Modality Combination Techniques for Continuous Sign Language Recognition

Jens Forster, Christian Oberdörfer, Oscar Koller, and Hermann Ney

Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{forster,oberdorfer,koller,ney}@cs.rwth-aachen.de

Abstract. Sign languages comprise parallel aspects and use several modalities to form a sign but so far it is not clear how to best combine these modalities in the context of statistical sign language recognition. We investigate early combination of features, late fusion of decisions, as well as synchronous combination on the hidden Markov model state level, and asynchronous combination on the gloss level. This is done for five modalities on two publicly available benchmark databases consisting of challenging real-life data and less complex lab-data, the state-of-the-art typically focusses on. Using modality combination, the best published word error rate on the SIGNUM database (lab-data) is improved from 11.9% to 10.7% and from 55% to 41.9% on the RWTH-PHOENIX-Weather database (challenging real-life data).

Keywords: Automatic Sign Language Recognition, Modality Combination, Hidden Markov Model, Single View Video, Speech Recognition.

1 Introduction and Related Work

Sign languages are natural, visual languages used by the Deaf and, in part, the hard-of-hearing communities world-wide as their main communication tool. In contrast to spoken languages, sign languages convey their meaning not only by one information channel/modality but by a number of modalities such as body pose, facial expression, as well as both hands which occur in parallel. This non-sequential aspect of sign languages poses a challenge to statistical sign language recognition systems which are typically based on speech recognition systems and expect signs to be composed of a sequence of sub-units.

The combination of not perfectly synchronous modalities has been tackled in the area of audio-visual speech recognition where audio information and visual information of the mouth region are combined to improve speech recognition in noisy environments. Verma et al. [17] investigated early and late integration of modalities finding that in this context late integration improves on early integration. Different structures of hidden Markov models (HMM) such as multi-stream and product HMM [9,10], pairwise and asynchronous HMM [2], and factorial and coupled HMM [12,11] have been investigated taking into account the dominant role of the audio stream. Although those results indicate that the

J.M. Sanches, L. Micó, and J.S. Cardoso (Eds.): IbPRIA 2013, LNCS 7887, pp. 89–99, 2013.

coupled HMM is a good model choice, this finding cannot be transferred directly to sign language recognition because here modalities are more strongly decoupled than in audio-visual speech recognition and there is no "master" modality.

Addressing sign language respective gesture recognition, Tran et al. [16] use multi-stream HMM and early fusion of posture features for continuous action recognition showing improved detection accuracy for a set of five actions. Vogler and Metaxas [18] investigate parallel hidden Markov models (PaHMM) for recognition of continuous American Sign Language (ASL) using cyber-gloves for feature extraction. They report an improvement from 6.7% to 5.8% word error rate (WER) for 22 signs using 400 training and 99 test sentences. Deng and Tsui [3] used PaHMMs for isolated ASL sign recognition and combined hand trajectories and right hand postures. Wang et al. [19] applied PaHMMs to the task of isolated, large vocabulary, recognition of Chinese sign language with 2435 signs using data gloves for data acquisition and hand orientations, shapes and positions as features achieving a recognition error rate of 16%. Theodorakis et al. [15] evaluated product HMMs for the recognition of 93 isolated, Greek sign language signs and reported that an asynchronous combination of features outperformed a synchronous combination. Ong et al. [13] proposed using boosted sequential trees for isolated sign recognition. Although not explicitly tested, the feature selection process allows to combine different modalities within the learned models.

So far, work on modality combination for sign language has a number of shortcomings due to the difficulty of the problem and a lack of suitable databases. First, the majority of work considers only the recognition of isolated signs instead of continuous sign language. Second, the used databases are typically created in a research lab specifically for pattern recognition or linguistic research which is typically removed from sign language in the wild. Third, data/feature acquisition is often simplified by using cyber or colored gloves, body markers, or calibrated stereo cameras while in the majority of real-life scenarios only a single camera is available. To tackle part of the above shortcomings, we compare four modality combination techniques for the recognition of continuous German sign language from single-view video. The modality combination techniques are early combination of the modalities in the feature space (feature combination), late combination of system decisions (system combination), as well as the combination on state level for HMMs (synchronous combination) and the combination on gloss level (asynchronous combination). Experimental results on the publicly available, large vocabulary, German sign language databases SIGNUM (lab-data) and RWTH-PHOENIX-Weather (challenging "real-life" data recorded from public TV) indicate that synchronous and asynchronous combinations are the methods of choice outperforming the best published WERs on both databases.

