assumption that the separate streams evolve independently from one another with independent output. As a consequence, it is possible to train the single-stream HMMs completely independently from the other streams, and put streams together at the recognition time using the appropriate stream weights (see Fig. 3).

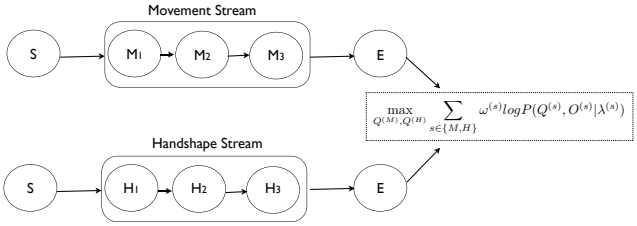


Fig. 3. Example of a Parallel HMM with 2 streams and 3 states in each stream.

2.2.2. *Product HMM*

Somewhere between the completely synchronous fusion scheme and Parallel HMMs lies the so-called product HMM [8, 9] (see Fig. 4), which has been quite successfully applied in audiovisual speech

recognition. It allows streams to be in asynchrony within the model but forces them to be in synchrony at the model boundaries. As is shown in Fig. 4 each state of PHMM is a combination of the two streams. The Product-HMM is essentially obtained as the 'product' of the single-stream HMMs.

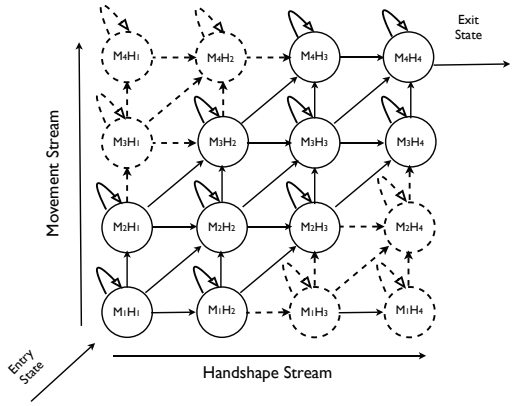


Fig. 4. Example of a product HMM with 2 streams and 4 states in each stream. The movement and handshape streams are denoted by M_x and H_y, where x,y are the states of the movement and handshape stream model respectively.

Product HMMs also permit the control of the degree of asynchrony between streams by discarding states from the lattice, e.g., in Fig. 4, if we discard from the lattice all the states and transitions marked with dashed lines we restrict the asynchrony between the two streams to be up to 1 state. In the two extreme cases when only the states that lie on the diagonal or all the states in the lattice are kept, the model becomes equivalent to the synchronous multi-stream and Parallel HMM respectively.

3. EXPERIMENTS

Evaluation of the presented fusion schemes was performed experimentally. Our sign database comprises 691 training and 110 test sign instances over a vocabulary of 93 signs. These were randomly selected and results were acquired for 10 repetitions (repeated hold-out method). We performed all training and testing using the Hidden Markov Model Toolkit (HTK) [11].

3.1. System Overview

In our system, a single camera in front of the signer is used for video acquisition. Then an image processing system is applied for segmentation and different features are extracted [10]. We divide our features into two different channels (movement and handshape). The movement channel consists of the location and movements of the signer's hands and head. Movement feature vectors are of dimension 6 and comprise the position coordinates of the hand normalized with reference to the position coordinates of the head, the first derivatives of these and the distance between the two hands. The handshape channel essentially includes features related to the shape of the dominant hand. Handshape feature vectors are of dimension 4 and are region-based features. These include the area of the handshape, its eccentricity, its compactness which is the ratio of its area and its perimeter squared and the ratio of its minor and major axis lengths (see Fig. 1).

3.2. Classification Experiments

As mentioned in Section 2.2, state synchrony between movement and handshape channels is quite restrictive, so we tried to model stream asynchrony using Product HMMs (PHMMs), taking advantage of the ability they have to control the degree of asynchrony between streams. Classification results per experiment repetition are depicted in Fig. 5 and allow the comparison among the various stream fusion approaches, i.e., using state-synchronous multi-stream HMMs, PaHMMs and PHMMs. As we can see, using PHMMs we obtain the best performance. Recognition accuracy is increased by 56% over handshape-only, 5,97% over movement-only models (these results are not shown in the graph), 1,37% over synchronous multi-stream and 1,19% over PaHMMs (absolute percentage differences).

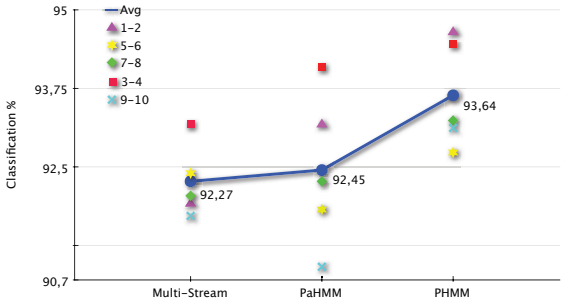


Fig. 5. Classification results per experiment repetition for fusion using state-synchronous Multi-stream HMM, Parallel HMM (PaHMM) and Product HMM (PHMM).

In our experiments using Product HMMs (PHMM) we first train each stream separately based on single-stream observations, then combine them creating the Product HMM and re-train both streams together. Stream weights $\lambda_m = 5$, $\lambda_h = 3$ and asynchrony between the two streams up to 2 states were applied.