2 System Overview and Features

The sign language recognition system used in this paper is based on the freely available open source speech recognition system RASR [14]. Given a sequence of features $x_1^T = x_1, \ldots, x_T$, the system searches for an unknown sequence of words

$w_1^N = w_1, \ldots, w_N$ for which the sequence of features x_1^T best fits the learned models. To this end, the posterior probability $p(w_1^N|x_1^T)$ over all possible word sequences w_1^N with unknown number of words N is maximized. Using Bayes' decision rule, casting the *visual model* $p(x_1^T|w_1^N)$ as the marginal over all possible HMM temporal state sequence $s_1^T = s_1, \ldots, s_T$ for word sequence w_1^N, as well as assuming a first order Markov dependency and maximum approximation,

$$x_1^T \rightarrow [w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}} \left\{ p(w_1^N) \max_{s_1^T} \left\{ p(x_t|s_t, w_1^N) \cdot p(s_t|s_{t-1}, w_1^N) \right\} \right\} \quad (1)$$

where $p(w_1^N)$ is the *language model*. Time-synchronous word-conditioned tree search with dynamic programming is used expanding all *state hypotheses* $Q_v(t, s)$ in all trees for each time step t.

$$Q_v(t, s) = \max_{\sigma} \{ p(x_t, s|\sigma) \cdot Q_v(t-1, \sigma) \} \quad (2)$$

denoting the joint probability for the best partial path up to time t ending in state s with the best predecessor state σ and predecessor word v. In case of a word end state the state hypotheses of the previous time step are weighted by the language model to obtain new state hypotheses for the next word.

$$Q_v(t, s = 0) = \max_u \{ p(v|u) \cdot Q_u(t, S_v) \}, \quad (3)$$

where u is the predecessor word at the previous time step, v is the predecessor of the new state hypothesis, S_v is the ending state of word v, $s = 0$ is the virtual starting state, and $p(v|u)$ is a bigram language model for simplification. A new state tree with predecessor word v and virtual starting state $s = 0$ is then started and the whole process repeated until the end of the current sentence is reached.

Features: In this work, the focus is on the combination of different modalities and not answering the questions which modalities or which feature to use for a given modality. Features and their respective parameters have been empirically optimized for the databases described in Section 4. Five modalities are addressed in this work:

a) Full Upper Body and Body Pose
b) Right Hand Shape and Orientation
c) Left Hand Shape and Orientation
d) Facial Expression
e) Movement Trajectory and Position

Modality **a** is represented by a PCA-reduced, temporal stack of $\pm w$ video frames encoding complete body-pose and pose change over time. PCA is applied for each color channel separately. For the SIGNUM database, original video frames are scaled to 32×32 pixels and to 53×65 pixels for the RWTH-PHOENIX-Weather (PHOENIX) database. w is set to 4 respectively 2 for SIGNUM and PHOENIX and the final feature dimension is 210.

Modalities **b** and **c** are represented by HoG3D features [8] extracted using a non-dense spatio-temporal grid from spatio-temporal volumes of ± 4 hand

patch images cropped automatically using dynamic programming tracking [4] for SIGNUM (tracking error rate: 10.2% (right-hand), 32.4% (left-hand) on an annotated subset). Ground-truth coordinates are used for PHOENIX.

A person independent active appearance model of the face is fitted to each video frame resulting in 109 fitting parameters which form the features for modality **d**. Both, HoG3D and face features are stacked over ± 4 frames for SIGNUM and ± 2 frames for PHOENIX, and reduced to 200 dimensions via PCA.

Finally, modality **e** is represented by the eigenvectors and eigenvalues of the movement of the right hand within a time window of $2\delta + 1$ frames. Additionally, the hand's position w.r.t. the nose is added. Because the movement trajectory has only little discriminative power, the final feature vectors are concatenated to the feature vectors of modality **b**.

3 Modality Combination Techniques

In the following, we discuss feature, synchronous, asynchronous, and system combination. The training of visual models is identical for all strategies.

Feature Combination
The idea in feature combination is to concatenate I feature sets $(x_{t,1}, \ldots, x_{t,i}, \ldots, x_{t,I})$ to one vector $x_t = [x_{t,1} \cdots x_{t,i} \cdots x_{t,I}]^\top$ and use it to train a single HMM model. While the advantage of this approach is the ability to use the same approach for recognition as used for single feature sets, its main drawback is its inflexibility. For each combination of feature sets a new model has to be trained and evaluated. Furthermore, the feature dimension is strongly increased making dimension reduction a necessity. For the results reported in this work PCA reduction is always performed after feature concatenation.

Synchronous Combination
A more flexible approach to combine different modalities for recognition is the synchronous combination using *multi-stream* HMMs. In contrast to feature combination, modalities are trained separately and only joined in the recognition process. At each time step t, all visual models are evaluated, each with its own feature set, and the visual probabilities are combined to one probability:

$$x_1^T \rightarrow [w_1^N]_{\text{opt}} = \underset{w_1^N}{\text{argmax}} \left\{ p(w_1^N) \cdot \max_{s_1^T} \left\{ \prod_{t=1}^{T} \prod_{i=1}^{I} p(x_{t,i}|s_t, w_1^N)^{\gamma_i} \cdot p(s_t|s_{t-1}, w_1^N) \right\} \right\} \tag{4}$$

where $x_{t,i}$ is the feature set of model i at time step t, and γ_i is the weighting factor of modality i with $\sum_i^I \gamma_i = 1$.

The advantage of this approach is the ability to model each feature set separately. This allows faster individual training and optimization of the model

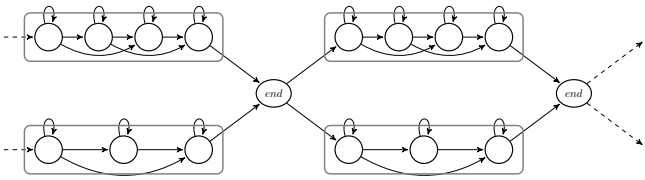

Fig. 1. Asynchronous Combination of HMM-based systems using parallel HMM

parameters for each modality. On the other hand, the different models are connected *lock-step*-wise, meaning that all models always are in the same HMM state and must have the same HMM topology, which is also restricting.

Asynchronous Combination

The idea of asynchronous combination is to recognize each word individually by each model and join and synchronize the decisions only at word boundaries. Each model can be freely aligned to the features of a word by choosing its own sequence of state transitions. Asynchronous combination is especially suited for sign language recognition because of signs being composed of several, not perfectly synchronous components.

As depicted in Fig. 1, the asynchronous strategy is able to use models with different HMM topologies. For simplification and better comparison between different approaches we chose to use only models of the same topology in this work. To formalize the asynchronous combination, we want to find the best sequence of states $s_{t_{n-1}+1,i}^{t_n,i}$ for each model i going from time step $t_{n-1}+1$ to t_n for word n where t_n is the time step where word w_n ends. For simplification, we put all states in time step t into a vector of states $\lambda_t = [s_{t,1}, \cdots, s_{t,i}, \cdots, s_{t,I}]^\top$. Leading to the recognition formula:

$$x_1^T \rightarrow [w_1^N]_{\text{opt}} = \underset{w_1^N}{\operatorname{argmax}} \max_{t_1^N} \left\{ \prod_{n=1}^{N} \left[p(w_n|w_1^{n-1}) \right.\right.$$

$$\left.\left. \cdot \max_{\lambda_{t_{n-1}+1}^{t_n}} \left\{ \prod_{t=t_{n-1}+1}^{t_n} p(x_t, \lambda_t|\lambda_{t-1}, w_t) \right\} \right] \right\} \tag{5}$$

The word conditioned tree search introduced in Section 2 is modified to match the asynchronous combination strategy. The probability $p(x_t, s|\sigma)$ depends now on state vector λ and the predecessor state vector $\nu = [\sigma_1, \cdots, \sigma_i, \cdots, \sigma_I]^\top$. So with $p(x_t, \lambda|\nu) = p(x_t|\lambda) \cdot p(\lambda|\nu)$ follows that recursion Equation 2 changes to

$$Q_v(t, \lambda) = \max_{\sigma} \{p(x_t|\lambda) \cdot p(\lambda|\nu) \cdot Q_v(t-1, \nu)\} \tag{6}$$

with

$$p(x_t, \lambda) = \prod_{i=1}^{I} p(x_t, \lambda_i)^{\gamma_i} \quad \text{and} \quad p(\lambda, \nu) = \prod_{i=1}^{I} p(\lambda_i, \sigma_i)^{\gamma_i} \tag{7}$$