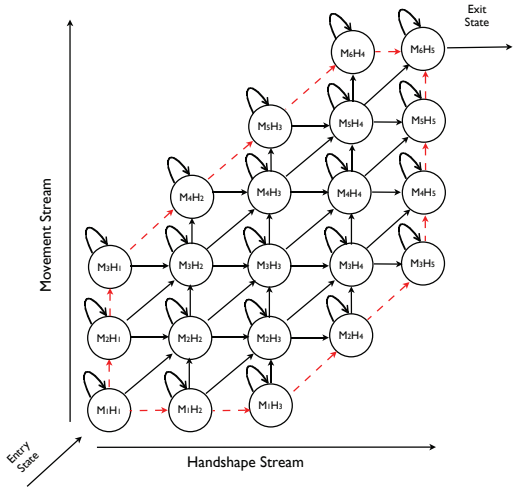


Fig. 6. A Product HMM with 2 streams using 5 and 6 states for each stream respectively and asynchrony between the two streams up to 2 states. The red dashed line follows the common paths in the Product HMM.

In Fig. 6 we show with dashed line the paths mostly followed by the test data in the Product HMM when only 2-state asynchrony has been allowed between the two streams. As we can see extreme paths are most often followed, taking advantage of Product HMMs property that allows state asynchrony between streams. By increasing the degree of asynchrony between the streams over 2 states, we observe that more centered paths are followed and recognition accuracy decreases. We may thus assume that movement and handshape streams are neither synchronous neither completely independent and the Product HMMs offers sufficient modeling flexibility to account for this specifity.

3.3. Isolated Sign Recognition Experiments

Beyond classification, we also experiment with isolated sign recognition. We test feature fusion method (FF) as we describe in Section 2.1.1, fusion using one state-synchronous multi-stream HMM (MS1) training each stream separately based on single-stream observations and subsequently combine them as in Eq.(3) as well as one state-synchronous multi-stream HMM (MS2) training both streams together using a multi-stream like the one in Fig. 2. Recognition results (repeated holdout method) are depicted in Fig. 7. Error bars indicate the range of accuracies from the maximum to the minimum one in the repeated holdout method.

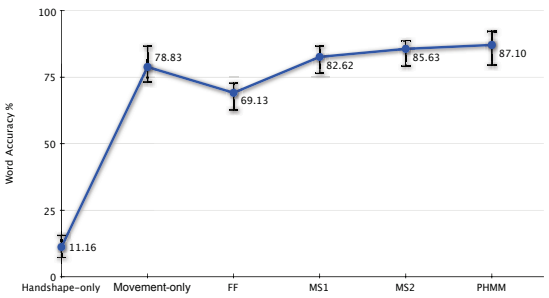


Fig. 7. Isolated sign recognition results for Handshape-only, Movement-only models and fusion using feature fusion (FF), synchronous Multi-stream HMMs trained in two ways (MS1, MS2), and Product HMM (PHMM). We also depict the maximum and the minimum performance from the repeated holdout method (in the form of errorbars).

As we can see in Fig.7, feature fusion decreases recognition accuracy related to movement-only models because movement and handshape streams carry different kind of information and also because we don't use stream weights and both streams contribute equally. For both MS1 and MS2 models, stream weights $\lambda_m = 5$, $\lambda_h = 3$ were used and as expected the MS2 models outperformed the MS1 because in MS1 models, the two single-streams HMMs are trained asynchronously whereas Eq.(3) assumes that the HMM streams are state synchronous. When using Product HMMs we obtain the best performance increasing recognition accuracy 8,27% over movement-only models, 18% over feature fusion, 4,49% over MS1 and 1,47% over MS2. Our results show a modest but promising increase which may be made more prominent by integrating more different modalities. This objective is part of our current work in continuous sign language recognition.

4. CONCLUSIONS

We have proposed the application of the so-called Product HMMs for efficient multistream fusion for sign language recognition. We have investigated alternative fusion schemes and we have demonstrated superior performance of the proposed system. Product HMMs allow for information stream asynchrony but at the same time also account for possible interstream interactions. As opposed to parallel HMMs, they do not consider the involved streams as completely independent. We have reported experiments in sign classification and isolated sign recognition for the Greek Sign Language with promising results. Currently we are working on extending this approach for continuous sign language recognition.

5. ACKNOWLEDGMENTS

The authors would like to thank O. Diamanti from the National Technical University of Athens for the visual processing frontend, G. Caridakis and D. Dimitriadis for useful discussions, and E. Efthimiou and S.-E. Fotinea from the Institute for Language and Speech Processing for providing the Greek Sign Language database.

6. REFERENCES

[1] D.McNeill, "Hand and mind: what gestures reveal about thought," *University of Chicago Press*, Chicago, 1992.

[2] S.C.W. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *Trans. Pattern Anal. Mach. Intellig.*, vol. 27, no. 6, pp. 873–891, 2005.

[3] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," in *Proc. 8th European Conf. Computer Vision*, 2004.

[4] B. Bauer, H. Hienz, and K-L. Kraiss, "Video-based continuous sign language recognition using statistical methods," *Int'l Conf. Pattern Recognition*, 2000.

[5] C. Vogler and D. Metaxas, "Handshapes and movements: Multiple-channel american sign language recognition," in *Gesture Workshop*, 2003, pp. 247–258.

[6] S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," *Sign Language Studies*, vol. 64, pp. 195 – 277, 1989.

[7] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Tr. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[8] C. Neti G. Gravier, G. Potamianos, "Asynchrony modeling for audio-visual speech recognition," *Proc. of the 2 int'l conf. on Human Language Technology Research*, 2002.

[9] G. Potamianos J. Luettin and C. Neti, "Asynchromous stream modeling for large vocabulary audio-visual speech recognition," *ICASSP 2001 IEEE Int'l Conf.*, vol. 1, pp. 169–172, 2001.

[10] O. Diamanti and P. Maragos, "Geodesic active regions for segmentation and tracking of human gestures in sign language videos," in *Proc. ICIP*, 2008.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Ltd., Cambridge,1999.