Fig. 2. Example frames from SIGNUM (left) and PHOENIX (right)

Table 1. Single Signer Statistics of SIGNUM and PHOENIX

	SIGNUM		**PHOENIX**	
	Training	Test	Training	Test
# frames	416,620	114,230	46,282	6751
# sentences	1809	531	304	47
# running glosses	11,109	2805	3309	487
vocabulary size	455	-	266	-
# singletons	0	0	90	-
# out-of-vocabulary [%]	-	3.6	-	1.6
perplexity (3-gram)	17.8	72.2	15.9	34.9

satisfying $\sum_i^I \gamma_i = 1$. The second recursion Equation 3 accordingly changes to

$$Q_v(t, \lambda = \mathbf{0}) = \max_u \{p(v|u) \cdot Q_u(t, \Lambda_v)\} \tag{8}$$

with the vector of all ending states $\Lambda_v = [S_{v,1}, \cdots, S_{v,i}, \cdots, S_{v,I},]^\top$.

System Combination
The idea of system combination is to fuse decisions of individual systems on the sentence level. To this end an acyclic graph is created during the recognition process containing all possible word sequences which can be detected in the test data set by the used model. Each edge is labeled with a word and the probability to choose this word for the current path. The graphs of different systems are combined to choose the best common recognition path for each sentence using a modified version of the recognizer output voting error reduction ((i)ROVER) [7]. The great advantage of system combination is the ability to apply each recognizer isolated to the data with separately optimized parameters.

4 Databases

Modality combination techniques have been tested on the single signer setups of the SIGNUM database [1], representing 'lab-data' created for pattern recognition purposes, and the recently created RWTH-PHOENIX-Weather database (PHOENIX) [5], representing challenging 'real-life' data recorded from German public TV. Both databases contain videos of continuous signing in German sign language recorded by a single camera in frontal view and are annotated using

Table 2. Multi Signer Statistics of SIGNUM

	SIGNUM MS	
	Training	Test
# frames	3,618,630	996,270
# sentences	15,075	4425
# running glosses	92,575	23,350
vocabulary size	455	-
# singletons	0	0
# out-of-vocabulary [%]	-	3.6
perplexity (3-gram)	17.8	72.2

gloss notation effectively labeling the meaning of a sign rather than the actual visual appearance. Example frames from the single signer setups of both databases are shown in Fig. 2 and key statistics are subsumed in Table 1.

SIGNUM Database
In the SIGNUM database, a native signer wearing black clothes in front of a dark blue background signs predefined sentences taken from the domain 'daily life'. Videos are recorded at 780×580 pixels and 30 frames per second (fps). Special to the SIGNUM database is that every sentence of the 603 unique training and 177 unique testing sentences is performed thrice by the signer.

PHOENIX Database
Videos (25 fps, resolution 210×260 pixels) of the PHOENIX database show a hearing interpreter in dark clothes (short and long-sleeve) in front of a grayish, artificial background interpreting on-the-fly the spoken weather forecast of the German public TV-station PHOENIX. Containing 'real-life' data, the videos of PHOENIX pose a strong challenge to computer vision and recognition systems because of high signing speed, motion blur, out-of-plane rotations, strong facial expressions, and classifier signs.

5 Experimental Results

Experiments have been carried out using the setups described in Section 4. All error rates are reported in WER measuring the minimum number of insertions, deletions and substitutions needed to transform the recognized into the ground-truth sentence. For both databases, left-to-right HMMs with Bakis topology and Gaussian mixture densities trained using EM-algorithm and maximum likelihood criterion are used. The optimal number of HMM states for each gloss has been learned from single gloss, ground-truth annotation for PHOENIX and estimated from the state alignment of a baseline system for SIGNUM. The number of Gaussians has been optimized for the single modality systems for both databases. A trigram language model using modified Kneser-Ney discounting is used in all experiments. The language model scale (weighting of language versus visual model) is optimized for each experiment. In case of synchronous, asynchronous

Table 3. Single Modality Baseline Results on PHOENIX and SIGNUM

	WER%			WER%	
	PHOENIX	SIGNUM		PHOENIX	SIGNUM
Full Frame	80.1	31.6	Face	62.6	89.3
Right Hand	45.2	12.5	Right Hand Traj.	42.1	14.2
Left Hand	63.9	51.0			

Table 4. Modality Combination Results on PHOENIX and SIGNUM

	WER%		WER%	
	PHOENIX	SIGNUM	PHOENIX	SIGNUM
	Feature Combination		System Combination	
Full Frame + Right Hand	62.8	16.1	49.5	12.4
Full Frame + Left Hand	70.2	27.7	66.7	30.5
Right Hand + Left Hand	51.1	13.3	48.0	12.5
Right Hand + Face	45.8	33.8	48.3	12.5
Full Frame + Right Hand Traj	61.2	17.0	46.2	13.5
Right Hand Traj + Left Hand	45.6	14.4	46.2	13.9
Right Hand Traj + Face	45.0	28.8	46.2	13.9
	Synchronous Combination		Asynchronous Combination	
Full Frame + Right Hand	45.2	10.8	45.2	**10.7**
Full Frame + Left Hand	63.7	25.9	63.9	25.3
Right Hand + Left Hand	42.9	12.0	43.3	12.0
Right Hand + Face	45.0	12.5	44.4	12.5
Full Frame + Right Hand Traj	42.1	12.9	42.1	12.9
Right Hand Traj + Left Hand	**41.9**	13.5	**41.9**	13.5
Right Hand Traj + Face	**41.9**	14.2	42.1	14.2

and system combination the trained models of respective single modality systems are used and the weighting parameters of the modalities are optimized.

In the following, modalities **a** to **e** are referred to as Full Frame, Right Hand, Left Hand, Face and Right Hand Trajectory. Table 3 subsumes the baseline results obtained using only one modality. Comparing the results on PHOENIX to the results of SIGNUM the difference in recognition performance is striking. Containing 'real-life' data recorded outside the research lab, PHOENIX poses a harder challenge than SIGNUM. For example, considering Full Frame, the signer in SIGNUM is centered in the video and hardly changes his body pose while PHOENIX contains strong variations in body pose leading to a WER of 80.1%. Results for Left Hand are worse than Right Hand results for both databases because both signers are right dominant. In contrast to SIGNUM where Right Hand Trajectory does not improve over Right Hand due to slow signing, PHOENIX results are improved by the movement modality. Furthermore, SIGNUM contains hardly any facial expressions leading to a high WER for this modality.

Going beyond single modalities, combination experiments have been carried out using all single modality systems listed in Table 3. In this work only

combinations of two modalities are considered without loss of generality and the results are listed in Table 4. Experimental results using more than two modalities indicate similar findings and are omitted here for brevity. Please note that we are interested in the question of how to combine modalities and not in the question which modalities to combine to obtain optimal recognition performance.

All modality combination strategies investigated improve recognition results for all modality combinations considered over the baseline of the poorer of both modalities. Only synchronous and asynchronous combination yield results that are consistently either better than or equal to the best single modality system used in the respective combination. The combination of modalities into a joint feature space is too strict and suffers from the curse of dimensionality. System combination suffers from the low number of combined systems as well as systems making too similar recognition errors.

Furthermore, there is hardly any difference in results between synchronous and asynchronous combination but the time complexity of the asynchronous combination as presented here scales polynomial in the number of modalities while synchronous combination scales linearly. Currently, this makes the synchronous combination the method of choice since the added flexibility of the asynchronous combination due to different HMM topologies for each modality is not investigated in this work.

After evaluating the different combination strategies in single signer setups, recognition was also performed on the SIGNUM multi signer database (Table 2), consisting of 24 additional signers. Findings were similar to the single signer case albeit of higher recognition error rates. For example, asynchronous combination of full frame (49.5% single stream) and right-hand (23.6% single stream) modalities achieved 23.4% WER improving slightly over the best single stream result. For better comparison, the same parameters were applied to single and multi signer setups, better results should be achieved by optimizing parameters for the multi signer database.

6 Conclusions and Future Work

We investigated four modality combination techniques for five different modalities. The technique of asynchronous combination has been incorporated into a state-of-the-art, large-vocabulary sign language recognition system using word conditioned tree search and dynamic programming. Modality combination techniques have been evaluated on the two publicly available, large-vocabulary, sign language databases SIGNUM (lab-data) and PHOENIX (challenging real-life data). Synchronous and asynchronous combination strategies were found to outperform feature combination and system combination approaches in the context of combination of two modalities. Using modality combination strategies the best, to the best of our knowledge, published recognition results of 11.9% WER [6] on the SIGNUM database and 55.0% WER [5] on PHOENIX were improved to 10.7% WER respectively 41.9% WER. While synchronous and asynchronous combination differ hardly in results in the presented comparison,

the asynchronous combination has been restricted to the same HMM topology for all modalities as the synchronous combination for the sake of comparison. In future work, we will investigate the influence of the added flexibility of the asynchronous combination on recognition results for large vocabulary continuous sign language recognition.

Acknowledgments. This work has been funded by the European Community's Seventh Framework Programme FP7-ICT-2007-3 under grant agreement 231424 - SignSpeak and the Janggen-Pöhn-Stiftung. Special thanks to Thomas Hoyoux (Technical University of Innsbruck) and Yvan Richard (Centre de Recerca i Innovacio de Catalunya (CRIC)) for providing AAM and HoG3D features.

References

1. van Agris, U., Knorr, M., Kraiss, K.F.: The significance of facial features for automatic sign language recognition. In: FG, pp. 1–6 (September 2008)
2. Bengio, S.: An asynchronous hidden Markov model for audio-visual speech recognition. In: NIPS (2003)
3. Deng, J., Tsui, H.T.: A Two-step Approach based on PaHMM for the Recognition of ASL. In: ACCV (January 2002)
4. Dreuw, P., Deselaers, T., Rybach, D., Keysers, D., Ney, H.: Tracking using dynamic programming for appearance-based sign language recognition. In: FG, pp. 293–298 (2006)
5. Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., Ney, H.: Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In: LREC (May 2012)
6. Gweth, Y., Plahl, C., Ney, H.: Enhanced continuous sign language recognition using PCA and neural network features. In: CVPR 2012 Workshop on Gesture Recognition (June 2012)
7. Hoffmeister, B., Schlüter, R., Ney, H.: Icnc and Irover: The limits of improving system combination with classification? In: Interspeech, pp. 232–235 (September 2008)
8. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC, pp. 995–1004 (September 2008)
9. Luettin, J., Potamianos, G., Neti, C.: Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In: ICASSP, pp. 169–172 (2001)
10. Nakamura, S., Kumatani, K., Tamura, S.: Multi-modal temporal asynchronicity modeling by product HMMs for robust audio-visual speech recognition. In: Multimodal Interfaces, pp. 305–309 (2002)
11. Nefian, A.V., Liang, L., Pi, X., Liu, X., Murphy, K.: Dynamic bayesian networks for audio-visual speech recognition. EURASIP J. Appl. Signal Process. 2002(1), 1274–1288 (2002)
12. Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.: A coupled hmm for audio-visual speech recognition. In: ICASSP, pp. 2013–2016 (2002)
13. Ong, E.J., Cooper, H., Pugeault, N., Bowden, R.: Sign language recognition using sequential pattern trees. In: CVPR, June 16-21, pp. 2200–2207 (2012)
14. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., Ney, H.: The RWTH Aachen University open source speech recognition system. In: INTERSPEECH, pp. 2111–2114 (2009)

15. Theodorakis, S., Katsamanis, A., Maragos, P.: Product-HMMs for Automatic Sign Language Recognition. In: ICASSP, pp. 1601–1604 (2009)
16. Tran, K., Kakadiaris, I.A., Shah, S.K.: Fusion of human posture features for continuous action recognition. In: ECCV Workshop on Sign, Gesture and Activity, SGA (2011)
17. Verma, A., Faruquie, T., Neti, C., Basu, S., Senior, A.: Late integration in audio-visual continuous speech recognition. In: ASRU (1999)
18. Vogler, C., Metaxas, D.: Parallel Hidden Markov Models for American Sign Language Recognition. In: ICCV, pp. 116–122 (1999)
19. Wang, C., Chen, X., Gao, W.: Expanding Training Set for Chinese Sign Language Recognition. In: FG (2006)